

ARTICLE COMPILATIONS IN COMPLEX SYSTEMS

Transition from Chaotic to Nonchaotic Behavior in Randomly Driven Systems

S. Fahy

Department of Physics, University of Michigan, Ann Arbor, Michigan 48109-1120

D. R. Hamann

AT&T Bell Laboratories, Murray Hill, New Jersey 07974

(Received 18 February 1992)

We discuss the explicit dependence of the final trajectory on initial conditions for randomly driven nonlinear dynamical systems which are stopped and restarted with random velocities at regular intervals (a Brownian-type motion). We find a transition from chaotic behavior for long intervals between stops to nonchaotic behavior for short intervals between stops. For short intervals, the Lyapunov exponent is related to the thermal average square force due to the potential. The consequences for "hybrid molecular-dynamics Monte Carlo" sampling methods are discussed.

PACS numbers: 05.45.+b, 05.40.+j

In the study of nonlinear dynamical problems, it is a familiar fact that the trajectory of a particle may show extreme sensitivity to initial conditions, i.e., the system exhibits chaotic behavior [1]. The deterministic evolution of such a system appears to display a random character, and modifying its dynamics by adding truly random forces might be expected to make its behavior "more random." We recently discovered a counterintuitive contradiction of this notion: When an ensemble of particles with *different* initial conditions are driven by an *identical* sequence of random forces designed to simulate Brownian motion, their *trajectories may become identical* at long times. (Here, and in the rest of this paper, when we say that trajectories become identical, we mean that the average distance between them converges exponentially to zero.) It is well known that no matter what the initial position of a particle undergoing Brownian motion in a fixed external potential $V(x)$, the statistical distribution of its positions at long times is simply given by the Boltzmann distribution [2], proportional to $\exp[-V(x)/k_B T]$ for the appropriate temperature T . Our result entails a much stronger statement than the observation that the *statistical* distributions of Brownian trajectories become independent of initial positions; the ensemble of trajectories becomes point by point identical in time, following a single final trajectory, which is, however, highly erratic and random.

We consider a particle of mass m which moves according to Newton's equations (without friction) in a potential $V(x)$, except that at regular time intervals τ it is stopped and the components of its velocity are reset to random values chosen from a Gaussian distribution of variance $k_B T/m$ (i.e., the velocity is reset at regular intervals from a Maxwell distribution for temperature T). This motion is in many respects similar to Brownian motion of the particle at a temperature T . It can be shown that the distribution of positions of the particle for long times is just the Boltzmann distribution, independent of the value of τ chosen [3]. Indeed, this approach is frequently used in Monte Carlo simulations [3,4] to sample

points from a probability distribution $P(x)$ by choosing the classical potential $V(x) = -k_B T \ln[P(x)]$. (In the numerical simulations presented here we will set $k_B T$ and m equal to unity [5].)

In typical Monte Carlo applications, x represents a vector with many components, corresponding to motion of a particle in a high-dimensional space. While the phenomenon we describe was discovered in such a situation, the dimension of the space in which the particle moves does not appear to be crucial. For ease of visualization, we will present in detail here the behavior of a two-dimensional system chosen to be a "bad case" in a sense discussed below. This system has the quartic-plus-sinusoidal potential

$$V(x, y) = \sin(2\pi x)/2\pi x + \sin(2\pi y)/2\pi y + r^4/16\pi,$$

shown in Fig. 1. Shown in Fig. 2 is the mean-square distance $\langle r_{12}^2 \rangle$ between pairs of identically driven particles in this potential, first for 150 steps with a time between stops of $\tau = 2.5$, and then for 150 steps with $\tau = 1.0$. (By

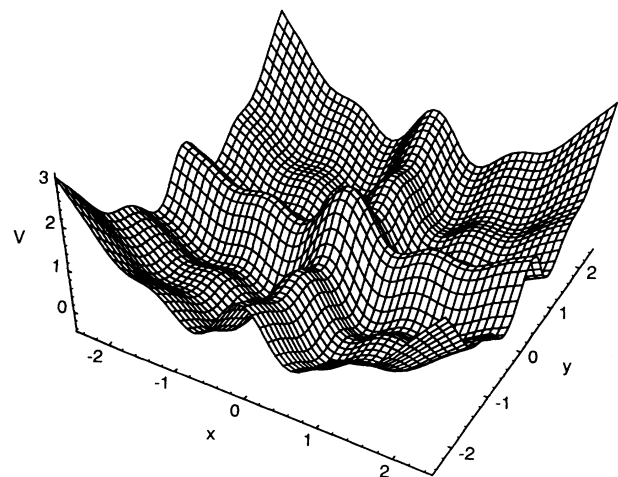


FIG. 1. The potential $V(x, y) = \sin(2\pi x)/2\pi x + \sin(2\pi y)/2\pi y + r^4/16\pi$.

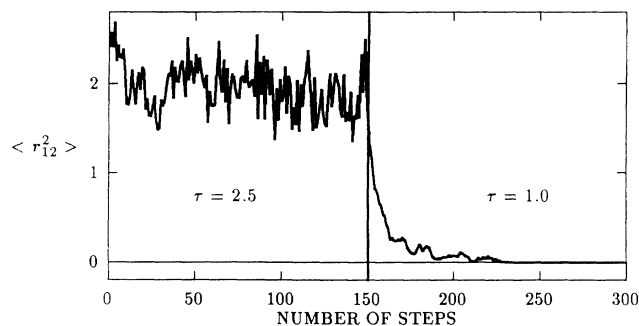


FIG. 2. The mean-square distance $\langle r_{12}^2 \rangle$ between pairs of identically driven particles vs number of steps, for the potential in Fig. 1. The particle coordinates were initially independently distributed from the Boltzmann distribution. The time τ between stops was 2.5 for the first 150 steps and was 1.0 for the second 150 steps. 100 independent simulations were averaged for the curve shown.

“identically driven,” we mean that both particles of each pair were given an identical, randomly chosen velocity [6] at the start of each step of length τ .)

The behavior of the particles shown in steps 150–300 of Fig. 2 is an illustration of a phenomenon which may be loosely stated as follows: *If the time interval τ between steps is lower than a threshold value τ_c , the final trajectory of the particle is entirely independent of the initial conditions to any required level of accuracy.* (For the motion shown in Fig. 2, the threshold τ_c is between 1 and 2.5.) When $\tau < \tau_c$, if two particles are started at entirely different positions but are “driven” by the same particular choice of velocities, they will (with probability 1) end up traveling along exactly the same trajectory at the same time. The final trajectory depends only on the choice of velocities, not on the initial position of the particle. We have observed this behavior in simulations of a variety of bounded systems (i.e., systems where the Boltzmann distribution effectively confines the particles to a finite region), with as many as 3000 coordinates, as often arise in Monte Carlo applications. We have not yet been able to find a bounded system for which the rule is violated [7]. We will prove the result for bounded one-dimensional systems, by showing that the appropriately defined Lyapunov exponent for the random motion is negative when τ is sufficiently short and is simply related to the thermal average of the square of the acceleration due to V .

Thus, although the trajectory of the particles is highly erratic, the system is *not* chaotic when $\tau < \tau_c$, because the final path (though random) is independent of the initial conditions. The threshold value τ_c and the rate of convergence of the trajectories for any given value of τ depend on the potential energy function $V(x)$ as well as on the variance $k_B T/m$ of the velocity components. If the time interval τ is greater than τ_c , the motion of initially uncorrelated particles subjected to an identical choice of

driving velocities is correlated at large times (as in the first 150 steps of Fig. 2) but never becomes identical. Indeed, particles that started out very close together will not exhibit any greater correlation in their motion at large times than particles that started far apart. In this regime, the system is chaotic in the usual sense of extreme sensitivity to initial conditions.

Among unbounded systems, an obvious counterexample to the result is for a potential $V(x)$ which is constant everywhere. In a periodic potential, the statement can only be true modulo a period of the potential. For example, in simulations of the two-dimensional periodic system with

$$V(x, y) = \{\cos(2\pi x) + \cos(2\pi y) + \cos^2[2\pi(x - y)]\}/2\pi,$$

we observe that particles started at random positions but driven by the same set of velocities end up following exactly identical motion except for a constant shift of a random number of complete periods of the potential in the x and y directions. In general, we suspect that the original statement will only be true for systems where the particle is confined to a finite region, either by the potential or by the geometry of the problem (e.g., by folding a periodic potential onto a torus).

For bounded linear systems, where $V(x)$ has a quadratic form, the final motion is independent of the initial motion for all values of τ ; i.e., the value of τ_c is infinite. The action of regularly stopping and restarting the particle can be thought of as an infinite damping force “turned on” for an infinitesimal time at intervals τ apart, followed by a driving force with the appropriate impulse also applied at intervals of τ . Because harmonic oscillators obey linear superposition, we see that after the damping term causes the initial conditions to decay, the final motion depends only on the velocities chosen to drive the particle. The damping analogy can be used for nonlinear systems also, but linear superposition is not obeyed, and the effects of initial conditions need not vanish at large times.

We now show analytically that for any one-dimensional potential V which confines particles to a finite region, and for short enough intervals τ between stops, the average rate of contraction γ of the distance between two particles initially close together (i.e., the negative of the Lyapunov exponent) is given by $\gamma = \gamma_0 + O(\tau^2)$, where

$$\gamma_0 = \tau \langle (\partial V / \partial x)^2 \rangle / 2mk_B T. \quad (1)$$

The angle brackets denote the thermal average (with respect to the Boltzmann distribution). To prove this, consider two particles initially at points x_0 and x'_0 close together. Let the two particles be started with equal velocity \dot{x} . For short times, the positions of the particles are given by $x(t) = x_0 + \dot{x}t + \ddot{x}t^2/2 + \dots$ and $x'(t) = x'_0 + \dot{x}t + \ddot{x}'t^2/2 + \dots$. The contraction of the distance between the particles at the end of one interval τ is then given by

$$\frac{x'(\tau) - x(\tau)}{x'_0 - x_0} \approx 1 + \frac{\tau^2}{2} \frac{\ddot{x}' - \ddot{x}}{x'_0 - x_0} \approx 1 - \frac{\tau^2}{2m} \frac{\partial^2 V}{\partial x^2},$$

since $m\ddot{x} = -\partial V/\partial x$. If the curvature of the potential $\partial^2 V/\partial x^2$ is greater than zero, the particles move closer together, and if it is less than zero, they move farther apart during the motion. So in general the particles do *not* always move closer together in each individual step. However, the probability that a particle is at a position x during its motion is proportional to $\exp[-V(x)/k_B T]$. Thus, the *average* contraction of the distance between identically driven particles in one step of length τ is given by

$$1 - \frac{\tau^2}{2Zm} \int \frac{\partial^2 V}{\partial x^2} \exp[-V(x)/k_B T] dx,$$

where Z is the normalization factor for the Boltzmann distribution. The average rate of contraction over many such steps is then

$$\gamma_0 = \frac{\tau}{2Zm} \int \frac{\partial^2 V}{\partial x^2} \exp[-V(x)/k_B T] dx.$$

Integration by parts (with the assumption that boundary terms vanish) gives the result $\gamma_0 = \tau \langle (\partial V/\partial x)^2 \rangle / 2mk_B T$. It is clear how this result breaks down for larger values of τ ; the proof relies on the expansion of the trajectories to second order in τ , which is not valid for large values of τ .

Intuitively, we may consider that the harmonic-oscillator argument above applies when the particle makes a sufficient number of stops within a region of positive curvature. Since positive curvature occurs near minima, and negative curvature near maxima of the potential, the Boltzmann weight favors the former regions. The potential in Fig. 1 was chosen to have a number of accessible regions of negative curvature so that it would clearly illustrate this point.

Although the proof in one dimension does not extend in an obvious way to higher-dimensional systems [8], the numerical evidence suggests that the qualitative nature of the result is true for bounded systems of arbitrary dimension. In fact, for most of the potentials we have investigated, the generalization of Eq. (1) to higher dimensions gives a reasonable estimate (within an order of magnitude) of the average rate of contraction of the distance between pairs of identically driven points in the limit of small τ , as shown in Fig. 3 [9]. Note that Fig. 3 shows the *asymptotic* contraction rate γ . For $\tau < \tau_c$, $-\gamma$ is the Lyapunov exponent, but for $\tau > \tau_c$ it is identically zero (because the particles are confined to a finite region) and is different from the Lyapunov exponent, which is positive. It is intriguing to find that the value of τ_c is similar for both systems shown in Fig. 3. However, we do not clearly understand at this stage the factors which determine τ_c . Moreover, although γ_0 gives the *average* rate of convergence of two close trajectories, fluctuations in the rate are an important aspect of the behavior of the system, especially near threshold τ_c .

Apart from its intrinsic interest, this result has some important consequences for Monte Carlo applications us-

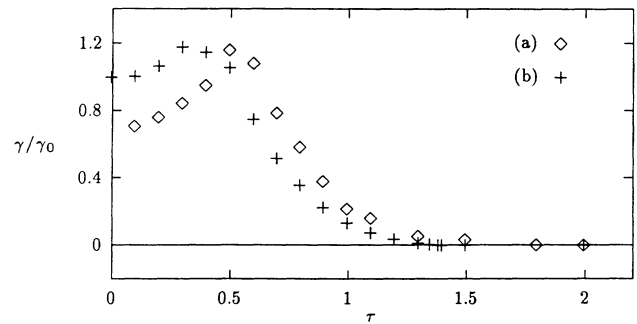


FIG. 3. Asymptotic average contraction rate γ , scaled by γ_0 from Eq. (1), vs τ for (a) the potential shown in Fig. 1 and (b) the one-dimensional Duffing potential, $V(x) = x^4 - x^2$.

ing the sampling idea described above. In many such applications, one must be concerned that measurements of a quantity at nearby times in the simulation will be strongly correlated with one another, substantially reducing the rate at which “effectively independent” samples are generated [4]. The phenomenon of critical slowing down is associated with such correlations becoming very long ranged. However, the autocorrelation time (the time required to generate “effectively” uncorrelated samples [10] from the Boltzmann distribution) depends on the quantity being measured. For example, near a phase transition, measurements of the order parameter relevant to the transition tend to have very long autocorrelation times, whereas the measurements of other quantities may be relatively well behaved [10]. It is clear that the present result (when applicable) ensures that the time required for trajectories to become identical is an absolute upper bound on the autocorrelation time for measurements of *all* quantities. Thus, it may be used to test for critical slowing down, even when the nature of the transition and the associated order parameter are unknown. However, we emphasize that this result provides no *solution* to such critical slowing down.

We have empirically observed the independence of initial conditions to occur in other, less natural, forms of driven motion for which the particles do not have a Boltzmann distribution. For example, independence of initial conditions also occurs for short time steps if some of the coordinates x_i (instead of the corresponding components \dot{x}_i of the velocity) are reset at random from a Gaussian distribution at the end of each step. This seems to suggest that something other than the Boltzmann distribution explicitly used in the derivation of Eq. (1) may be at the heart of the behavior.

It is possible to modify the stop-start motion in other ways and still observe the independence of the final trajectory on initial conditions. For a fixed value of τ one may mix in some of the “old” velocity \mathbf{v}_{old} with the random velocity \mathbf{v}_{ran} to get a new starting velocity, $\mathbf{v}_{\text{new}} = \alpha \mathbf{v}_{\text{old}} + \beta \mathbf{v}_{\text{ran}}$, where $\alpha^2 + \beta^2 = 1$. As α increases from 0 to 1, the motion changes gradually from the stop-start

kind already discussed to an uninterrupted conservative motion. A threshold value of $\alpha = \alpha_c$ separates nonchaotic behavior of the motion (for $\alpha < \alpha_c$) from chaotic behavior (for $\alpha > \alpha_c$). The value of α_c depends on τ . Simulations of motion in the periodic potential given above under Gaussian random driving forces applied at regular intervals with the addition of a constant damping proportional to velocity (but without stopping the particle at regular intervals) reveal a similar transition from nonchaotic behavior for large damping to chaotic behavior for small damping.

Finally, we note that the nature of this result (i.e., the exponential convergence of trajectories, with a well-defined Lyapunov exponent) ensures that it is insensitive to round-off and truncation errors in the numerical simulations. Of course, at what point the machine representations of two trajectories become truly identical depends on the machine precision.

S.F. wishes to thank D. Kessler, L. Sander, and R. Savit for useful discussions.

-
- [1] See, for example, G. L. Baker and J. P. Gollub, *Chaotic Dynamics: An Introduction* (Cambridge Univ. Press, Cambridge, 1990); also, *Noise and Chaos in Nonlinear Dynamical Systems*, edited by F. Moss, L. A. Lugiato, and W. Schleich (Cambridge Univ. Press, Cambridge, 1990).
 - [2] See, for example, R. Balescu, *Equilibrium and Non-Equilibrium Statistical Mechanics* (Wiley, New York, 1975).
 - [3] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, Phys. Lett. B **195**, 216 (1987), and references therein.
 - [4] R. T. Scalettar, D. J. Scalapino, and R. L. Sugar, Phys.

Rev. B **34**, 7911 (1986); S. R. White and J. W. Wilkins, Phys. Rev. B **37**, 5024 (1989); D. R. Hamann and S. Fahy, Phys. Rev. B **41**, 11 352 (1990).

- [5] Usually in Monte Carlo simulations there is an "accept-reject" part of each step also (see Ref. [3]), to allow for inexact energy conservation in integrating the equations of motion. This does not alter our results, and we have removed it in the simulations shown here, having verified that our results are insensitive to the step length used in numerical integration.
- [6] We have verified, using various subtractive and linear congruential pseudorandom number generators in choosing velocities, that our results do not depend on the details of the generator used.
- [7] There is no reason for the usual considerations limiting passage through energy barriers large compared to the temperature to be circumvented in the present problem [see S. Chandrasekar, Rev. Mod. Phys. **15**, 1 (1943)]. Presumably, for particles started in different regions, separated by large potential barriers, the time for them to "find" each other must be at least as great as the typical barrier penetration time.
- [8] The argument breaks down because pairs of particles entering a region tend to align themselves to some extent along the most slowly contracting direction in that region. Thus, the naive formal extension of the one-dimensional result to higher dimensions gives only an upper bound on the contraction rate as $\tau \rightarrow 0$ and does not rule out an average dilation.
- [9] The only potential we investigated for which Eq. (1) did not give a reasonable estimate was the quartic-plus-quadratic potential $V(x,y) = x^2 y^2 + (x^2 + y^2)/n$ for large values of n . As $n \rightarrow \infty$, this potential becomes nonbounding along the x and y axes, and the observed contraction rate tends to zero.
- [10] See, for example, K. Binder and D. W. Heermann, *Monte Carlo Simulation in Statistical Physics* (Springer, Berlin, 1988).

Stochastic resonance

Luca Gammaitoni

Dipartimento di Fisica, Università di Perugia, and Istituto Nazionale di Fisica Nucleare, Sezione di Perugia, VIRGO-Project, I-06100 Perugia, Italy

Peter Hänggi

Institut für Physik, Universität Augsburg, Lehrstuhl für Theoretische Physik I, D-86135 Augsburg, Germany

Peter Jung

Department of Physics and Astronomy, Ohio University, Athens, Ohio 45701

Fabio Marchesoni

Department of Physics, University of Illinois, Urbana, Illinois 61801 and Istituto Nazionale di Fisica della Materia, Università di Camerino, I-62032 Camerino, Italy

Over the last two decades, stochastic resonance has continuously attracted considerable attention. The term is given to a phenomenon that is manifest in nonlinear systems whereby generally feeble input information (such as a weak signal) can be amplified and optimized by the assistance of noise. The effect requires three basic ingredients: (i) an energetic activation barrier or, more generally, a form of threshold; (ii) a weak coherent input (such as a periodic signal); (iii) a source of noise that is inherent in the system, or that adds to the coherent input. Given these features, the response of the system undergoes resonance-like behavior as a function of the noise level; hence the name stochastic resonance. The underlying mechanism is fairly simple and robust. As a consequence, stochastic resonance has been observed in a large variety of systems, including bistable ring lasers, semiconductor devices, chemical reactions, and mechanoreceptor cells in the tail fan of a crayfish. In this paper, the authors report, interpret, and extend much of the current understanding of the theory and physics of stochastic resonance. They introduce the readers to the basic features of stochastic resonance and its recent history. Definitions of the characteristic quantities that are important to quantify stochastic resonance, together with the most important tools necessary to actually compute those quantities, are presented. The essence of classical stochastic resonance theory is presented, and important applications of stochastic resonance in nonlinear optics, solid state devices, and neurophysiology are described and put into context with stochastic resonance theory. More elaborate and recent developments of stochastic resonance theory are discussed, ranging from fundamental quantum properties—being important at low temperatures—over spatiotemporal aspects in spatially distributed systems, to realizations in chaotic maps. In conclusion the authors summarize the achievements and attempt to indicate the most promising areas for future research in theory and experiment. [S0034-6861(98)00101-9]

CONTENTS

I. Introduction	224	D. Weak-noise limit of stochastic resonance—power spectra	244
II. Characterization of Stochastic Resonance	226	V. Applications	246
A. A generic model	226	A. Optical systems	246
1. The periodic response	226	1. Bistable ring laser	246
2. Signal-to-noise ratio	228	2. Lasers with saturable absorbers	248
B. Residence-time distribution	229	3. Model for absorptive optical bistability	248
1. Level crossings	229	4. Thermally induced optical bistability in semiconductors	249
2. Input-output synchronization	230	5. Optical trap	250
C. Tools	231	B. Electronic and magnetic systems	251
1. Digital simulations	231	1. Analog electronic simulators	251
2. Analog simulations	231	2. Electron paramagnetic resonance	253
3. Experiments	231	3. Superconducting quantum interference devices	253
III. Two-State Model	232	C. Neuronal systems	254
IV. Continuous Bistable Systems	234	1. Neurophysiological background	254
A. Fokker-Planck description	234	2. Stochastic resonance, interspike interval histograms, and neural response to periodic stimuli	255
1. Floquet approach	234	3. Neuron firing and Poissonian spike trains	257
B. Linear-response theory	236	4. Integrate-and-fire models	259
1. Intrawell versus interwell motion	238	5. Neuron firing and threshold crossing	260
2. Role of asymmetry	239		
3. Phase lag	240		
C. Residence-time distributions	240		

VI. Stochastic Resonance—Carried On	261
A. Quantum stochastic resonance	261
1. Quantum corrections to stochastic resonance	262
2. Quantum stochastic resonance in the deep cold	263
B. Stochastic resonance in spatially extended systems	267
1. Global synchronization of a bistable string	267
2. Spatiotemporal stochastic resonance in excitable media	268
C. Stochastic resonance, chaos, and crisis	270
D. Effects of noise color	272
VII. Sundry Topics	274
A. Devices	274
1. Stochastic resonance and the dithering effect	274
B. Stochastic resonance in coupled systems	274
1. Two coupled bistable systems	274
2. Collective response in globally coupled bistable systems	275
3. Globally coupled neuron models	275
C. Miscellaneous topics on stochastic resonance	275
1. Multiplicative stochastic resonance	275
2. Resonant crossing	276
3. Aperiodic stochastic resonance	277
D. Stochastic resonance—related topics	277
1. Noise-induced resonances	277
2. Periodically rocked molecular motors	278
3. Escape rates in periodically driven systems	279
VIII. Conclusions and Outlook	279
Acknowledgments	281
Appendix: Perturbation Theory	281
References	283

I. INTRODUCTION

Users of modern communication devices are annoyed by any source of background hiss. Under certain circumstances, however, an extra dose of noise can in fact help rather than hinder the performance of some devices. There is now even a name for the phenomenon: *stochastic resonance*. It is presently creating a buzz in fields such as physics, chemistry, biomedical sciences, and engineering.

The mechanism of stochastic resonance is simple to explain. Consider a heavily damped particle of mass m and viscous friction γ , moving in a symmetric double-well potential $V(x)$ [see Fig. 1(a)]. The particle is subject to fluctuational forces that are, for example, induced by coupling to a heat bath. Such a model is archetypal for investigations in reaction-rate theory (Hänggi, Talkner, and Borkovec, 1990). The fluctuational forces cause transitions between the neighboring potential wells with a rate given by the famous Kramers rate (Kramers, 1940), i.e.,

$$r_K = \frac{\omega_0 \omega_b}{2\pi\gamma} \exp\left(-\frac{\Delta V}{D}\right). \quad (1.1)$$

with $\omega_0^2 = V''(x_m)/m$ being the squared angular frequency of the potential in the potential minima at $\pm x_m$, and $\omega_b^2 = |V''(x_b)|/m$ the squared angular frequency at the top of the barrier, located at x_b ; ΔV is the height of

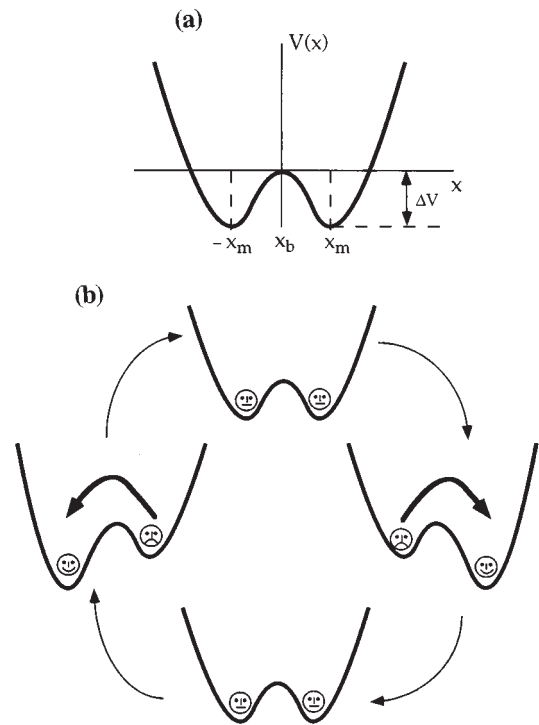


FIG. 1. Stochastic resonance in a symmetric double well. (a) Sketch of the double-well potential $V(x) = (1/4)bx^4 - (1/2)ax^2$. The minima are located at $\pm x_m$, where $x_m = (a/b)^{1/2}$. These are separated by a potential barrier with the height given by $\Delta V = a^2/(4b)$. The barrier top is located at $x_b = 0$. In the presence of periodic driving, the double-well potential $V(x, t) = V(x) - A_0 x \cos(\Omega t)$ is tilted back and forth, thereby raising and lowering successively the potential barriers of the right and the left well, respectively, in an antisymmetric manner. This cyclic variation is shown in our cartoon (b). A suitable dose of noise (i.e., when the period of the driving approximately equals twice the noise-induced escape time) will make the “sad face” happy by allowing synchronized hopping to the globally stable state (strictly speaking, this holds true only in the statistical average).

the potential barrier separating the two minima. The noise strength $D = k_B T$ is related to the temperature T .

If we apply a weak periodic forcing to the particle, the double-well potential is tilted asymmetrically up and down, periodically raising and lowering the potential barrier, as shown in Fig. 1(b). Although the periodic forcing is too weak to let the particle roll periodically from one potential well into the other one, noise-induced hopping between the potential wells can become synchronized with the weak periodic forcing. This statistical synchronization takes place when the average waiting time $T_K(D) = 1/r_K$ between two noise-induced interwell transitions is comparable with half the period T_Ω of the periodic forcing. This yields the *time-scale matching condition* for stochastic resonance, i.e.,

$$2T_K(D) = T_\Omega. \quad (1.2)$$

In short, stochastic resonance in a symmetric double-well potential manifests itself by a synchronization of activated hopping events between the potential minima

with the weak periodic forcing (Gammaitoni, Marchesoni, *et al.*, 1989). For a given period of the forcing T_Ω , the time-scale matching condition can be fulfilled by tuning the noise level D_{\max} to the value determined by Eq. (1.2).

The concept of stochastic resonance was originally put forward in the seminal papers by Benzi and collaborators (Benzi *et al.*, 1981, 1982, 1983) wherein they address the problem of the periodically recurrent ice ages. This very suggestion that stochastic resonance might rule the periodicity of the primary cycle of recurrent ice ages was raised independently by C. Nicolis and G. Nicolis (Nicolis, 1981, 1982, 1993; Nicolis and Nicolis, 1981). A statistical analysis of continental ice volume variations over the last 10^6 yr shows that the glaciation sequence has an average periodicity of about 10^5 yr. This conclusion is intriguing because the only comparable astronomical time scale in earth dynamics known so far is the modulation period of its orbital eccentricity caused by planetary gravitational perturbations. The ensuing variations of the solar energy influx (or solar constant) on the earth surface are exceedingly small, about 0.1%. The question climatologists (still) debate is whether a geodynamical model can be devised, capable of enhancing the climate sensitivity to such a small external periodic forcing. Stochastic resonance provides a simple, although not conclusive answer to this question (Matteucci, 1989, 1991; Winograd *et al.*, 1992; Imbrie *et al.*, 1993). In the model of Benzi *et al.* (1981, 1982, 1983), the global climate is represented by a double-well potential, where one minimum represents a small temperature corresponding to a largely ice-covered earth. The small modulation of the earth's orbital eccentricity is represented by a weak periodic forcing. Short-term climate fluctuations, such as the annual fluctuations in solar radiation, are modeled by Gaussian white noise. If the noise is tuned according to Eq. (1.2), synchronized hopping between the cold and warm climate could significantly enhance the response of the earth's climate to the weak perturbations caused by the earth's orbital eccentricity, according to arguments by Benzi *et al.* (1981, 1982).

A first experimental verification of the stochastic resonance phenomenon was obtained by Fauve and Heslot (1983), who studied the noise dependence of the spectral line of an ac-driven Schmitt trigger. The field then remained somewhat dormant until the modern age of stochastic resonance was ushered in by a key experiment in a bistable ring laser (McNamara, Wiesenfeld, and Roy, 1988). Soon after, prominent dynamical theories in the adiabatic limit (Gammaitoni, Marchesoni, Menichella-Saetta, and Santucci, 1989; McNamara and Wiesenfeld, 1989; Presilla, Marchesoni, and Gammaitoni, 1989; Hu *et al.*, 1990) and in the full nonadiabatic regime (Jung and Hänggi, 1989, 1990, 1991a) have been proposed. Moreover, descriptions in terms of the linear-response approximation have frequently been introduced to characterize stochastic resonance (Dykman *et al.*, 1990a, 1990b; Gammaitoni *et al.*, 1990; Dykman, Haken, *et al.*, 1993; Jung and Hänggi, 1991a; Hu, Haken, and Ning, 1992).

Over time, the notion of stochastic resonance has been widened to include a number of different mechanisms. The unifying feature of all these systems is the increased sensitivity to small perturbations at an optimal noise level. Under this widened notion of stochastic resonance, the first non-bistable systems discussed were excitable systems (Longtin, 1993). In contrast to bistable systems, excitable systems have only one stable state (the rest state), but possess a threshold to an excited state which is not stable and decays after a relatively long time (in comparison to the relaxation rate of small perturbations around the stable state) to the rest state. Soon afterwards, threshold detectors (see Sec. V.C, which presents cartoon versions of excitable systems) were discovered as a class of simple systems exhibiting stochastic resonance (Jung, 1994; Wiesenfeld *et al.* 1994; Gingl, Kiss, and Moss, 1995; Gammaitoni, 1995a; Jung, 1995). In the same spirit, stochastic-resonance-like features in purely autonomous systems have been reported (Hu, Ditzinger, *et al.*, 1993; Rappel and Strogatz, 1994).

The framework developed for excitable and threshold dynamical systems has paved the way for stochastic resonance applications in neurophysiology: stochastic resonance has been demonstrated in mechanoreceptor neurons located in the tail fan of crayfish (Douglass *et al.*, 1993) and in hair cells of crickets (Levin and Miller, 1996).

In the course of an ever-increasing flourishing of stochastic resonance, new applications with novel types of stochastic resonance have been discovered, and there seems to be no end in sight. Most recently, the notion of stochastic resonance has been extended into the domain of microscopic and mesoscopic physics by addressing the quantum analog of stochastic resonance (Löfstedt and Coppersmith, 1994a, 1994b; Grifoni and Hänggi, 1996a, 1996b) and also into the world of spatially extended, pattern-forming systems (spatiotemporal stochastic resonance) (Jung and Mayer-Kress, 1995; Löcher *et al.*, 1996; Marchesoni *et al.*, 1996; Wio, 1996; Castelpoggi and Wio, 1997; Vilar and Rubí, 1997). Other important extensions of stochastic resonance include stochastic resonance phenomena in coupled systems, reviewed in Sec. VII.B, and stochastic resonance in deterministic systems exhibiting chaos (see Sec. VI.C).

Stochastic resonance is by now a well-established phenomenon. In the following sections, the authors have attempted to present a comprehensive review of the present status of stochastic resonance theory, applications, and experimental evidences. After having introduced the reader into different quantitative measures of stochastic resonance, we outline the theoretical tools. A series of topical applications that are rooted in the physical and biomedical sciences are reviewed in some detail.

The authors trust that with the given selection of topics and theoretical techniques a reader will enjoy the tour through the multifaceted scope that underpins the physics of stochastic resonance. Moreover, this comprehensive review will put the reader at the very forefront of present and future stochastic resonance studies. The reader may also profit by consulting other, generally

more confined reviews and historical surveys, which in several aspects complement our work and/or provide additional insights into topics covered herein. In this context we refer the reader to the accounts given by Moss (1991, 1994), Jung (1993), Moss, Pierson, and O'Gorman (1994), Moss and Wiesenfeld (1995a, 1995b), Wiesenfeld and Moss (1995), Dykman, Luchinsky, *et al.* (1995), Bulsara and Gammaitoni (1996), as well as to the comprehensive proceedings of two recent conferences (Moss, Bulsara, and Shlesinger, 1993; Bulsara *et al.*, 1995).

II. CHARACTERIZATION OF STOCHASTIC RESONANCE

Having elucidated the main physical ideas of stochastic resonance in the preceding section, we next define the observables that actually quantify the effect. These observables should be physically motivated, easily measurable, and/or be of technical relevance. In the seminal paper by Benzi *et al.* (1981), stochastic resonance was quantified by the intensity of a peak in the power spectrum. Observables based on the power spectrum are indeed very convenient in theory and experiment, since they have immediate intuitive meaning and are readily measurable. In the neurophysiological applications of stochastic resonance another measure has become fashionable, namely the interval distributions between activated events such as those given by successive neuronal firing spikes or consecutive barrier crossings.

We follow here the historical development of stochastic resonance and first discuss important quantifiers of stochastic resonance based on the power spectrum. Along with the introduction of the quantifiers, we demonstrate their properties for two generic models of stochastic resonance; the periodically driven bistable two-state system and the double-well system. The detailed mathematical analysis of these models is the subject of Secs. III, IV, and the Appendix. Important results therein are used within this section to support a more intuitive approach. In a second part, we discuss quantifiers that are based on the interval distribution; these latter measures emphasize the synchronization aspect of stochastic resonance. We finish the section with a list of other, alternative methods and tools that have been used to study stochastic resonance. In addition, we present a list of experimental demonstrations.

A. A generic model

We consider the overdamped motion of a Brownian particle in a bistable potential in the presence of noise and periodic forcing

$$\dot{x}(t) = -V'(x) + A_0 \cos(\Omega t + \varphi) + \xi(t), \quad (2.1)$$

where $V(x)$ denotes the reflection-symmetric quartic potential

$$V(x) = -\frac{a}{2}x^2 + \frac{b}{4}x^4. \quad (2.2a)$$

By means of an appropriate scale transformation, cf.

Sec. IV.A, the potential parameters a and b can be eliminated such that Eq. (2.2a) assumes the dimensionless form

$$V(x) = -\frac{1}{2}x^2 + \frac{1}{4}x^4. \quad (2.2b)$$

In Eq. (2.1) $\xi(t)$ denotes a zero-mean, Gaussian white noise with autocorrelation function

$$\langle \xi(t)\xi(0) \rangle = 2D\delta(t) \quad (2.3)$$

and intensity D . The potential $V(x)$ is bistable with minima located at $\pm x_m$, with $x_m = 1$. The height of the potential barrier between the minima is given by $\Delta V = \frac{1}{4}$ [see Fig. 1(a)].

In the absence of periodic forcing, $x(t)$ fluctuates around its local stable states with a statistical variance proportional to the noise intensity D . Noise-induced hopping between the local equilibrium states with the Kramers rate

$$r_K = \frac{1}{\sqrt{2}\pi} \exp\left(-\frac{\Delta V}{D}\right) \quad (2.4)$$

enforces the mean value $\langle x(t) \rangle$ to vanish.

In the presence of periodic forcing, the reflection symmetry of the system is broken and the mean value $\langle x(t) \rangle$ does not vanish. This can be intuitively understood as the consequence of the periodic biasing towards one or the other potential well.

Filtering all the information about $x(t)$, except for identifying in which potential well the particle resides at time t (known as two-state filtering), one can achieve a binary reduction of the two-state model (McNamara and Wiesenfeld, 1989). The starting point of the two-state model is the master equation for the probabilities n_{\pm} of being in one of the two potential wells denoted by their equilibrium positions $\pm x_m$, i.e.,

$$\dot{n}_{\pm}(t) = -W_{\mp}(t)n_{\pm} + W_{\pm}(t)n_{\mp}, \quad (2.5)$$

with corresponding transition rates $W_{\mp}(t)$. The periodic bias toward one or the other state is reflected in a periodic dependence of the transition rates; see Eq. (3.3) below.

1. The periodic response

For convenience, we choose the phase of the periodic driving $\varphi = 0$, i.e., the input signal reads explicitly $A(t) = A_0 \cos(\Omega t)$. The mean value $\langle x(t) | x_0, t_0 \rangle$ is obtained by averaging the inhomogeneous process $x(t)$ with initial conditions $x_0 = x(t_0)$ over the ensemble of the noise realizations. Asymptotically ($t_0 \rightarrow -\infty$), the memory of the initial conditions gets lost and $\langle x(t) | x_0, t_0 \rangle$ becomes a periodic function of time, i.e., $\langle x(t) \rangle_{as} = \langle x(t + T_{\Omega}) \rangle_{as}$ with $T_{\Omega} = 2\pi/\Omega$. For small amplitudes, the response of the system to the periodic input signal can be written as

$$\langle x(t) \rangle_{as} = \bar{x} \cos(\Omega t - \bar{\phi}), \quad (2.6)$$

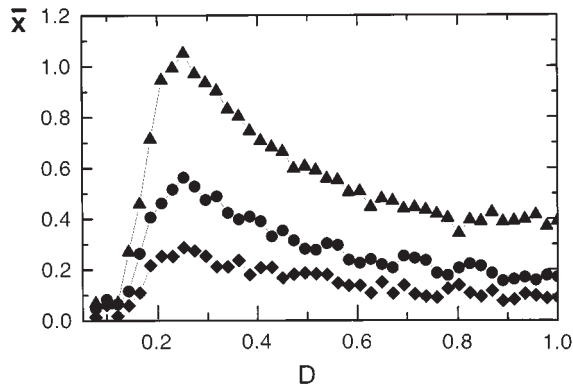


FIG. 2. Amplitude $\bar{x}(D)$ of the periodic component of the system response (2.6) vs the noise intensity D (in units of ΔV) for the following values of the input amplitude: $A_0 x_m / \Delta V = 0.4$ (triangles), $A_0 x_m / \Delta V = 0.2$ (circles), and $A_0 x_m / \Delta V = 0.1$ (diamonds) in the quartic double-well potential (2.2a) with $a = 10^4 \text{ s}^{-1}$, $x_m = 10$ (in units $[x]$ used in the experiment), and $\Omega = 100 \text{ s}^{-1}$.

with amplitude \bar{x} and a phase lag $\bar{\phi}$. Approximate expressions for the amplitude and phase shift read

$$\bar{x}(D) = \frac{A_0 \langle x^2 \rangle_0}{D} \frac{2r_K}{\sqrt{4r_K^2 + \Omega^2}} \quad (2.7a)$$

and

$$\bar{\phi}(D) = \arctan\left(\frac{\Omega}{2r_K}\right), \quad (2.7b)$$

where $\langle x^2 \rangle_0$ is the D -dependent variance of the stationary unperturbed system ($A_0 = 0$). Equation (2.7) has been shown to hold in leading order of the modulation amplitude $A_0 x_m / D$ for both discrete and continuous one-dimensional systems (Nicolis, 1982; McNamara and Wiesenfeld, 1989; Presilla, Marchesoni, and Gammaitoni, 1989; Hu, Nicolis, and Nicolis, 1990). While postponing a more accurate discussion of the validity of the above equations for \bar{x} and $\bar{\phi}$ to Sec. IV.B, we notice here that Eq. (2.7a) allows within the two-state approximation, i.e., $\langle x^2 \rangle_0 = x_m^2$, a direct estimate for the noise intensity D_{SR} that maximizes the output \bar{x} versus D for fixed driving strength and driving frequency.

The first and most important feature of the amplitude \bar{x} is that it *depends* on the noise strength D , i.e., the *periodic* response of the system can be manipulated by changing the noise level. At a closer inspection of Eq. (2.7), we note that the amplitude \bar{x} first increases with increasing noise level, reaches a maximum, and then decreases again. This is the celebrated *stochastic resonance* effect. In Fig. 2, we show the result of a simulation of the double-well system [Eqs. (2.1)–(2.3)] for several weak amplitudes of the periodic forcing A_0 . Upon decreasing the driving frequency Ω , the position of the peak moves to smaller noise strength (see Fig. 6, below).

Next we attempt to assign a physical meaning to the value of D_{SR} . The answer was given originally by Benzi and co-workers (Benzi *et al.* 1981, 1982, 1983): an unper-

turbed bistable system with $A_0 = 0$ switches spontaneously between its stable states with rate r_K . The input signal modulates the symmetric bistable system, making successively one stable state less stable than the other over half a period of the forcing. Tuning the noise intensity so that the random-switching frequency r_K is made to agree closely with the forcing angular frequency Ω , the system attains the maximum probability for an escape out of the less stable state into the more stable one, before a random back switching event takes place. When the noise intensity D is too small ($D \ll D_{SR}$), the switching events become very rare; thus the periodic component(s) of the interwell dynamics are hardly visible. Under such circumstances, the periodic component of the output signal $x(t)$ is determined primarily by motion around the potential minima—the intrawell motion. A similar loss of synchronization happens in the opposite case when $D \gg D_{SR}$: The system driven by the random source flips too many times between its stable states within each half forcing period for the forced components of the interwell dynamics to be statistically relevant.

In this spirit, the time-scale matching condition in Eq. (1.2), which with $T_K = 1/r_K$ is recast as $\Omega = \pi r_K$, provides a reasonable condition for the maximum of the response amplitude \bar{x} . Although the time-scale matching argument yields a value for D_{SR} that is reasonably close to the exact value it is important to note that it is *not exact* (Fox and Lu, 1993). Within the two-state model, the value D_{SR} obeys the transcendental equation

$$4r_K^2(D_{SR}) = \Omega^2(\Delta V/D_{SR} - 1), \quad (2.8)$$

obtained from Eq. (2.7a). The time-scale matching condition obviously does not fulfill Eq. (2.8); thus underpinning its approximate nature.

The phase lag $\bar{\phi}$ exhibits a transition from $\bar{\phi} = \pi/2$ at $D = 0^+$ to $\bar{\phi} \propto \Omega$ in the vicinity of D_{SR} . By taking the second derivative of the function $\bar{\phi}$ in Eq. (2.7b) and comparing with Eq. (2.8) one easily checks that D_{SR} lies on the right-hand side of the point of inflection of $\bar{\phi}$, being $\bar{\phi}''(D_{SR}) > 0$.

It is important to note that the variation of the angular frequency Ω at a fixed value of the noise intensity D does not yield a resonance-like behavior of the response amplitude. This behavior is immediately evident from Eq. (2.7a) and also from numerical studies (for those who don't trust the theory). A more refined analysis (Thorwart and Jung, 1997) shows that the decomposition of the susceptibility into its real and imaginary parts restores a nonmonotonic frequency dependence—see also the work on dynamical hysteresis and stochastic resonance by Phillips and Schulten (1995), and Mahato and Shenoy (1994).

Finally, we introduce an alternative interpretation of the quantity $\bar{x}(D)$ due to Jung and Hänggi (1989, 1991a): the integrated power p_1 stored in the delta-like spikes of $S(\omega)$ at $\pm \Omega$ is $p_1 = \pi \bar{x}^2(D)$. Analogously, the modulation signal carries a total power $p_A = \pi A_0^2$. Hence the spectral amplification reads

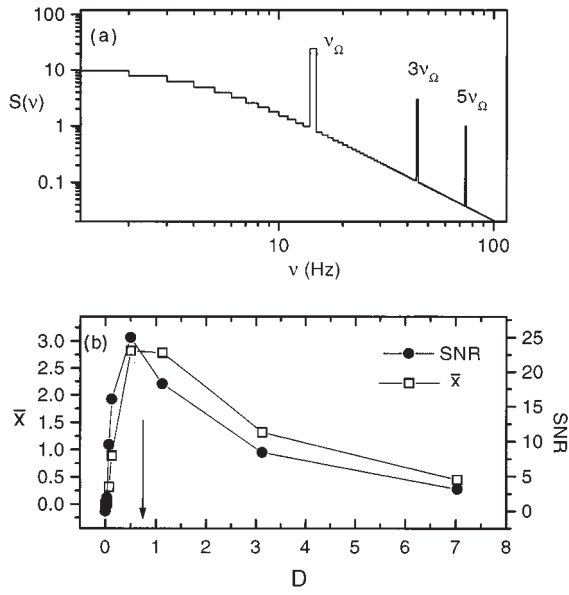


FIG. 3. Characterization of stochastic resonance. (a) A typical power spectral density $S(\nu)$ vs frequency ν for the case of the quartic double-well potential in Eq. (2.2a). The delta-like spikes at $(2n+1)\nu_\Omega$, $\Omega=2\pi\nu_\Omega$, with $n=0, 1$, and 2 , are displayed as finite-size histogram bins. (b) Strength of the first delta spike, Eq. (2.12), and the signal-to-noise ratio SNR , Eq. (2.13), vs D (in units of ΔV). The arrow denotes the D value corresponding to the power spectral density plotted in (a). The other parameters are $Ax_m/\Delta V=0.1$, $a=10^4 \text{ s}^{-1}$, and $x_m=10$ (in units $[x]$ used in the experiment).

$$\eta \equiv p_1/p_A = [\bar{x}(D)/A_0]^2. \quad (2.9)$$

In the linear-response regime of Eq. (2.7), η is independent of the input amplitude. This spectral amplification η will frequently be invoked in Sec. IV, instead of $\bar{x}(D)$.

2. Signal-to-noise ratio

Instead of taking the ensemble average of the system response, it sometimes can be more convenient to extract the relevant *phase-averaged* power spectral density $S(\omega)$, defined here as (see Secs. III and IV.A)

$$S(\omega) = \int_{-\infty}^{+\infty} e^{-i\omega\tau} \langle \langle x(t+\tau)x(t) \rangle \rangle d\tau, \quad (2.10)$$

where the inner brackets denote the ensemble average over the realizations of the noise and outer brackets indicate the average over the input initial phase φ . In Fig. 3(a) we display a typical example of $S(\nu)$ ($\omega=2\pi\nu$) for the bistable system. Qualitatively, $S(\omega)$ may be described as the superposition of a background power spectral density $S_N(\omega)$ and a structure of delta spikes centered at $\omega=(2n+1)\Omega$ with $n=0, \pm 1, \pm 2, \dots$. The generation of only *odd* higher harmonics of the input frequency are typical fingerprints of periodically driven symmetric nonlinear systems (Jung and Hänggi, 1989). Since the strength (i.e., the integrated power) of such spectral spikes decays with n according to a power law

such as A_0^{2n} , we can restrict ourselves to the first spectral spike, being consistent with the linear-response assumption implicit in Eq. (2.6). For small forcing amplitudes, $S_N(\omega)$ does not deviate much from the power spectral density $S_N^0(\omega)$ of the unperturbed system. For a bistable system with relaxation rate $2r_K$, the hopping contribution to $S_N^0(\omega)$ reads

$$S_N^0(\omega) = 4r_K \langle x^2 \rangle_0 / (4r_K^2 + \omega^2). \quad (2.11)$$

The spectral spike at Ω was verified experimentally (Debnath, Zhou, and Moss, 1989; Gammaitoni, Marchesoni, *et al.*, 1989; Gammaitoni, Menichella-Saetta, Santucci, Marchesoni, and Presilla, 1989; Zhou and Moss, 1990) to be a delta function, thus signaling the presence of a periodic component with angular frequency Ω in the system response [Eq. (2.6)]. In fact, for $A_0x_m \ll \Delta V$ we are led to separate $x(t)$ into a noisy background (which coincides, apart from a normalization constant, with the unperturbed output signal) and a periodic component with $\langle x(t) \rangle_{as}$ given by Eq. (2.6) (Jung and Hänggi, 1989). On adding the power spectral density of either component, we easily obtain

$$S(\omega) = (\pi/2) \bar{x}(D)^2 [\delta(\omega - \Omega) + \delta(\omega + \Omega)] + S_N(\omega), \quad (2.12)$$

with $S_N(\omega) = S_N^0(\omega) + \mathcal{O}(A_0^2)$ and $\bar{x}(D)$ given in Eq. (2.7a). In Fig. 3(b) the strength of the delta-like spike of $S(\omega)$ (more precisely \bar{x}) is plotted as a function of D .

Stochastic resonance can be envisioned as a particular problem of signal extraction from background noise. It is quite natural that a number of authors tried to characterize stochastic resonance within the formalism of data analysis, most notably by introducing the notion of signal-to-noise ratio (SNR) (McNamara *et al.*, 1988; Debnath *et al.*, 1989; Gammaitoni, Marchesoni, *et al.*, 1989; Vemuri and Roy, 1989; Zhou and Moss, 1990; Gong *et al.*, 1991, 1992). We adopt here the following definition of the signal-to-noise ratio

$$SNR = 2 \left[\lim_{\Delta\omega \rightarrow 0} \int_{\Omega-\Delta\omega}^{\Omega+\Delta\omega} S(\omega) d\omega \right] / S_N(\Omega). \quad (2.13)$$

Hence on combining Eqs. (2.11) and (2.12), the SN ratio for a symmetric bistable system reads in leading order

$$SNR = \pi(A_0x_m/D)^2 r_K. \quad (2.14)$$

Note that the factor of 2 in the definition (2.13) was introduced for convenience, in view of the power spectral density symmetry $S(\omega) = S(-\omega)$. The SN ratio SNR for the power spectral density plotted in Fig. 3(a) versus frequency ν ($\omega=2\pi\nu$) is displayed in Fig. 3(b). The noise intensity \bar{D}_{SR} at which SNR assumes its maximum *does not coincide* with the value D_{SR} that maximizes the response amplitude \bar{x} , or equivalently the strength of the delta spike in the power spectrum given by Eq. (2.12). As a matter of fact, if the prefactor of the Kramers rate is independent of D , we find that the SN ratio of Eq. (2.14) has a maximum at

$$\bar{D}_{SR} = \Delta V/2. \quad (2.15)$$

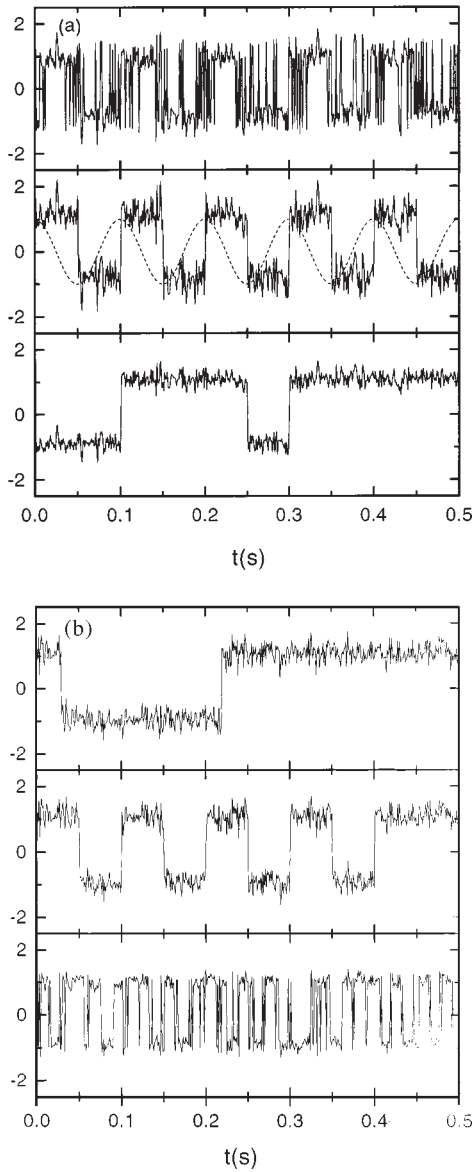


FIG. 4. Example of input/output synchronization in the symmetric bistable system of Eqs. (2.1)–(2.2a). (a) Varying the noise intensity D with Ω held constant. The sampled signal shown with dashes is the input $A(t)$ (arbitrary units). The remaining trajectories are the corresponding system output (in units of x_m) for increasing D values (from bottom to top). (b) Effect of varying Ω with D held constant. The three output samples $x(t)$ (in units of x_m) are displayed for increasing Ω values (from top to bottom). The parameters for (a) and (b) are $Ax_m/\Delta V=0.1$, $a=10^4 \text{ s}^{-1}$, and $x_m=(a/b)^{1/2}=10$, cf. in Fig. 2.

B. Residence-time distribution

In Sec. II.A we interpreted the resonant-like dependence of the amplitude $\bar{x}(D)$ of the periodic response on the noise intensity D by means of a synchronization argument, originally formulated by Benzi and co-workers (Benzi *et al.*, 1981). Moreover, we pointed out that the response amplitude does not show this synchronization if the driving frequency Ω is tuned against the

escape rate r_K . However, any experimentalist who ever tried to reproduce stochastic resonance in a real system (including here the analog circuits) knows by experience that a synchronization phenomenon takes place any time the condition $r_K \sim \Omega$ is established by varying either D or Ω . In Figs. 4(a) and 4(b) we depict the typical input-output synchronization effect in the bistable system Eqs. (2.1)–(2.3). In Fig. 4(a) the noise intensity is increased from low (rare random switching events) up to very large values, crossing the resonance values D_{SR} of Eqs. (2.8). In the latter case the output signal $x(t)$ becomes tightly locked to the periodic input. In Fig. 4(b), the noise intensity D is kept fixed and the forcing frequency Ω is increased. At low values of Ω , we notice an alternate asymmetry of the output signal towards either positive or negative values, depending on the sign of the input signal. However, many switches occur in both directions within any half forcing period. At large values of Ω , the effect of the time modulation is averaged out and the symmetry of the output signal seems to be fully restored. Finally, at $\Omega \sim r_K$ the synchronization mechanism is established with clear resemblance to Fig. 4(a). In the following subsection we characterize stochastic resonance as a “resonant” synchronization phenomenon, resulting from the combined action of noise and periodic forcing in a bistable system. The tool employed to this purpose is the residence-time distribution. Introduced as a tool (Gammaitoni, Marchesoni, *et al.*, 1989; Zhou and Moss, 1990; Zhou, *et al.*, 1990; Löfstedt and Coppersmith, 1994b; Gammaitoni, Marchesoni, and Santucci, 1995), such a notion proved useful for applications in diverse areas of natural sciences (Bulsara *et al.*, 1991; Longtin *et al.*, 1991; Simon and Libchaber, 1992; Carroll and Pecora, 1993b; Gammaitoni, Marchesoni, *et al.* 1993; Mahato and Shenoy, 1994; Mannella *et al.*, 1995; Shulgin *et al.*, 1995).

1. Level crossings

A deeper understanding of the mechanism of stochastic resonance in a bistable system can be gained by mapping the continuous stochastic process $x(t)$ (the system output signal) into a *stochastic point process* $\{t_i\}$. The symmetric signal $x(t)$ is converted into a point process by setting two crossing levels, for instance at $x_{\pm} = \pm c$ with $0 \leq c \leq x_m$. On sampling the signal $x(t)$ with an appropriate time base, the times t_i are determined as follows: data acquisition is triggered at time $t_0=0$ when $x(t)$ crosses, say, x_- with negative time derivative [$x(0)=-c$, $\dot{x}(0)<0$]; t_1 is the subsequent time when $x(t)$ first crosses x_+ with positive derivative [$x(t_1)=c$, $\dot{x}(t_1)>0$]; t_2 is the time when $x(t)$ switches back to negative values by recrossing x_- with negative derivative, and so on. The quantities $T(i)=t_i-t_{i-1}$ represent the residence times between two subsequent switching events. For simplicity and to make contact with the theory of Sec. IV.C, we set $c=x_m$. The statistical properties of the stochastic point process $\{t_i\}$ are the subject of intricate theorems of probability theory (Rice, 1944; Papoulis, 1965; Blake and Lindsey, 1973). In particular, no systematic way is known to find the distribution of

threshold crossing times. An exception is the symmetric bistable system: here, the *long* intervals T of consecutive crossings obey Poissonian statistics with an exponential distribution (Papoulis, 1965)

$$N(T) = (1/T_K) \exp(-T/T_K). \quad (2.16)$$

The distribution (2.16) is important for the forthcoming discussion, because it describes to a good approximation the first-passage time distribution between the potential minima in unmodulated bistable systems [see also Hänggi, Talkner, and Borkovec (1990), and references therein].

2. Input-output synchronization

In the absence of periodic forcing, the residence time distribution has the exponential form of Eq. (2.16). In the presence of the periodic forcing (Fig. 5), one observes a series of peaks, centered at odd multiples of the half driving period $T_\Omega = 2\pi/\Omega$, i.e., at $T_n = (n - \frac{1}{2})T_\Omega$, with $n = 1, 2, \dots$. The heights of these peaks decrease exponentially with their order n . These peaks are simply explained: the best time for the system to switch between the potential wells is when the relevant potential barrier assumes a minimum. This is the case when the potential $V(x, t) = V(x) - A_0 x \cos(\Omega t + \varphi)$ is tilted most extremely to the right or the left (in whichever well the system is residing). If the system switches at this time into the other well it then takes *half a period* waiting time in the other well until the new relevant barrier assumes a minimum. Thus $T_\Omega/2$ is a preferred residence interval. If the system “misses” a “good opportunity” to jump, it has to wait another full period until the relevant potential barrier for a switch again assumes the minimum. The second peak in the residence-time distribution is therefore located at $3/2T_\Omega$. The location of the other peaks is evident. The peak heights decay exponentially because the probabilities of the system to jump over a minimal barrier are statistically independent. We now argue that the strength P_1 of the first peak at $T_\Omega/2$ (the area under the peak) is a measure of the synchronization between the periodic forcing and the switching between the wells: If the mean residence time of the system in one potential well is much larger than the period of the driving, the system is not likely to jump the first time the relevant potential barrier assumes its minimum. The escape-time distribution exhibits in such a case a large number of peaks where P_1 is small. If the mean residence-time of the system in one well is much shorter than the period of the driving, the system will not “wait” with switching until the relevant potential barrier assumes its maximum and the residence-time distribution has already decayed practically to zero before the time $T_\Omega/2$ is reached and the weight P_1 is again small. Optimal synchronization, i.e., a maximum of P_1 , is reached when the mean residence time matches half the period of the driving frequency, i.e., our old time-scale matching condition Eq. (1.2). This resonance condition can be achieved by varying either Ω or D . This is demonstrated in Figs. 5(a) and 5(b). In the insets we

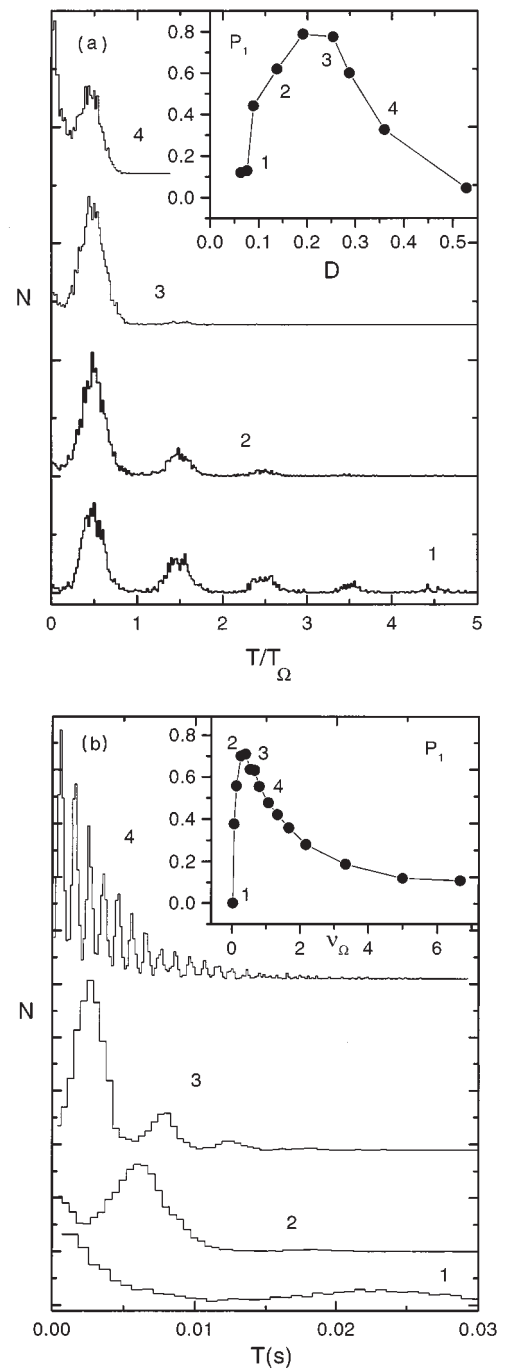


FIG. 5. Residence time distributions $N(T)$ for the symmetric bistable system of Eqs. (2.1)–(2.2a). (a) Increasing D (from below) with Ω held constant; inset: the strength P_1 of the first peak of $N(T)$ vs D (in units of ΔV). The definition of P_1 is as in Eq. (4.67) with $\alpha = 1/4$. (b) Increasing Ω (from below) with D held constant; inset: P_1 versus ν_Ω . Here, Ω is in units of r_K . The numbers 1–4 in the insets correspond to the D values (a) and the Ω values (b) of the distributions on display. The parameters for (a) and (b) are $Ax_m/\Delta V = 0.1$, $a = 10^4 \text{ s}^{-1}$, and $x_m = (a/b)^{1/2} = 10$, cf. Fig. 2.

have plotted the strength of the peak at $T_\Omega/2$ as a function of the noise strength D [Fig. 5(a)] and as a function of the driving frequency Ω [Fig. 5(b)]. In passing, we

anticipate that also the remaining peaks of $N(T)$ at T_n with $n > 1$ exhibit stochastic resonance [see Sec. IV.C].

We conclude with a comment on the multi-peaked residence-time distributions: the existence of peaks in the residence-time distribution $N(T)$ at T_n with $n > 1$ should not mislead the reader to think that the power spectral density $S(\omega)$ exhibits subharmonics of the fundamental frequency Ω (i.e., delta spikes at frequencies smaller than Ω). Although it may happen that the system waits for an odd number of half forcing periods (i.e., an integer number of extra “wait loops”) before switching states, such occurrences are randomly spaced in time and, therefore, do not correspond to any definite spectral component (Papoulis, 1965).

C. Tools

The seminal paper by Benzi *et al.* (1981) provoked no immediate reaction in the literature. Apart from a few early theoretical studies by Nicolis (1982), Eckmann and Thomas (1982), and Benzi *et al.* (1982, 1983, 1985), only one experimental paper (Fauve and Heslot, 1983) addressed the phenomenon of stochastic resonance. One reason may be the technical difficulty of treating nonstationary Fokker-Planck equations with time-dependent coefficients (Jung, 1993). Moreover, extensive numerical computations were not yet everyday practice. The experimental article by McNamara, Wiesenfeld, and Roy (1988) marked a *renaissance* of stochastic resonance, which has flourished and developed ever since in different directions. Our present knowledge of this topic has been reached through a variety of investigation tools. In the following paragraphs, we outline the most popular ones, with particular attention given to their advantages and limitations.

1. Digital simulations

The first evidence of stochastic resonance was produced by simulating the Budyko-Sellers model of climate change (Benzi *et al.*, 1982) on a Digital Instruments mainframe computer (model PDP 1000), an advanced computer at the time! Nowadays accurate digital simulations of either continuous or discrete stochastic processes can be carried out at home on unsophisticated personal desktop computers. Regardless of the particular algorithm adopted in the diverse cases, digital simulations proved particularly useful in the study of stochastic resonance in numerous cases (Nicolis *et al.*, 1990; Dayan *et al.*, 1992; Mahato and Shenoy, 1994; Masoliver *et al.*, 1995) and in chaotic (Carroll and Pecora, 1993a, 1993b; Hu, Haken, and Ning, 1993; Ippen *et al.*, 1993; Anishchenko *et al.*, 1994) or spatially extended systems (Neiman and Schimansky-Geier, 1994, 1995; Jung and Mayer-Kress, 1995; Lindner *et al.*, 1995). A decisive contribution to the understanding of the stochastic resonance phenomenon was produced in Augsburg (Germany) by Jung and Hänggi (1989, 1990, 1991a, 1991b, 1993), who encoded the matrix-continued fraction algorithm (Risken, 1984; Risken and Vollmer, 1989). Convergence problems at low noise intensities

and small driving frequencies, due to the truncation procedure, are the main limitations of this algorithm.

2. Analog simulations

This type of simulation allows more flexibility than digital simulation and for this reason has been preferred by a number of researchers. Rather than quoting all of them individually, we mention here the prominent groups, including those active in Perugia (Italy) (Gammaitoni, Marchesoni, *et al.*, 1989, 1993, 1994, 1995; Gammaitoni, Menichella-Saetta, *et al.*, 1989), in St. Louis (USA) (Debnath *et al.*, 1989; Zhou and Moss, 1990; Moss, 1991, 1994), in Lancaster (England) (Dykman, Mannella, *et al.*, 1990b, 1992) and in Beijing (China) (Hu *et al.*, 1991; Gong *et al.*, 1991, 1992). Analog simulators of stochastic processes are easy to design and assemble. Their results are not as accurate as digital simulations, but offer some advantages: (a) a large range of parameter space can be explored rather quickly; (b) high-dimensional systems may be simulated more readily than by computer, though systematic inaccuracies must be estimated and treated carefully. The block diagram of the Perugia simulator is illustrated in Sec. V.B.1 for the case of a damped quartic double-well oscillator subjected to both noisy and periodic driving. In passing, we mention here that Figs. 1–5 were actually obtained by means of that simulation circuit. In order to give the reader an idea of the reliability of analog simulation, we point out that all directly measured quantities are given with a maximum error of about 5%.

3. Experiments

By now, stochastic resonance has been repeatedly observed in a large variety of experiments. The first experiment was on an electronic circuit. Fauve and Heslot (1983) used a Schmitt trigger to demonstrate the effect. A first *in situ* physical experiment by McNamara, Wiesenfeld, and Roy (1988) used a bistable ring laser to demonstrate stochastic resonance in the noise-induced switching between the two counter-propagating laser modes. Stochastic resonance has also been demonstrated optically in a semiconductor feedback laser (Iannelli *et al.*, 1994), in a unidirectional photoreactive ring resonator (Jost and Sahleh, 1996), and in optical heterodyning (Dykman, Golubev, *et al.*, 1995). Relevance of stochastic resonance in electronic paramagnetic resonance has been identified by Gammaitoni, Martinelli, Pardi, and Santucci (1991, 1993). Simon and Libchaber (1992) observed SR in a beautifully designed optical trap, where a dielectric particle moves in the field of two overlapping Gaussian laser-beams that try to pull the particle into their center. Spano and collaborators (1992) have observed stochastic resonance in a paramagnetically driven bistable buckling ribbon. Magnetostochastic resonance in ferrite-garnet films has been measured by Grigorenko *et al.* (1994) and in yttrium-iron spheres by Reibold *et al.* (1997). I and Liu (1995) observed stochastic resonance in weakly ionized magnetoplasmas, and Claes and van den Broeck (1991) for dis-

persions of particles suspended in time-periodic flows. A first demonstration of stochastic resonance in a semiconductor device, more precisely in the low-temperature ionization breakdown in p-type germanium, has been reported by Kittel *et al.* (1993). Stochastic resonance has been observed in superconducting quantum interference devices (SQUID) by Hibbs *et al.* (1995) and Rouse *et al.* (1995). Furthermore, Rouse *et al.* provided the first experimental evidence of noise-induced resonances (see Sec. VII.D.1) in their SQUID system. In recent experiments by Mantegna and Spagnolo (1994, 1995, 1996), stochastic resonance was demonstrated in yet another semiconductor device, a tunnel diode. Stochastic resonance has also been observed in modulated bistable chemical reaction dynamics (Minimal-Bromate and Belousov-Zhabotinsky reactions) by Hohmann, Müller, and Schneider (1996).

Undoubtedly, the neurophysiological experiments on stochastic resonance constitute a cornerstone in the field. They have triggered the interest of scientists from biology and biomedical engineering to medicine. Longtin, Bulsara, and Moss (1991) have demonstrated the surprising similarity between interspike interval histograms of periodically stimulated sensory neurons and residence-time distributions of periodically driven bistable systems (see Secs. II.B and IV.C). Stochastic resonance in a living system was first demonstrated by Douglass *et al.* (1993) (see also in Moss *et al.*, 1994) in the mechanoreceptor cells located in the tail fan of crayfish. A similar experiment using the sensory hair cells of a cricket was performed by Levin and Miller (1996). A cricket can detect an approaching predator by the coherent motion of the air although the coherence is buried under a huge random background. Fairly convincing arguments had been given by Levin and Miller that stochastic resonance is actually responsible for this capability of the cricket. Since the functionality of neurons is based on gating ion channels in the cell membrane, Bezrukov and Vodyanoy (1995) have studied the impact of stochastic resonance on ion-channel gating. Stochastic resonance has been studied in visual perceptions [Riani and Simonotto, 1994, 1995; Simonotto *et al.* (1997); Chialvo and Apkarian (1993)] and in the synchronized response of neuronal assemblies to a global low-frequency field (Gluckman *et al.*, 1996).

III. TWO-STATE MODEL

In this section we discuss the simplest model that epitomizes the class of symmetric bistable systems introduced in Sec. II.A. Such a *discrete* model was proposed originally as a stochastic resonance study case by McNamara and Wiesenfeld (1989), who also pointed out that under certain restrictions it renders an accurate representation of most *continuous* bistable systems. For this reason, we discuss the two-state model in some detail. Most of the results reported below are of general validity and provide the reader with a preliminary analytical scheme on which to rely.

Let us consider a symmetric unperturbed system that switches between two discrete states $\pm x_m$ with rate W_0 out of either state. We define $n_{\pm}(t)$ to be the probabilities that the system occupies either state \pm at time t , that is $x(t) = \pm x_m$. In the presence of a periodic input signal $A(t) = A_0 \cos(\Omega t)$, which biases the state \pm alternatively, the transition probability densities $W_{\mp}(t)$ out of the states $\pm x_m$ depend periodically on time. Hence the relevant master equation for $n_{\pm}(t)$ reads

$$\dot{n}_{\pm}(t) = -W_{\mp}(t)n_{\pm} + W_{\pm}(t)n_{\mp} \quad (3.1a)$$

or, making use of the normalization condition $n_{+} + n_{-} = 1$,

$$\dot{n}_{\pm}(t) = -[W_{\pm}(t) + W_{\mp}(t)]n_{\pm} + W_{\pm}(t). \quad (3.1b)$$

The solution of the rate equation (3.1) is given by

$$n_{\pm}(t) = g(t) \left[n_{\pm}(t_0) + \int_{t_0}^t W_{\pm}(\tau) g^{-1}(\tau) d\tau \right],$$

$$g(t) = \exp \left(- \int_{t_0}^t [W_{+}(\tau) + W_{-}(\tau)] d\tau \right), \quad (3.2)$$

with unspecified initial condition $n_{\pm}(t_0)$. For the transition probability densities $W_{\mp}(t)$, McNamara and Wiesenfeld (1989) proposed to use periodically modulated escape rates of the Arrhenius type

$$W_{\mp}(t) = r_K \exp \left[\mp \frac{A_0 x_m}{D} \cos(\Omega t) \right]. \quad (3.3)$$

On assuming, as in Sec. II.A, that the modulation amplitude is small, i.e., $A_0 x_m \ll D$, we can use the following expansions in the small parameter $A_0 x_m / D$,

$$W_{\mp}(t) = r_K \left[1 \mp \frac{A_0 x_m}{D} \cos(\Omega t) + \frac{1}{2} \left(\frac{A_0 x_m}{D} \right)^2 \cos^2(\Omega t) \mp \dots \right],$$

$$W_{+}(t) + W_{-}(t) = 2r_K \left[1 + \frac{1}{2} \left(\frac{A_0 x_m}{D} \right)^2 \cos^2(\Omega t) + \dots \right]. \quad (3.4)$$

In Sec. IV.C, we discuss in more detail the validity of the expression (3.3) for the rates. The most important condition is a small driving frequency (adiabatic assumption). The integrals in Eq. (3.2) can be performed analytically to first order in $A_0 x_m / D$,

$$n_{+}(t|x_0, t_0) = 1 - n_{-}(t|x_0, t_0) = \frac{1}{2} \{ \exp[-2r_K(t-t_0)] \times [2\delta_{x_0, x_m} - 1 - \kappa(t_0)] + 1 + \kappa(t) \}, \quad (3.5)$$

where $\kappa(t) = 2r_K(A_0 x_m / D) \cos(\Omega t - \bar{\phi}) / \sqrt{4r_K^2 + \Omega^2}$, and $\bar{\phi} = \arctan[\Omega / (2r_K)]$. The quantity $n_{+}(t|x_0, t_0)$ in Eq. (3.5) should be read as the conditional probability that $x(t)$ is in the state $+$ at time t , given that its initial state is $x_0 \equiv x(t_0)$. Here the Kronecker delta δ_{x_0, x_m} is 1 when the system is initially in the state $+$.

From Eq. (3.5), any statistical quantity of the discrete process $x(t)$ can be computed to first order in $A_0 x_m / D$, namely:

(a) The time-dependent response $\langle x(t)|x_0, t_0 \rangle$ to the periodic forcing. From the definition $\langle x(t)|x_0, t_0 \rangle = \int x P(x, t|x_0, t_0) dx$ with $P(x, t|x_0, t_0) \equiv n_+(t) \delta(x - x_m) + n_-(t) \delta(x + x_m)$, it follows immediately that in the asymptotic limit $t_0 \rightarrow -\infty$,

$$\lim_{t_0 \rightarrow -\infty} \langle x(t)|x_0, t_0 \rangle \equiv \langle x(t) \rangle_{as} = \bar{x}(D) \cos[\Omega t - \bar{\phi}(D)], \quad (3.6)$$

with

$$\bar{x}(D) = \frac{A_0 x_m^2}{D} \frac{2r_K}{\sqrt{4r_K^2 + \Omega^2}} \quad (3.7a)$$

and

$$\bar{\phi}(D) = \arctan\left(\frac{\Omega}{2r_K}\right). \quad (3.7b)$$

Equation (3.7) coincides with Eq. (2.7) for $\langle x^2 \rangle_0 = x_m^2$.

(b) The autocorrelation function $\langle x(t+\tau)x(t)|x_0, t_0 \rangle$. The general definition

$$\begin{aligned} \langle x(t+\tau)x(t)|x_0, t_0 \rangle &= \int \int xy P(x, t+\tau|y, t) \\ &\quad \times P(y, t|x_0, t_0) dx dy \end{aligned} \quad (3.8)$$

greatly simplifies in the stationary limit $t_0 \rightarrow -\infty$,

$$\begin{aligned} \lim_{t_0 \rightarrow -\infty} \langle x(t+\tau)x(t)|x_0, t_0 \rangle &= \langle x(t+\tau)x(t) \rangle_{as} = x_m^2 \exp(-2r_K|\tau|) [1 - \kappa(t)^2] \\ &\quad + x_m^2 \kappa(t+\tau) \kappa(t). \end{aligned} \quad (3.9)$$

In Eq. (3.9) we can easily separate an exponentially decaying branch due to randomness and a periodically oscillating tail driven by the periodic input signal. Note that even in the stationary limit $t_0 \rightarrow -\infty$, the output-signal autocorrelation function depends on both times $t+\tau$ and t . However, in real experiments t represents the time set for the trigger in the data acquisition procedure. Typically, the averages implied by the definition of the autocorrelation function are taken over many sampling records of the signal $x(t)$, triggered at a large number of times t within one period of the forcing T_Ω . Hence, the corresponding phases of the input signal, $\theta = \Omega t + \varphi$, are uniformly distributed between 0 and 2π . This corresponds to averaging $\langle x(t+\tau)x(t) \rangle_{as}$ with respect to t uniformly over an entire forcing period, whence

$$\begin{aligned} \langle \langle x(t+\tau)x(t) \rangle \rangle &= x_m^2 \exp(-2r_K|\tau|) \left[1 - \frac{1}{2} \left(\frac{A_0 x_m}{D} \right)^2 \frac{4r_K^2}{4r_K^2 + \Omega^2} \right] \\ &\quad + \frac{x_m^2}{2} \left(\frac{A_0 x_m}{D} \right)^2 \frac{4r_K^2}{4r_K^2 + \Omega^2} \cos(\Omega \tau), \end{aligned} \quad (3.10)$$

where the outer brackets $\langle \dots \rangle$ stay for $(1/T_\Omega) \int_0^{T_\Omega} [\dots] dt$.

(c) The power spectral density $S(\omega)$. The power spectral density commonly reported in the literature is the Fourier transform of Eq. (3.10) [see Eq. (2.10)]

$$\begin{aligned} S(\omega) &= \left[1 - \frac{1}{2} \left(\frac{A_0 x_m}{D} \right)^2 \frac{4r_K^2}{4r_K^2 + \Omega^2} \right] \frac{4r_K x_m^2}{4r_K^2 + \omega^2} \\ &\quad + \frac{\pi}{2} \left(\frac{A_0 x_m}{D} \right)^2 \frac{4r_K^2 x_m^2}{4r_K^2 + \Omega^2} [\delta(\omega - \Omega) + \delta(\omega + \Omega)], \end{aligned} \quad (3.11)$$

which has the same form as the expression for $S(\omega)$ derived in Eq. (2.12). As a matter of fact, $S_N(\omega)$ is the product of the Lorentzian curve obtained with no input signal $A_0 = 0$ and a factor that depends on the forcing amplitude A_0 , but is smaller than unity. The total output power, signal plus noise, for the two-state model discussed here, is $2\pi x_m^2$, independent of the input-signal amplitude A_0 and frequency Ω . Hence the effect of the input signal is to transfer power from the broadband noise background into the delta spike(s) of the power spectral density. Finally, the SN ratio follows as

$$SNR = \pi \left(\frac{A_0 x_m}{D} \right)^2 r_K + \mathcal{O}(A_0^4). \quad (3.12)$$

To leading order in $A_0 x_m/D$, Eq. (3.12) coincides with Eq. (2.14).

The residence-time distribution $N(T)$ for the two-state model was calculated by Zhou, Moss, and Jung (1990), and by Löfstedt and Coppersmith (1994a, 1994b) within a two-state model, yielding in leading order of $A_0 x_m/D$ [cf. Sec. IV.D],

$$\begin{aligned} N(T) &= \mathcal{N}_0 [1 - (1/2)(A_0 x_m/D)^2 \cos(\Omega T)] r_K \\ &\quad \times \exp(-r_K T), \end{aligned} \quad (3.13)$$

with $\mathcal{N}_0^{-1} = 1 - (1/2)(A_0 x_m/D)^2 [1 + (\Omega/r_K)^2]$. Note that $N(T)$ exhibits the peak structure of Fig. 5, with $T_n = (n - 1/2)T_\Omega$. Furthermore, the strength P_1 of the first peak can be easily calculated by integrating $N(T)$ over an interval $[(\frac{1}{2} - \alpha), (\frac{1}{2} + \alpha)]T_\Omega$, with $0 < \alpha \leq 1/4$. Skipping the details of the integration, one realizes that P_1 is a function of the ratio Ω/r_K and attains its maximum for $r_K \approx 2\nu_\Omega$ as illustrated in the inset of Fig. 5(b).

In this section we detailed the symmetric two-state model as an archetypal system that features stochastic resonance. We profited greatly from the analytical study of McNamara and Wiesenfeld (1989). The two-state model can be regarded as an *adiabatic approximation* to any continuous bistable system, like the overdamped quartic double-well oscillator of Eqs. (2.1)–(2.3), provided that the input-signal frequency is low enough for the notion of transition rates [Eq. (3.4)] to apply.

In general, the difficulty lies in the derivation of time-dependent transition rates in a continuous model. A systematic method consists of finding the unstable periodic orbits in the absence of noise, since they act as basin boundaries in an extended phase-space description (Jung and Hänggi, 1991b). Rates in periodically driven systems can be defined as the transition rates across

those basin boundaries and correspond to the lowest-lying Floquet eigenvalue of the time-periodic Fokker-Planck operator (Jung, 1989, 1993)—see also the Sec. VII.D.3. Depending on the degree of approximation needed, the intrawell dynamics may become significant and more sophisticated formalisms may be required.

IV. CONTINUOUS BISTABLE SYSTEMS

A two-state description of stochastic resonance is of limited use for a number of reasons. First, the dynamics is reduced to the switching mechanism between two metastable states only. Thus it neglects the short-time dynamics that takes place within the immediate neighborhood of the metastable states themselves. Moreover, our goal is to describe both the linear as well as the nonlinear stochastic resonance response in the whole regime of modulation frequencies, extending from exponentially small Kramers rates up to intrawell frequencies, and higher. Put differently, a more elaborate approach has to model the nonadiabatic regime of driving in the whole accessible state space of the dynamical process $x(t)$. This goal will be presented within the class of continuous-state random systems (Stratonovitch, 1963; Hänggi and Thomas, 1982; Risken, 1984; van Kampen, 1992), which can be modeled in terms of a Fokker-Planck equation.

A. Fokker-Planck description

As a generic system modeling stochastic resonance we shall consider a Brownian particle of mass m that moves in a bistable potential $V(x)$ and is subjected to thermal noise $\xi(t)$ of the Nyquist type at temperature T . Moreover, we perturb the particle with a periodically varying force, i.e., we start from the Langevin equation

$$m\ddot{x} = -m\gamma\dot{x} - V'(x) + mA_0 \cos(\Omega t + \varphi) + \sqrt{2m\gamma kT}\xi(t). \quad (4.1)$$

Here $\xi(t)$ denotes a Gaussian white noise with zero average and autocorrelation function $\langle \xi(t)\xi(s) \rangle = \delta(t-s)$. The external forcing term is characterized by an amplitude A_0 , an angular frequency Ω , and an arbitrary but fixed initial phase φ . The statistically equivalent description for the corresponding probability density $p(x, v = \dot{x}, t; \varphi)$ is governed by the two-dimensional Fokker-Planck equation

$$\begin{aligned} \frac{\partial}{\partial t} p(x, v, t; \varphi) = & \left\{ -\frac{\partial}{\partial x} v + \frac{\partial}{\partial v} \right. \\ & \times [\gamma v + f(x) - A_0 \cos(\Omega t + \varphi)] \\ & \left. + \gamma D \frac{\partial^2}{\partial v^2} \right\} p(x, v, t; \varphi), \end{aligned} \quad (4.2)$$

where we introduced $f(x) = -V'(x)/m$ and the diffusion strength $D = kT/m$. For large values of the friction coefficient γ we can simplify the above inertial dynamics through adiabatic elimination of the velocity variable $\dot{x} = v$ (Marchesoni and Grigolini, 1983; Risken, 1984;

Grigolini and Marchesoni, 1985) to arrive at the periodically modulated Langevin equation

$$\gamma\dot{x} = f(x) + A_0 \cos(\Omega t + \varphi) + \sqrt{2\gamma D}\xi(t). \quad (4.3)$$

With the choice $f(x) = (ax - bx^3)/m$, where $a > 0$, $b > 0$, we recover the bistable quartic double-well potential $V(x) = -(1/2)ax^2 + (1/4)bx^4$ of Fig. 1. On making use of the rescaled variables:

$$\begin{aligned} \bar{x} &= x/x_m, \quad \bar{t} = at/\gamma, \quad \bar{A}_0 = A_0/ax_m, \\ \bar{D} &= D/ax_m^2, \quad \bar{\Omega} = \gamma\Omega/a, \end{aligned} \quad (4.4)$$

where $\pm x_m = \sqrt{a/b}$ locate the minima of $V(x)$, the relevant Fokker-Planck equation takes on a dimensionless form. Dropping, for the sake of convenience, all overbars one recovers the Smoluchowski limit of Eq. (4.2); i.e., in terms of an operator notation we obtain

$$\frac{\partial}{\partial t} p(x, t; \varphi) = \mathcal{L}(t)p(x, t; \varphi) \equiv [\mathcal{L}_0 + \mathcal{L}_{ext}(t)]p(x, t; \varphi). \quad (4.5)$$

Here, the Fokker Planck operator

$$\mathcal{L}_0 = -\frac{\partial}{\partial x}(x - x^3) + D\frac{\partial^2}{\partial x^2} \quad (4.6)$$

describes the unperturbed dynamics in the rescaled bistable potential

$$V(x) = -\frac{1}{2}x^2 + \frac{1}{4}x^4, \quad (4.7)$$

with barrier height $\Delta V = \frac{1}{4}$. The operator $\mathcal{L}_{ext}(t)$ denotes the gradient-type perturbation

$$\mathcal{L}_{ext}(t) = -A_0 \cos(\Omega t + \varphi) \frac{\partial}{\partial x}. \quad (4.8)$$

1. Floquet approach

The inertial, as well as the overdamped Brownian dynamics in Eqs. (4.2) and (4.5) describe a nonstationary Markovian process where the symmetry under time translation is retained in a discrete manner only. Since the Fokker-Planck operators in Eqs. (4.2) and (4.5) are invariant under the discrete time translations $t \rightarrow t + T_\Omega$, where $T_\Omega = 2\pi/\Omega$ denotes the modulation period, the Floquet theorem (Floquet, 1883; Magnus and Winkler, 1979) applies to the corresponding partial differential equation. For a general periodic Fokker-Planck operator such as $\mathcal{L}(t) = \mathcal{L}(t + T_\Omega)$, defined on the multidimensional space of state vectors $X(t) = (x(t); v(t) = \dot{x}(t); \dots)$, one finds that the relevant Floquet solutions are functions of the type

$$p(X, t; \varphi) = \exp(-\mu t) p_\mu(X, t; \varphi) \quad (4.9)$$

with Floquet eigenvalue μ and periodic Floquet modes p_μ ,

$$p_\mu(X, t; \varphi) = p_\mu(X, t + T_\Omega; \varphi). \quad (4.10)$$

The periodic Floquet modes $\{p_\mu\}$ are the (right) eigenfunctions of the Floquet operator

$$\left[\mathcal{L}(t) - \frac{\partial}{\partial t} \right] p_\mu(X, t; \varphi) = -\mu p_\mu(X, t; \varphi). \quad (4.11)$$

Here the Floquet modes $\{p_\mu\}$ are elements of the product space $L_1(X) \oplus T_\Omega$, where T_Ω is the space of functions that are periodic in time with period T_Ω , and $L_1(X)$ is the linear space of the functions that are integrable over the state space. In view of the identity

$$\begin{aligned} \exp(-\mu t) p_\mu(X, t; \varphi) &= \exp[-(\mu + ik\Omega)t] p_\mu(X, t; \varphi) \\ &\times \exp(ik\Omega t) \\ &\equiv \exp(-\hat{\mu}t) \hat{p}_\mu(X, t; \varphi), \end{aligned} \quad (4.12)$$

where $\hat{\mu} = \mu + ik\Omega$, $k=0, \pm 1, \pm 2, \dots$, and $\hat{p}_\mu(X, t; \varphi) = p_\mu(X, t; \varphi) \exp(ik\Omega t) = \hat{p}_\mu(X, t + T_\Omega; \varphi)$, we observe that the Floquet eigenvalues $\{\mu_n\}$ can be defined only mod $(i\Omega)$. Likewise, we introduce the set of Floquet modes of the adjoint operator $\mathcal{L}^\dagger(t)$, that is

$$\left[\mathcal{L}^\dagger(t) + \frac{\partial}{\partial t} \right] p_\mu^\dagger(X, t; \varphi) = -\mu p_\mu^\dagger(X, t; \varphi). \quad (4.13)$$

Here the sets $\{p_\mu\}$ and $\{p_\mu^\dagger\}$ are bi-orthogonal, obeying the equal-time normalization condition

$$\frac{1}{T_\Omega} \int_0^{T_\Omega} dt \int dX p_{\mu_n}(X, t; \varphi) p_{\mu_m}^\dagger(X, t; \varphi) = \delta_{n,m}. \quad (4.14)$$

Eqs. (4.11) and (4.13) allow for a spectral representation of the time inhomogeneous conditional probability $P(X, t|Y, s)$: With $t > s$ we find

$$\begin{aligned} P(X, t|Y, s) &= \sum_{n=0}^{\infty} p_{\mu_n}(X, t; \varphi) p_{\mu_n}^\dagger(Y, s; \varphi) \\ &\times \exp[-\mu_n(t-s)] \\ &= P(X, t + T_\Omega|Y, s + T_\Omega). \end{aligned} \quad (4.15)$$

With all real parts $\text{Re}[\mu_n] > 0$ for $n > 0$, the limit $s \rightarrow -\infty$ of Eq. (4.15) yields the ergodic, time-periodic probability

$$p_{as}(X, t; \varphi) = p_{\mu=0}(X, t; \varphi). \quad (4.16)$$

The asymptotic probability $p_{as}(X, t; \varphi)$ can be expanded into a Fourier series, i.e.,

$$p_{as}(X, t; \varphi) = \sum_{m=-\infty}^{\infty} a_m(X) \exp[im(\Omega t + \varphi)]. \quad (4.17)$$

With the arbitrary initial phase being distributed *uniformly*, i.e., with the probability density for φ given by $w(\varphi) = (2\pi)^{-1}$, the time average of Eq. (4.17) equals the phase average (Jung and Hänggi, 1989, 1990; Jung, 1993). Hence

$$\begin{aligned} \bar{p}_{as}(X) &= \frac{1}{2\pi} \int_0^{2\pi} p_{as}(X, t; \varphi) d\varphi \\ &= \frac{1}{T_\Omega} \int_0^{T_\Omega} p_{\mu=0}(X, t; \varphi) dt = a_0(X). \end{aligned} \quad (4.18)$$

At this stage it is worth pointing out a peculiarity of all

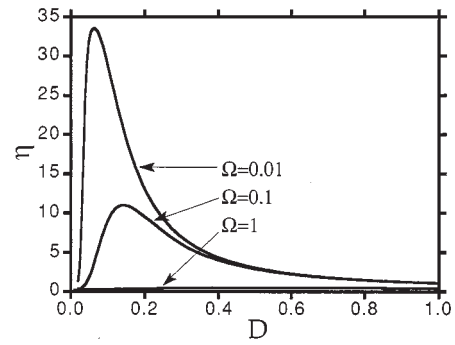


FIG. 6. The spectral amplification η for stochastic resonance in the symmetric bistable quartic double well is depicted vs the dimensionless noise strength D at a fixed modulation amplitude $A_0=0.2$ for three different values of the frequency Ω . The results were evaluated with the nonadiabatic Floquet theory for the corresponding time-periodic Fokker-Planck equation in Eq. (4.5). After Jung and Hänggi (1991a).

periodically driven stochastic systems: With $\theta = \Omega t + \varphi$ we could as well embed a periodic N -dimensional Fokker-Planck equation into a Markovian $(N+1)$ -dimensional, time-homogeneous Fokker-Planck equation by noting that $\dot{\theta} = \Omega$. With the corresponding stationary probability $p_{as}(x, \theta)$ not explicitly time dependent, an integration over θ does *not* yield the ergodic probability $p_{as}(x, t; \varphi)$ in Eq. (4.16) but rather the time-averaged result \bar{p}_{as} of Eq. (4.18), notwithstanding a claim to the contrary (Hu *et al.*, 1990).

Given the spectral representation (4.15) for the conditional probability, we can evaluate mean values and correlation functions. Of particular importance for stochastic resonance is the asymptotic expectation value

$$\langle X(t) \rangle_{as} = \langle X(t) | Y_0, t_0 \rightarrow -\infty \rangle, \quad (4.19)$$

where $\langle X(t) | Y_0, t_0 \rangle$ is the conditional average $\langle X(t) | Y_0, t_0 \rangle = \int dX X P(X, t | Y_0, t_0)$. With $P(X, t | Y_0, t_0 \rightarrow -\infty)$ approaching the asymptotic time-periodic probability, the relevant asymptotic average $\langle X(t) \rangle_{as}$ is also periodic in time and thus admits the Fourier series representation

$$\langle X(t) \rangle_{as} = \sum_{n=-\infty}^{\infty} M_n \exp[in(\Omega t + \varphi)]. \quad (4.20)$$

The complex-valued amplitudes $M_n \equiv M_n(\Omega, A_0)$ depend nonlinearly on both the forcing frequency Ω and the modulation amplitude A_0 . Within a linear-response approximation (see Sec. IV.B), only the contributions with $|n|=0, 1$ contribute to Eq. (4.20). Nonlinear contributions to the stochastic resonance observables, both for M_1 and higher-order harmonics with $|n| > 1$ have been evaluated numerically by Jung and Hänggi (1989, 1991a) by implementing the Floquet approach for the Fokker-Planck equation of the overdamped driven quartic double-well potential. The spectral amplification η of Eq. (2.9), i.e., the integrated power in the time-averaged power spectral density at $\pm \Omega$ (Jung and Hänggi, 1989, 1991a) [note also Eq. (4.24) below], is expressed in terms of $|M_1|$, i.e.,

$$\eta = \left(\frac{2|M_1|}{A_0} \right)^2. \quad (4.21)$$

Its behavior versus the noise strength D is depicted for different angular driving frequencies in Fig. 6. We observe that for a fixed modulation amplitude A_0 the stochastic resonance behavior of the spectral power amplification η decreases upon increasing the forcing frequency Ω . The behavior of η versus increasing Ω at fixed noise strength D is generally that of a monotonically decreasing function. An exception occurs in a symmetric bistable potential composed of two square wells and a square barrier with one well depth modulated periodically. For this case the SNR versus Ω has been shown to be nonmonotonic (Berdichevsky and Gitterman, 1996). The dependence on the modulation amplitude A_0 at fixed forcing frequency Ω is depicted in Fig. 7. We note that the maximum of the spectral amplification decreases with increasing amplitude A_0 . Hence nonlinear response effects tend to diminish the stochastic resonance phenomenon. For a small, fixed noise strength D (so that the driving frequency Ω exceeds the Kramers rate r_K), the spectral amplification η exhibits, however, a maximum as a function of the forcing amplitude A_0 —note the behavior in Fig. 7 below $D \sim 0.15$, and Fig. 36 in Sec. V.C.5.

The analog of the correlation function of a stationary process is the asymptotic time-inhomogeneous correlation

$$\langle X(t)X(t') \rangle_{as} = K(t, t'; \varphi) = \int \int XYP(X, t|Y, t') \times p_{as}(Y, t'; \varphi) dXdY, \quad (4.22)$$

where $t = t' + \tau$, with $\tau \geq 0$ and $t' \rightarrow \infty$. An additional averaging procedure (indicated by the double brackets) over a uniformly distributed initial phase φ for $K(t, t'; \varphi)$ (or equivalently, a time average over one modulation cycle) yields a time-homogeneous, stationary correlation function

$$\bar{K}(\tau) = \langle \langle X(t)X(t') \rangle \rangle_{as} \equiv \frac{1}{2\pi} \int K(t, t'; \varphi) d\varphi. \quad (4.23)$$

In terms of the Fourier amplitudes $\{M_n\}$ of Eq. (4.20), the long-time limit of $\bar{K}(\tau)$ assumes the oscillatory expression

$$\begin{aligned} \bar{K}(\tau) &\xrightarrow{\tau \rightarrow \infty} \equiv \bar{K}_{as}(\tau) = \langle \langle X(t+\tau) \rangle_{as} \langle X(t) \rangle_{as} \rangle \\ &= \sum_{n=-\infty}^{\infty} |M_n|^2 \exp(in\Omega\tau) \\ &= 2 \sum_{n=1}^{\infty} |M_n|^2 \cos(n\Omega\tau). \end{aligned} \quad (4.24)$$

In the last equality we used the fact that $M_0 = 0$ for a reflection-symmetric potential. Note that this asymptotic result is independent of the initial phase φ (no phase lag here!). This is in contrast with $\langle X(t) \rangle_{as}$, where the

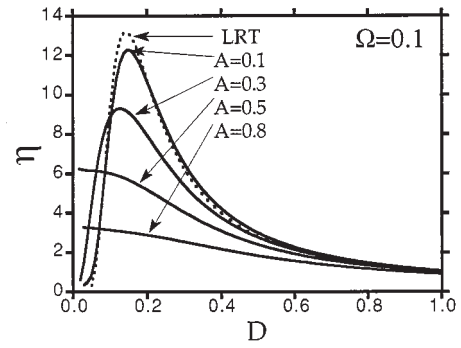


FIG. 7. The spectral amplification η versus the noise intensity D at a fixed modulation frequency $\Omega = 0.1$ is depicted for four values of the driving amplitude $A \equiv A_0$. The result of the linear response approximation in Eq. (4.51) is depicted by the dotted line. From Jung and Hänggi (1991a).

complex-valued amplitudes $\{M_n\}$ bring in an additional phase lag $\bar{\phi}_n$ for each Fourier component (see Sec. IV.B below).

This oscillatory, asymptotic long-time behavior yields in turn sharp δ spikes at multiples of the driving angular frequency Ω for the power spectral density of $\bar{K}(\tau)$. Depending on the symmetry properties of the Floquet operator, one finds that some of the amplitudes M_n assume vanishing weights (Jung and Hänggi, 1989; Hänggi *et al.*, 1993). In particular, for a symmetric double well, all even-numbered amplitudes M_{2n} assume zero weight; likewise a multiplicative driving $x A_0 \cos(\Omega t)$ in Eq. (4.2) in a symmetric double well yields identically vanishing weights for all $n = 0, \pm 1, \pm 2, \dots$.

Before we proceed by introducing the linear-response theory (LRT), we also point out that the result for the corresponding conditional probability (4.15) for zero forcing $A_0 = 0$ boils down to the time-homogeneous conditional probability density, i.e., with $\tau = (t - s) > 0$

$$P_0(X, \tau|Y, 0) = \sum_{n=0}^{\infty} \psi_n(X) \varphi_n(Y) \exp(-\lambda_n \tau). \quad (4.25)$$

Here, for $A_0 \rightarrow 0$ the set $\{\mu_n\}$ (with $k = 0$) reduces to the set of eigenvalues $\{\lambda_n\}$ of \mathcal{L}_0 , the set $\{p_{\mu_n}(X, t)\}$ reduces to the right eigenfunction $\{\psi_n\}$ of \mathcal{L}_0 , and $\{p_{\mu_n}^\dagger(Y, s)\}$ to the right eigenfunctions $\{\varphi_n(Y)\}$ of \mathcal{L}_0^\dagger , respectively.

B. Linear-response theory

As detailed in the Introduction, the prominent role of the stochastic resonance phenomenon is that it can be used to boost *weak* signals embedded in a noisy environment. Thus the linear-response concept, or more general, the concept of perturbation theory (see Appendix) for spectral quantities like the Floquet modes and the Floquet eigenvalues as discussed in the previous section are adequate methods for studying the basic physics that characterizes stochastic resonance. Both concepts have been repeatedly invoked and investigated in stochastic resonance studies by several research groups (Fox, 1989;

McNamara and Wiesenfeld, 1989; Presilla *et al.*, 1989; Dykman *et al.*, 1990a, 1990b; Hu *et al.*, 1990; Jung and Hänggi, 1991a, 1993; Dykman, Mannella, *et al.*, 1992; Hu, Haken, and Ning, 1992; Dykman, Luchinsky, *et al.*, 1995). Here we shall focus on the linear-response concept, which also emerges as a specific application of perturbation theory. In doing so, we shall rely on the linear-response theory pioneered by Kubo (1957, 1966) for equilibrium systems—and extended by Hänggi and Thomas (1982) to the wider class of stochastic processes that admit also nonthermal, stationary nonequilibrium states. This extension is of particular relevance because many prominent applications of stochastic resonance in optical, chemical, and biological systems operate *far* from thermal equilibrium. Without lack of generality, we confine the further analysis to a one-dimensional Markovian observable $x(t)$ subjected to an external weak periodic perturbation. Following Hänggi (1978) and Hänggi and Thomas (1982), the long-time limit of the response $\langle x(t) \rangle_{as}$ due to the perturbation $A(t) = A_0 \cos(\Omega t)$, i.e., we set $\varphi = 0$, assumes up to first order the form

$$\langle x(t) \rangle_{as} = \langle x(t) \rangle_0 + \int_{-\infty}^t ds \chi(t-s) A_0 \cos(\Omega s), \quad (4.26)$$

where $\langle x(t) \rangle_0$ denotes the stationary average of the unperturbed process. The memory kernel $\chi(t)$ of Eq. (4.26) is termed, hereafter, the *response function*. For an external perturbation operator of the general form

$$\mathcal{L}_{ext}(t) \equiv A_0 \cos(\Omega t) \Gamma_{ext}, \quad (4.27)$$

$\chi(t)$ is expressed by

$$\chi(t) = H(t) \int \int \int dx dy dz P_0(x, t|y, 0) x \Gamma_{ext}(y, z) p_0(z). \quad (4.28)$$

$H(t)$ denotes the (Heaviside) step function expressing causality of the response, $p_0(z)$ is the stationary probability density of the corresponding unperturbed, generally nonthermal equilibrium process, $P_0(x, t|y, 0)$ denotes the conditional probability density, and $\Gamma_{ext}(x, y)$ denotes the kernel of the operator Γ_{ext} that describes the perturbation in the master operator (either an integral operator or a differential operator such as in the Fokker-Planck case of Sec. IV.A, e.g., $\Gamma_{ext}(y, z) = \delta'(z - y)$ for Eq. (4.8)). An appealing form of the response function can be obtained by introducing the fluctuation $\zeta(x(t))$ defined by

$$\int dy \Gamma_{ext}(x, y) p_0(y) = - \int dz \mathcal{L}_0(x, z) \zeta(z) p_0(z), \quad (4.29)$$

where $\mathcal{L}_0(x, z)$ is the kernel of the unperturbed Fokker-Planck operator. We note that $\zeta(x(t))$ is indeed a fluctuation, i.e., it satisfies $\langle \zeta(x(t)) \rangle_0 = 0$. The response function (4.28) can then be expressed through the *fluctuation theorem*

$$\chi(t) = -H(t) \frac{d}{dt} \langle x(t) \zeta(x(0)) \rangle_0. \quad (4.30)$$

For $\delta x(t) = x(t) - \langle x(t) \rangle_0$ this can be recast as

$$\chi(t) = -H(t) \frac{d}{dt} \langle \delta x(t) \zeta(x(0)) \rangle_0. \quad (4.31)$$

This result is intriguing: the linear-response function can be obtained as the time derivative of a stationary, generally nonthermal correlation function between the two unperturbed fluctuations $\delta x(t)$ and $\zeta(x(t))$. From the spectral representation of the time-homogeneous conditional probability (4.25), it follows immediately that (on assuming that the eigenvalue $\lambda_0 = 0$ is not degenerate)

$$\chi(t) = H(t) \sum_{n=1}^{\infty} g_n \lambda_n \exp(-\lambda_n t). \quad (4.32)$$

The coefficients $\{g_n\}$ are given by

$$g_n = \langle \delta x \psi_n(x) \rangle_0 \langle \zeta(y) \varphi_n(y) \rangle_0. \quad (4.33)$$

The corresponding Fourier transform $\int_0^{\infty} \exp(-i\omega\tau) \chi(\tau) d\tau$ will be denoted by $\chi(\omega)$, with $\chi(\omega) = \chi'(\omega) - i\chi''(\omega)$. Generally, the eigenvalues of the real-valued operator \mathcal{L}_0 are complex valued and occur by the pair, e.g., λ_n and λ_n^* with the corresponding eigenfunctions $\psi_n(x)$ and $\phi_n(x)$ introduced above. Hence the contribution of each pair of complex conjugate eigenvalues is a real-valued quantity, thus yielding an overall real expression for $\chi(t)$ in Eq. (4.32). Upon substituting Eq. (4.32) into Eq. (4.26) we find the linear-response approximation

$$\begin{aligned} \langle \delta x(t) \rangle &= \langle x(t) \rangle_{as} - \langle x(t) \rangle_0 = \frac{A_0}{2} \sum_{n=1}^{\infty} \lambda_n g_n \\ &\times \left[\frac{e^{i\Omega t}}{\lambda_n + i\Omega} + \frac{e^{-i\Omega t}}{\lambda_n - i\Omega} \right]. \end{aligned} \quad (4.34)$$

Moreover, on Fourier transforming Eq. (4.32) we derive the spectral representation of $\chi(t)$, i.e.,

$$\chi(\omega) = \chi'(\omega) - i\chi''(\omega) = \sum_{n=1}^{\infty} \frac{\lambda_n g_n}{\lambda_n + i\omega}. \quad (4.35)$$

Therefore, the result of Eq. (4.34) can be cast into the form

$$\langle \delta x(t) \rangle = 2|M_1| \cos(\Omega t - \bar{\phi}), \quad (4.36)$$

where the spectral amplitude $|M_1|$ is given by

$$|M_1| = \frac{A_0}{2} |\chi(\Omega)|, \quad (4.37)$$

and the retarded positive phase shift $\bar{\phi}$ reads

$$\bar{\phi} = \arctan \left[\frac{\chi''(\Omega)}{\chi'(\Omega)} \right]. \quad (4.38)$$

The above results are valid for a general nonthermal stationary system. The fluctuation $\zeta(x(t))$ can be evaluated in a straightforward manner for all one-dimensional systems modeled by a Fokker-Planck equation. Examples include the stochastic resonance for absorptive optical bistability, Sec. V.A.3, or that for colored noise-driven bistable systems in Sec. VI.D.

For the case of the quartic double-well potential [Eqs. (2.1)–(2.3)], where the unmodulated system admits thermal equilibrium, the perturbation operator $\mathcal{L}_{ext}(t)$ is of the gradient type: from Eq. (4.8), $\mathcal{L}_{ext}(t) = A_0 \cos(\Omega t) [-\partial/\partial x]$. This, in turn, implies that the response function obeys the well-known fluctuation-dissipation theorem known from classical equilibrium statistical mechanics (Kubo, 1957, 1966), i.e.,

$$\chi(t) = -[H(t)/D] \frac{d}{dt} \langle \delta x(t) \delta x(0) \rangle_0, \quad (4.39)$$

where the corresponding fluctuation ζ reads $\zeta(x(0)) = \delta x(0)/D$. Note that this result holds true irrespective of the detailed form of the equilibrium dynamics.

1. Intrawell versus interwell motion

Given the spectral representation (4.35) of the response function $\chi(t)$, we can express the two stochastic resonance quantifiers, namely the spectral amplification η of Eqs. (2.9), Eq. (4.21), and the signal-to-noise ratio of Eq. (2.13) in terms of the spectral amplitude $|M_1|$. From Eq. (4.36) we find for the spectral amplification within linear response

$$\eta = (2|M_1|/A_0)^2 = |\chi(\Omega)|^2. \quad (4.40)$$

In view of the unperturbed power spectral density $S_N^0(\Omega)$ of the fluctuations $\delta x(t)$, i.e.,

$$S_N^0(\omega) = \int_{-\infty}^{\infty} e^{-i\omega\tau} \langle \delta x(\tau) \delta x(0) \rangle_0 d\tau, \quad (4.41)$$

the linear response result for the SNR reads

$$SNR = 4\pi |M_1|^2 / S_N^0(\Omega) = \pi A_0^2 |\chi(\Omega)|^2 / S_N^0(\Omega). \quad (4.42)$$

Both stochastic resonance observables possess a spectral representation via the spectral representations of $|\chi(\omega)|$ and $S_N^0(\omega)$.

In the following we shall explicitly assume that the noise strength D is weak. This implies that for a general bistable dynamics there exists a clear-cut separation of time scales. These are the escape time scale to leave the corresponding wells, i.e., the exponentially large time scale for interwell hopping, and the time scale that characterizes local relaxation within a metastable state. The eigenvalue λ_1 that characterizes the intrawell dynamics is always real valued and of the Kramers type (Hänggi *et al.*, 1990), i.e.,

$$\lambda_1 = 2r_K = r_+ + r_- \equiv \lambda, \quad (4.43)$$

where r_{\pm} are the forward and backward transition rates, respectively. The rates r_{\pm} depend through the Arrhenius factor on the activation energies $\Delta\Phi_0^{\pm}$, where $\Phi_0(x)$ is the generalized (non-thermal-equilibrium) potential associated with the unperturbed stationary probability density

$$p_0(x) = Z^{-1}(x) \exp(-\Phi_0(x)/D). \quad (4.44)$$

The relevant intrawell relaxation rates in the two wells located at $x = x_{1,2}^0$ are estimated as the real part of the two smallest eigenvalues that describe the equilibration of the probability density in the vicinity of the two stable states x_m , $m=1,2$, respectively. For small noise intensities, these eigenvalues can be approximated as

$$\lambda_2 = \Phi_0''(x = x_1) \quad (4.45)$$

and

$$\lambda_3 = \Phi_0''(x = x_2). \quad (4.46)$$

Note that, here, the indices of λ_2 and λ_3 have been chosen for later convenience and do not necessarily coincide with the index ordering of the Fokker-Planck spectrum $\{\lambda_n\}$. Given these three dominant time scales, the response at weak noise is cast as the sum of three terms, i.e., for a driving phase $\varphi=0$ we have the weak-noise approximation

$$\langle \delta x(t) \rangle = \frac{A_0}{2} \sum_{m=1,2,3} \lambda_m g_m \left[\frac{e^{i\Omega t}}{\lambda_m + i\Omega} + \frac{e^{-i\Omega t}}{\lambda_m - i\Omega} \right], \quad (4.47)$$

yielding corresponding estimates for $\chi(\Omega)$ and the stochastic resonance quantifiers η and the SNR. The weights g_m can be evaluated from the corresponding approximate eigenfunctions (Hänggi and Thomas, 1982; Dykman, Haken, *et al.*, 1993), or from a three-term exponential ansatz for the response function (Jung, 1993).

For the overdamped, symmetric quartic double-well dynamics [Eq. (4.7)], the spectral amplification given by Eq. (4.40) has been evaluated in the literature by means of Eq. (4.37) to give (Jung and Hänggi, 1991a, 1993)

$$\eta = D^{-2} \left[\frac{4g_1^2 r_K^2}{4r_K^2 + \Omega^2} + \frac{g^2 \alpha^2}{\alpha^2 + \Omega^2} + \frac{4g_1 g \alpha r_K (2\alpha r_K + \Omega^2)}{(4r_K^2 + \Omega^2)(\alpha^2 + \Omega^2)} \right]. \quad (4.48)$$

where $\lambda_2 = \lambda_3 \equiv \alpha$, with $\alpha=2$, and $g_2 = g_3 \equiv g/2$. The relevant weights g_n for $D \rightarrow 0$ read

$$g_1 \approx 1 - (1 + \alpha^{-1})D + O(D^2),$$

$$g = D/\alpha + O(D^2), \quad (4.49)$$

and r_K is the steepest-descent approximation for the Kramers rate

$$r_K = (\sqrt{2}\pi)^{-1} \exp[-1/(4D)]. \quad (4.50)$$

Upon neglecting the intrawell motion, the leading-order contribution in Eq. (4.48) reproduces Eq. (2.7a), i.e.,

$$\eta \approx \frac{1}{D^2} \left[1 + \frac{\pi^2}{2} \Omega^2 \exp\left(\frac{1}{2D}\right) \right]^{-1}. \quad (4.51)$$

This approximation exhibits the typical bell-shaped sto-

chastic resonance behavior as a function of increasing noise intensity D —see again Fig. 7, and also Figs. 18 and 19 below. Likewise, we can evaluate the SNR for the potential under study. In the weak-noise limit we have

$$S_N^0(\Omega) \approx \frac{4r_K}{4r_K^2 + \Omega^2} + \frac{2g\lambda_2}{\lambda_2^2 + \Omega^2}, \quad (4.52)$$

whence yielding the linear-response result for the SNR (Hu *et al.*, 1992; Jung, 1993), i.e.,

$$SNR = \frac{\pi A_0^2}{2D^2} \frac{4g_1^2 r_K^2 (\alpha^2 + \Omega^2) + (g\alpha)^2 (4r_K^2 + \Omega^2) + 4\alpha g_1 r_K (2\alpha r_K + \Omega^2)}{2g_1 r_K (\alpha^2 + \Omega^2) + g\alpha (4r_K^2 + \Omega^2)}. \quad (4.53)$$

This result, when plotted vs D displays a bell-shaped behavior for Ω not too large (see Fig. 8). Moreover, note that the result for the SNR diverges as $D \rightarrow 0$, proportional to D^{-1} . This is due to the intrawell contributions in Eq. (4.53). This feature is in agreement with simulations (McNamara and Wiesenfeld, 1989). On neglecting intrawell contributions, i.e., by setting $g_2 = g_3 = 0$, one finds in leading order the result of Eq. (2.14) (Gammaitoni, Marchesoni, *et al.*, 1989; McNamara and Wiesenfeld, 1989; Presilla *et al.*, 1989; Dykman *et al.*, 1990b), i.e.,

$$SNR = (\pi A_0^2 / D^2) r_K = [A_0^2 / (\sqrt{2} D^2)] \exp[-1/(4D)]. \quad (4.54)$$

We remark that within this interwell approximation the SNR —contrary to the spectral amplification η in Eq. (4.51)—is no longer dependent on the angular modulation frequency Ω ! This effective two-state approximation also exhibits a bell-shaped behavior, typical for stochastic resonance. In contrast to Eq. (4.53) the SN ratio vanishes for $D \rightarrow 0$. It is also worthwhile to point out the difference in how r_K enters the two stochastic resonance observables. The leading order contribution to SNR in Eq. (4.54) is proportional to r_K , while η in Eq. (4.51) is proportional to r_K^2 .

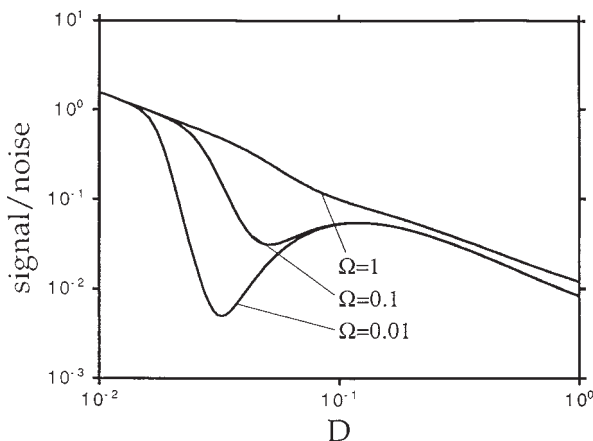


FIG. 8. The signal-to-noise ratio as function of the noise strength D for $A_0 = 0.1$ and different driving frequencies Ω . In contrast to the spectral amplification η —see Fig. 7—we note that the signal-to-noise ratio diverges as the noise strength $D \rightarrow 0$.

2. Role of asymmetry

In this subsection we study the effect of a potential asymmetry on stochastic resonance detectability. Here, the asymmetry of $\Phi_0(x)$ is characterized by the difference $\epsilon = \Delta\Phi_0^- - \Delta\Phi_0^+$ between the Arrhenius energies for backward and forward transitions. We shall assume that $\epsilon > 0$; thus the backward rate r_- is exponentially suppressed over the forward escape rate. Such an asymmetry implies also an exponential suppression of the corresponding weight $g_1 \sim 1 \rightarrow \exp(-\epsilon/D)$ [see Eq. (6.3.46) of Hänggi and Thomas, 1982]. As a consequence, the spectral amplification suffers an exponential suppression proportional to $[\exp(-\epsilon/D)]^2 = \exp(-2\epsilon/D)$, while the suppression of the SN ratio is weaker, being proportional to $\exp(-\epsilon/D)$.

On inspecting the leading order results in Eq. (4.51) for η and Eq. (4.54) for SNR , we note that the stochastic resonance maximum is located in the neighborhood where the monotonic decreasing function $y_1 = D^{-2}$ crosses the monotonic, exponentially increasing Arrhenius factor $y_2 = \exp(-\Delta\Phi_0/D)$ for the symmetric barrier with $\Delta\Phi_0^+ = \Delta\Phi_0^- = \Delta\Phi_0$. The suppression caused by asymmetry now modifies y_2 into $\exp(-\epsilon/D)y_2$. Hence the intersection point of y_1 and y_2 as functions of D is moved to larger noise intensities. Both the exponential decrease (induced by the asymmetry in activation barriers in the unperturbed potential) of the peak for η (and likewise for the SNR), as well as the shift to larger noise intensities of the peak position have been confirmed numerically for a nonequilibrium optical bistable system (Bartussek, Jung, and Hänggi, 1993; Bartussek, Hänggi, and Jung, 1994) (see also Sec. V.A.3) and again numerically in an asymmetric rf SQUID loop by Bulsara, Inchiosa, and Gammaitoni (1996). The detailed analysis for an asymmetric quartic bistable well with asymmetry energy ϵ , but identical curvatures, gives for the spectral amplification (Jung and Bartussek, 1996; Grifoni *et al.*, 1996)

$$\eta = \frac{1}{D^2} \left[\cosh^4 \left(\frac{\epsilon}{2D} \right) \left(1 + \frac{\Omega^2}{4r_K^2(\epsilon)} \right) \right]^{-1}, \quad (4.55)$$

with $r_K(\epsilon) = r_K \cosh(\epsilon/2D)$, while the corresponding result for the SNR in the presence of an asymmetry ϵ reads

$$SNR(\epsilon) = \frac{\pi A_0^2}{D^2} \frac{1}{\cosh^2(\epsilon/2D)} r_K(\epsilon). \quad (4.56)$$

An important feature of Eq. (4.55) is its universal shape for vanishing driving frequencies: In contrast to the symmetric case $\epsilon=0$, where the maximum of the spectral amplification increases with decreasing driving frequencies Ω (see Fig. 6 and Fig. 18 below for the optical bistability) the spectral amplification in asymmetric systems (see Fig. 19 below) approaches for $\Omega \rightarrow 0$ a limiting curve, with the stochastic resonance maximum assumed at a finite noise level. As a result, there exists no obvious time-scale matching condition in asymmetric systems.

3. Phase lag

The asymptotic probability $p_{as}(x,t;\varphi)$ [see Eq. (4.17)] depends periodically on the modulation phase $\theta = \Omega t + \varphi$. Moreover, due to the complex-valued amplitudes $a_m(x)$, the contribution to $p_{as}(x,t;\varphi)$ stemming from the pair of $\pm m$ introduces each its own additional phase lag $\bar{\phi}_m$. For periodically driven (linear) Gauss-Markov processes, only the terms with $m = \pm 1$ and $m = 0$ contribute to Eq. (4.17). The corresponding phase lag $\bar{\phi}_1$ for the asymptotic probability of a Brownian harmonic oscillator has been evaluated explicitly as a function of the friction coefficient γ by Jung and Hänggi (1990). Analogously, in Eq. (4.20) each nonlinear contribution to $\langle x(t) \rangle_{as}$ with power amplitude M_n introduces its own phase lag $\bar{\phi}_m$.

From the linear-response function $\chi(\Omega)$ we obtain the phase lag $\bar{\phi} \equiv \bar{\phi}_1$ of Eq. (4.38). Correspondingly, the linear-response approximation for $p_{as}(x,t;\varphi)$ in Eq. (4.17) yields such a phase lag in terms of the amplitude M_1 of Eqs. (4.20) and (4.37). Neglecting all the intrawell terms g_n with $n \geq 2$ in Eq. (4.32) gives the single-exponential approximation (2.7b) for the phase lag $\bar{\phi}$ (Nicolis, 1982; McNamara and Wiesenfeld, 1989; Gammaitoni, Marchesoni, *et al.*, 1991), i.e.,

$$\bar{\phi} = \arctan\left(\frac{\Omega}{2r_K}\right). \quad (4.57)$$

$\bar{\phi}$ decreases *monotonically* from $\pi/2$ at $D=0^+$ to zero (for $\Omega \rightarrow 0$) as D is made to grow to infinity. Note that this is the equivalent of the two-state approximation of Sec. III. The inclusion of intrawell terms (Dykman, Mannella, *et al.*, 1992; Gammaitoni and Marchesoni, 1993; Dykman, Mannella, McClintock, and Stocks, 1993) changes this monotonic behavior into a bell-shaped behavior, as long as the modulation amplitude A_0 remains small. This feature is consistent with the linear-response result [Eq. (4.47)], which accounts for the hopping term with rate r_K and the two intrawell terms with rates λ_2 and λ_3 . In particular, in a symmetric bistable potential with $D \rightarrow 0$, $\bar{\phi}$ approaches $\arctan(\Omega/\alpha)$ —see the definition of α below Eq. (4.48). The presence of the intrawell dynamics suppresses at low noise the influence of the interwell dynamics on the phase lag. Hence within the regime of validity of the linear-response approximation, the peak of the phase shift marks the turnover between the regimes dominated by hopping and intrawell motion. Note that this maximum is not physically related to

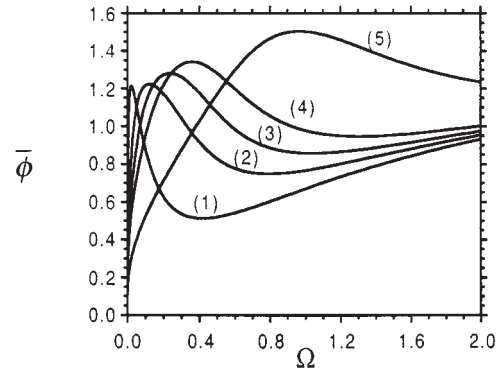


FIG. 9. Phase shift $\bar{\phi}$ of stochastic resonance in a periodically driven, overdamped quartic double well for a dimensionless noise strength $D=0.05$ vs driving frequency Ω for increasing driving amplitudes (1) $A_0=0.1$, (2) $A_0=0.3$, (3) $A_0=0.4$, (4) $A_0=0.5$, and (5) $A_0=1$. The lines are evaluated within a full nonadiabatic Floquet approach for the overdamped, time-periodic Fokker-Planck equation; see Jung and Hänggi (1993).

the maximum that characterizes stochastic resonance. Put differently, the noise value for the maximum of η , or SNR , is in no immediate relationship with the noise value that characterizes the maximum of $\bar{\phi}$: The stochastic resonance phenomenon is rooted in a physical *synchronization* effect between the interwell time scale and the period of the modulation signal, which acts here as an external “clock” (Jung and Hänggi, 1991a, 1993; Fox and Lu, 1993; Gammaitoni, Marchesoni, and Santucci, 1995). In contrast, a peak in $\bar{\phi}$ vs noise intensity D at small amplitude A_0 and small angular frequency Ω is due to the *competition* between hopping and intrawell dynamics—and should not be mistaken for a signature of stochastic resonance. The peak behavior of $\bar{\phi}$ vanishes with increasing modulation amplitude A_0 (Jung and Hänggi, 1993), where no clear-cut time-scale separation for hopping versus intrawell motion occurs. Such a dependence on the modulation amplitude is depicted in Fig. 9 as a function of the angular driving frequency Ω . The characteristic dependence of $\bar{\phi}$ on the driving strength A_0 (Ω held constant) clearly lies beyond the regime of validity of linear-response theory (Jung and Hänggi, 1993; Gómez-Ordóñez and Morillo, 1994).

C. Residence-time distributions

The residence-time distribution offers another possibility to characterize stochastic resonance. Historically, residence-time distributions were first employed in the stochastic resonance literature by Gammaitoni, Marchesoni, Menichella-Saetta, and Santucci (1989) and Zhou and Moss (1990) and then interpreted theoretically by Zhou, Moss, and Jung (1990).

Residence-time distributions were mentioned briefly in Sec. II. In this section, we discuss in more detail how these residence-time distributions can be obtained approximately and how stochastic resonance is manifested in their properties.

In the absence of periodic forcing, an escape-time distribution is defined as the distribution of times it takes for the system to escape out of a potential well. For weak noise, such a distribution is independent of the initial point apart from a small boundary layer around the basin boundary. In periodically driven systems, escape-time distributions depend additionally on the initial phase of the periodic forcing—they represent therefore a conditional escape-time distribution.

Residence-time distributions are defined as the distribution of time intervals Δ_n between two consecutive escape events, regardless of the phases of the periodic forcing $\phi = \Omega t$ at which these switching events occur. Each time interval Δ_n corresponds to a different switching phase ϕ_n which in turn depends on the prehistory of the process. The approximation strategy of Zhou, Moss, and Jung (1990) to obtain the residence-time distributions is to first compute the conditional escape-time distribution and then to average over the distribution function of the switching phases ϕ : i.e., the temporal modulation of the potential must be slow.

Although we refer to the quartic double-well potential [Eq. (2.2)] throughout, the conclusions we arrive at are of general validity since only hopping between the stable states is taken into account—relaxational motion within the potential wells is neglected. We therefore essentially resort to a two-state description, discussed in Sec. II. Furthermore, the subsequent analysis holds true only for low forcing frequencies according to the adiabatic approximation; this means that the temporal change of the adiabatic potential [Eq. (4.59) below] has to be slow in comparison to the intrawell relaxation. Such a restriction is needed here for the definition of the interwell transition rates $r_{\pm}(t)$ to make sense. On the other hand, it is clear (see below) that under such a circumstance, the two-state model is a good approximation to the continuous dynamics of a bistable process.

A detailed calculation of both the escape- and residence-time distributions has been reported in a recent paper by Choi, Fox, and Jung (1998). To help the reader interpret the results in Figs. 4 and 5, without bogging down in complicated algebraic manipulations, we outline here the simplified approach developed earlier by Zhou, Moss, and Jung (1990), with the caution that its validity is restricted to relatively large values of Ω . The starting point for the calculation of the conditional escape-time distribution $\rho_e(t)$ out of the left potential well is the instantaneous rate equation for the population in the left well $n_-(t, \phi)$, i.e.,

$$\dot{n}_-(t; \phi) = -r_+(\Omega t + \phi)n_-(t; \phi). \quad (4.58)$$

The quasistationary forward rate $r_+(\Omega t + \phi)$ denotes the adiabatic transition rate, obtained for a frozen potential

$$V_{ad}(x, t) = x^4/4 - x^2/2 - A_0 x \cos(\Omega t + \phi). \quad (4.59)$$

Initially, the system is in the left well, yielding the initial condition $n_-(0, \phi) = 1$. The quasistationary rate $r_+(\Omega t + \phi)$ can be obtained from the weakly driven

double-well potential upon combining the Kramers approach with the adiabatic assumption of a slow potential $V_{ad}(x, t)$ (Jung, 1989), i.e.,

$$r_+(\Omega t + \phi) = \frac{1}{2\pi} \sqrt{|V''_{ad}(x_b)|V''_{ad}(x_m)} \times \exp\left[-\frac{\Delta V_-(\Omega t + \phi)}{D}\right], \quad (4.60)$$

where the barrier height $\Delta V_-(\Omega t + \phi)$ and the curvatures of the adiabatic potential at the barrier top $V''(x_b)$ and in the left potential minimum $V''(x_m)$ can be obtained for small A_0 to give

$$r_+(\Omega t + \phi) \approx r_K \left[1 - \frac{3}{4}A_0 \cos(\Omega t + \phi)\right] \times \exp\left[\frac{A_0}{D} \cos(\Omega t + \phi)\right]. \quad (4.61)$$

The escape rate of the undriven system r_K is given in Eq. (2.4). As mentioned above, the applicability of Eq. (4.61) is restricted to the adiabatic regime, i.e., the frequency Ω has to be small compared to the local relaxation rate. In our scaled units this means $\Omega \ll 2$, as well as weak forcing, i.e., $A_0 \ll \sqrt{4/27}$. The two-state model rates $W_{\pm}(t)$ of Eq. (3.3) can thus be recovered from this adiabatic theory when prefactor corrections due to the forcing are neglected.

On coming back to Eq. (4.58), we note that the conditional escape-time distribution $\rho_e(t; \phi)$ can be written as

$$\rho_e(t; \phi) = -\dot{n}_-(t; \phi) = r_+(\Omega t + \phi) \times \exp\left[-\frac{1}{\Omega} \int_0^{\Omega t} r_+(\theta + \phi) d\theta\right]. \quad (4.62)$$

In order to obtain the residence-time distribution we have to find an expression for the distribution of the jump phase out of the left well $Y_-(\phi)$. It should be noted that this problem has not yet been solved systematically. For driving frequencies much smaller than the Kramers rate, there is no preferred phase and thus $Y(\phi) = 1/(2\pi)$. For larger frequencies, the following self-consistent approximation has been proposed:

$$Y_{\pm}(\phi) = \frac{1}{2\pi I_0(A_0/D)} \exp\left(\pm \frac{A_0}{D} \cos(\phi)\right), \quad (4.63)$$

with $I_0(x)$ being a modified Bessel function (Abramowitz and Stegun, 1965). The residence-time distribution in the symmetric bistable potential of Eq. (2.2) thus reads

$$N(T) = \langle N(t; \phi) \rangle = \int_0^{2\pi} Y_-(\phi) r_+(\Omega T + \phi) \times \exp\left[-\frac{1}{\Omega} \int_0^{\Omega T} r_+(\theta + \phi) d\theta\right] d\phi. \quad (4.64)$$

Performing a series expansion for small A_0/D in Eq. (4.64) we find after some algebra that

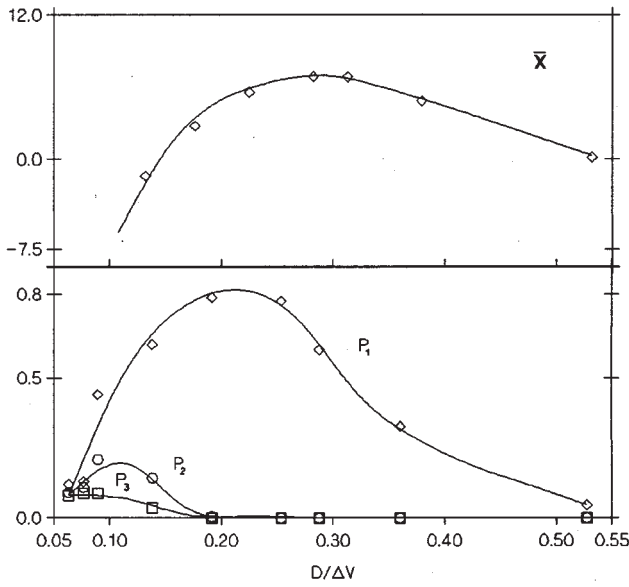


FIG. 10. Observable \bar{x} (arbitrary logarithmic scale) and height of the n th peak P_n with $n=1,2,3$ vs D for $\nu_\Omega=40$ Hz and $\alpha=\frac{1}{4}$. Data obtained by means of analog simulation of system of Eqs. (4.3)–(4.7) with $A_0x_m=0.5\Delta V$ and $a=3.2\times 10^4$ s $^{-1}$. After Gammaitoni, Marchesoni, and Santucci (1995).

$$N(T) = \mathcal{N}_0 \left[1 - \frac{1}{2} \left(\frac{A_0}{D} \right)^2 \cos(\Omega T) \right] r_K e^{-r_K T}, \quad (4.65)$$

with $\mathcal{N}_0^{-1} = 1 - \frac{1}{2} (A_0 x_m / D)^2 / [1 + (\Omega / r_K)^2]$. To indicate the dimensional dependence we use $1 = x_m$ throughout the remaining part of this section. The opposite limit, namely $\Delta V \gg A_0 x_m \gg D$, is reported here for completeness, though it will be used only in Sec. IV.D. The saddle-point approximation for the integral in Eq. (4.64) yields

$$N(T) \approx \exp[-(A_0/D)\cos(\Omega T)] \times \exp\{-(r_{\max}/2\Omega)(2\pi D/A_0)^{1/2}\} \times [2n+1 + \text{erf}(\sqrt{A_0/2D}(\Omega T - \pi))], \quad (4.66)$$

where $\text{erf}(x)$ denotes the error function (Abramowitz and Stegun, 1965), $r_{\max} = r_K \exp(A_0/D)$ and $\Omega T = \text{mod}(\Omega T, 2\pi)$. Both limiting expressions (4.65) and (4.66) for $N(T)$ exhibit a series of peaks centered at $T_n = (2n-1)(T_\Omega/2)$ (see Fig. 5). The location of the first peak is due to the fact that the clock was triggered at $T=0$, immediately after the system had crossed the barrier at $x=0$ and half a forcing period before the relevant escape barrier attained a minimum for the first time. The n th peak corresponds to the event that the system switches first after $n-1$ entire periods. Such “wait loops” do not correspond to any subharmonic component in the process power spectral distribution as pointed out in Sec. II.B.

We are now in a position to discuss the synchronization mechanism that occurs in the bistable potential of Eq. (2.2) subjected to a periodic driving. In order to quantify the strength of the n th peak of $N(T)$, we in-

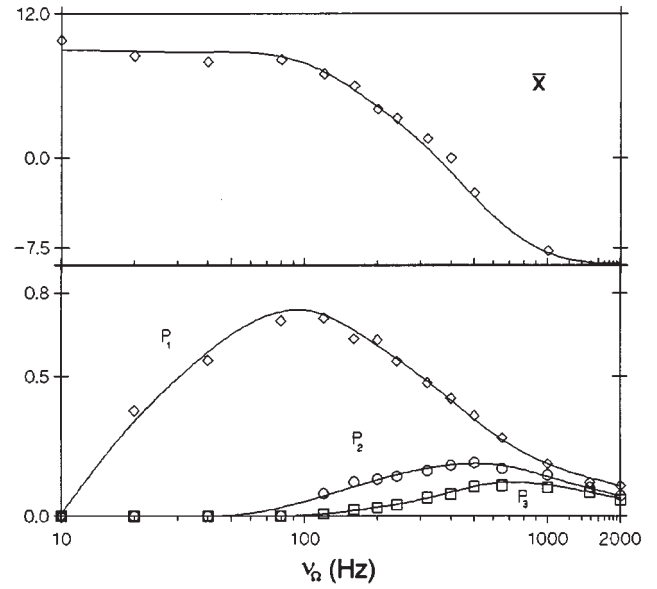


FIG. 11. Observable \bar{x} (arbitrary logarithmic scale) and height of the n th peak P_n with $n=1,2,3$ vs ν_Ω for $D=0.3\Delta V$ and $\alpha=1/4$. Other simulation circuit parameters are as in Fig. 10. An independent measurement yielded $\mu_K = 1.8 \times 10^{-2}a$. After Gammaitoni, Marchesoni, and Santucci (1995).

roduce the areas under the peaks

$$P_n = \int_{T_n - \alpha T_\Omega}^{T_n + \alpha T_\Omega} N(T) dT, \quad (4.67)$$

with $n=1,2,\dots$ and $0 < \alpha \leq \frac{1}{4}$. The actual value of the parameter α is immaterial for the behavior of P_1 . Let us focus now on the distribution of Eq. (4.65). In the regime of validity of Eq. (4.65), i.e., $r_K < \Omega \leq 2$, the background of the distribution $N(T)$ is negligible. The strength P_n of the n th peak is thus a function of the ratio r_K/Ω alone. As a consequence, P_n attains its maximum by setting either the forcing frequency ν_Ω to

$$\nu_n \approx (2n-1)r_K/2, \quad (4.68)$$

or tuning the noise, with constant ν_Ω , according to Eq. (4.68). The picture of SR as a “resonant” synchronization mechanism is thus fully established (Gammaitoni, Marchesoni, and Santucci, 1995). However, the reader should keep in mind that although the weight P_1 exhibits a maximum as a function of the frequency, the underlying mechanism is not a resonance in the sense of dynamical systems. While a dynamical resonance is due to the interaction of two degrees of freedom when their time scales agree, the nature of the peak of P_1 as a function of Ω is merely due to the coincidence of two time scales. For the sake of comparison, in Figs. 10 and 11 we show the dependence of P_n on D and Ω , as produced by means of an analog simulation. The basic properties of the synchronization mechanism are clearly visible: (i) For D tending to zero at fixed Ω (Fig. 10), all P_n approach a constant value independent on n , as expected in the nonadiabatic weak-noise limit of Sec. IV.D. (ii) Each curve $P_n = P_n(D)$ passes through a

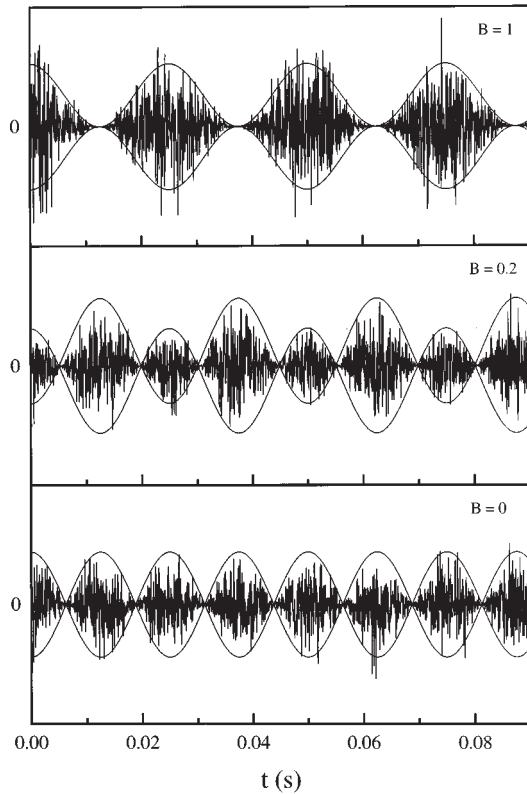


FIG. 12. Digitized time series for $f(t)$ with $\nu_\Omega = 40$ Hz, $A_0 = 1$, and different values of B . The envelope function $\pm |A_0 \cos(\Omega t) + B|$ is drawn for convenience. Ordinate units are arbitrary (Gammaitoni, Marchesoni, and Santucci, 1994).

maximum, the position D_n of which shifts progressively towards smaller values with the index n . (iii) The observables $\bar{x}(D)$ and P_1 peak at different D values, in agreement with the predictions for D_{SR} , Eq. (2.8), and D_1 , Eq. (4.68). (iv) Contrary to \bar{x} , which decreases monotonically with the forcing frequency, the curves $P_n = P_n(\nu_\Omega)$ of Fig. 11 exhibit a clear-cut resonant-like profile. The positions ν_n of the maxima of P_n are apparently odd multiples of the fundamental frequency $r_K/2$; (v) Moreover, we notice that the inequality $P_n(\nu_\Omega) > P_m(\nu_\Omega)$ holds for $n < m$ in the whole range of simulated forcing frequencies. Finally, we stress that for $\Omega \sim r_K$ or smaller, condition in Eq. (4.68) may be fulfilled by the exponential background of the distribution $N(T)$ alone, even in the absence of peaks at T_n . Of course, in this limit the quantities P_n provide no characterization of the synchronization mechanism.

As an application of the residence-time analysis, we now show how stochastic resonance may occur in the absence of symmetry breaking. Following Dykman, Luchinsky, *et al.* (1992), we consider the symmetric bistable process

$$\dot{x} = -V'(x) + f(t), \quad (4.69)$$

where the potential function $V(x)$ is as in Eq. (2.2) and

$$f(t) = [A_0 \cos(\Omega t + \varphi) + B] \xi(t), \quad (4.70)$$

with $A_0, B \geq 0$. On setting $\varphi = 0$ for convenience, the autocorrelation function of the noise reads

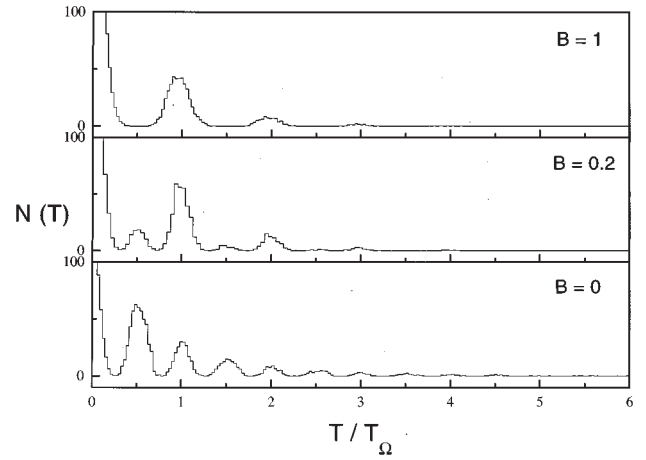


FIG. 13. Simulated residence-time distribution $N(T)$ for the system of Eqs. (4.69) and (4.70) with $\nu_\Omega = 40$ Hz, $A_0 = 1$, and different values of B . Other circuitual parameters are $a = 2 \times 10^4$ s $^{-1}$, $x_m = 1$, and $\bar{D} = 0.16 \Delta V$. After Gammaitoni, Marchesoni, and Santucci (1994).

$$\langle f(t)f(t') \rangle = 2D \delta(t-t') [A_0 \cos(\Omega t) + B]^2. \quad (4.71)$$

In Fig. 12 we show examples of digitized time series for $f(t)$ in the three typical cases discussed below, i.e., for $A_0 = 1$ and $B = 0$, for $B = 0.2$ and $B = 1$, and for fixed effective noise intensity $\bar{D} = D(B^2 + A_0^2/2)$.

Contrary to the process of Eqs. (2.1)–(2.3), the process under investigation here is symmetric under parity transformation $x \rightarrow -x$ at any time. As a consequence, $\langle x(t; \phi) \rangle_{as} = 0$ vanishes identically and no peak is detectable in the power spectrum $S(\omega)$.

Evidence of stochastic resonance effects can be detected through a synchronization-based analysis (Gammaitoni, Marchesoni, Menichella-Saetta, and Santucci, 1994). In Fig. 13, residence-time distributions are displayed for the same A_0 and B values as in Fig. 12. The series of exponentially decaying peaks is apparent. Moreover, the dependence of $N(T)$ on the offset parameter B [see Eq. (4.71)] is also noteworthy. As illustrated in Fig. 12 for $B \geq A_0$ the modulation period of $f(t)$ is T_Ω , whence the peaks of $N(T)$ are centered at $T_n = nT_\Omega$. For $B = 0$, instead, the $f(t)$ modulation period is half the forcing period and, accordingly, the peaks of $N(T)$ are located at $T_n = nT_\Omega/2$. In the intermediate case $0 < B < A_0$, two series of higher and lower peaks show up with maxima at the even and odd multiples of $T_\Omega/2$, respectively.

When the strength of the first peak is plotted against the noise intensity D , the curves $P_1(D)$ pass through a maximum for $D_1 = D_1(B)$. The B dependence of D_1 can be interpreted quantitatively as follows. For $A_0 \ll B$ the rate performs small oscillations about the unperturbed rate $r_K(B^2D)$, where B^2D is the effective intensity \bar{D} for $A_0 = 0$. Therefore, according to the stochastic resonance condition (4.68), the synchronization of switching events and periodic noise amplitude modulation is maximum for $\nu_\Omega = r_K(B^2D_1)/2$. For $A_0 \gg B$, the characteristic switching rate to compare with is the

Kramers rate r_K with effective noise intensity $\bar{D} = A_0^2 D_1/2$. On remembering that the period of the $f(t)$ amplitude modulation is now $T_\Omega/2$, we conclude that the relevant stochastic resonance condition is $2\nu_\Omega = r_K(A_0^2 D_1/2)/2$. Of course, such estimates for $D_1(B)$, which fit very closely the simulation results of Gammaitoni, Marchesoni, Menichella-Saetta, and Santucci (1994), could have been obtained by explicitly calculating the relevant distributions $N(T)$.

D. Weak-noise limit of stochastic resonance—power spectra

The most common approach in investigating stochastic resonance (apart from numerical solutions or analog simulations) is linear-response theory or the perturbation theory outlined in the Appendix. The condition for linear-response theory and perturbation theory to work well is that the effect of the periodic forcing can be treated as a small perturbation, i.e., $A_0 x_m \ll D$. The response of the periodic driving can then be described in terms of quantities of the unperturbed Fokker-Planck equation, such as its eigenfunctions and eigenvalues and/or some corresponding unperturbed correlation function.

In this section, we consider the complementary limit where $A_0 x_m \gg D$, i.e., linear-response theory is no longer valid. This weak noise limit, sometimes also termed nonlinear stochastic resonance limit, reveals some peculiar properties of the power spectrum that we shall discuss (Shneidman *et al.*, 1994a, 1994b; Stocks, 1995). The starting point for these investigations is, as in Sec. III, the two-state master equation Eq. (3.1) for the population dynamics. Assuming adiabatic conditions, i.e., the change of the adiabatic potential $V(x, t) = V_0(x) - A_0 x \cos(\Omega t)$ is slower than the thermal relaxation within a potential well (in our units $\Omega \ll 2$), the two stable states in the master equation are given by the local (time-dependent) minima $x_\pm(t)$ of the adiabatic potential $V(x, t)$. Fluctuations of the system within the potential wells will be neglected, or treated as a small perturbation.

We consider a situation in which the noise strength D is the smallest parameter. The driving frequency is large enough (though adiabatically slow) so that almost all escape events take place when the potential barriers assume their smallest values. Under these conditions, the transition probability densities $W_\pm(t)$ are sharply peaked at those instants when the escape times are minimal. For the continuous bistable systems under study, two adiabatic transition rates $r_\pm(t)$ were introduced in Eq. (4.61). As a starting point, on taking the limit $D/A_0 x_m \rightarrow 0$ of Eq. (4.61) for $A_0 x_m \ll \Delta V$, we set

$$W_\mp(t) = \alpha \delta[t - (2m + \delta_{\pm 1,1})(\pi/\Omega)]. \quad (4.72)$$

The constant α , not related to the quantity in Sec. IV.B.1, is the escape probability $(r_{\max}/\Omega)(2\pi D/A_0 x_m)^{1/2}$, with $r_{\max} = r_K \exp(A_0 x_m/D)$, and $m = 0, \pm 1, \pm 2, \dots$ (phase locking approximation).

Heading towards the power spectrum of this two-state process, we subtract the periodic oscillations of the asymptotic correlation (see Sec. IV.A) by introducing the newly defined correlation function, i.e., with φ set 0 we write

$$K(t + \tau, t) = \langle [x(t + \tau) - \langle x(t + \tau) \rangle][x(t) - \langle x(t) \rangle] \rangle. \quad (4.73)$$

The phase-averaged correlation function $\bar{K}(\tau)$, defined by averaging $K(t + \tau, t)$ over one period of t , has been obtained by solving the two-state master equation to yield

$$\bar{K}(\tau) = 4 \frac{x_m^2 \mu^{\tau/\Delta + 1}}{(1 + \mu)^2} \mu^{[\tau/\Delta]} (1 - (1 - \mu)[\tau/\Delta]), \quad (4.74)$$

where $\Delta = T_\Omega/2$, $\mu \equiv \exp(-\alpha)$, and $[\tau/\Delta]$ denotes the integer part of τ in units of Δ (Shneidman *et al.*, 1994a, 1994b). The autocorrelation function $\bar{K}(\tau)$ exhibits cusps at each multiple of $T_\Omega/2$. While such cusps of $\bar{K}(\tau)$ look to be a minor artifact of the phase-locking approximation for $D/A_0 x_m \rightarrow 0$, the periodicity of their recurrence will affect the corresponding power spectral density in a rather peculiar way. The Fourier transform of $\bar{K}(\tau)$ reads

$$S(\omega) = \frac{4x_m^2}{\Delta} \frac{\mu(1 - \mu)}{(1 + \mu)\omega^2} \times \frac{1 - \cos(\Delta\omega)}{[1 - \mu \cos(\Delta\omega)]^2 + \mu^2 \sin^2(\Delta\omega)}. \quad (4.75)$$

We next discuss this result and its regime of validity.

(i) $S(\omega)$ corresponds to the noise background component of the power spectral density of the periodically driven dynamics analyzed in Sec. IV.A. It depends on the forcing frequency Ω through a modulation factor and decays according to the same “universal” ω^{-2} power law as the unperturbed system does. Equation (4.75) was derived by assuming the deltalike transition rates of Eq. (4.72). On increasing the noise intensity or, equivalently, upon decreasing the forcing frequency, such an assumption becomes progressively invalid: The switching phase may no longer be locked to the phase of the input signal.

(ii) $S(\omega)$ exhibits sharp dips at even multiples of the driving frequency. In the asymptotic approximation [Eq. (4.75)] the power spectral density vanishes at $2n\Omega$, that is $S(2n\Omega) = 0$. These zeros in the profile of $S(\omega)$ are the fingerprints of the periodic structure of nonanalytical cusps in the correlation function $\bar{K}(\tau)$.

(iii) The experimental observability of the $S(\omega)$ dips is a delicate matter. For finite noise intensities these dips broaden and undergo a “wash out” effect to an extent that we estimate only at the end of this subsection. Here, we limit ourselves to noticing that the width of the sharp peaks in the transition rates $r_\pm(t)$ is, in fact, of order $(D/A_0 x_m)^{1/2} \Omega^{-1}$. It follows that for spectral frequencies ω smaller than $\Omega(A_0 x_m/D)^{1/2}$ the deltalike approximation (4.72) is sound, and the predictions of the present analysis are physically correct. For larger frequencies,

i.e., a finer temporal resolution, the cusps of $\bar{K}(\tau)$ become increasingly unphysical. The condition of observability for the m th dip is thus

$$(2m)^2 \ll A_0 x_m / D. \quad (4.76)$$

Note that detecting even the fundamental dip requires a sufficiently large value of the ratio $A_0 x_m / D$. This explains why dips have not been predicted within the linear-response treatment where $A_0 x_m / D \ll 1$ (see Secs. IV.B and Appendix). The fact that sometimes (at weak noise) the dips were seen experimentally (termed “unexplained” or “strange”—see Zhou and Moss, 1990; Bulsara *et al.*, 1991; Kiss *et al.*, 1993), and sometimes (at strong noise) not, has continued to baffle experimentalists and theorists at the heyday of early stochastic resonance simulations.

(iv) In the derivation of $S(\omega)$, the modulation of the quasiequilibrium states $x_{\pm}(t)$ has been neglected. This effect is significant at large frequencies, when the system, far from being synchronized, may sojourn many forcing cycles in one well. The approximation $x_+(t) - x_-(t) = 2x_m$ employed to derive Eq. (4.74) ought to be improved. Without specializing our analysis to any potential, we observe that the next-to-leading order correction to the difference $x_+(t) - x_-(t)$ is proportional to $(A_0 x_m / \Delta V)^2 \cos^2(\Omega t + \phi)$. It follows that only corrections to fourth order in $A_0 x_m / \Delta V$, i.e., proportional to $(A_0 x_m / \Delta V)^4 \cos(2\Omega t) \cos[2\Omega(t + \tau)]$, may become important in the computation of the phase-averaged correlation function $\bar{K}(\tau)$ and its Fourier transform $S(\omega)$. Shneidman, Jung, and Hänggi (1994a, 1994b) concluded that for

$$\Omega \gg r_{\max}(2\pi D / A_0 x_m)^{1/2} (\Delta V / A_0 x_m)^2, \quad (4.77)$$

a *finite-width* intrawell peak may become detectable at 2Ω , and may even dominate over the relevant $S(\omega)$ dip. Furthermore, on accounting for higher-order terms of the time modulation $x_{\pm}(t)$, we can generate a harmonic structure of intrawell peaks at all even multiples of the forcing frequency.

In the weak-noise limit, the residence-time distribution can be calculated analytically without much effort. We can either proceed as for the two-state model, or adopt the approach of the foregoing Sec. IV.C. In both cases, consistently with the assumptions used above, we can assume a phase-switch distribution given by $Y_{\pm}(\phi) = \delta(\phi - \pi \delta_{\pm 1,1})$. Not surprisingly, the final result coincides with Eq. (4.66), the low-noise expression for $N(T)$ in the adiabatic approximation. It follows immediately that the ratio of any two consecutive $N(T)$ peaks is given by

$$N(T_{n+1}) / N(T_n) = \exp(-\alpha/2), \quad (4.78)$$

with $T_n = (n - 1/2)T_{\Omega}$, and α defined below Eq. (4.72). Thus, as anticipated in Sec. IV.C, the strengths P_n of the $N(T)$ peaks approach one another in the asymptotic limit $\alpha \rightarrow 0$. Moreover, the width of such peaks is of the order of $(D / A_0 x_m)^{1/2} \Omega^{-1}$, namely the same as that of the sharp peaks of the actual transition densities $r_{\pm}(t)$ at low noise, cf. Eq. (4.61).

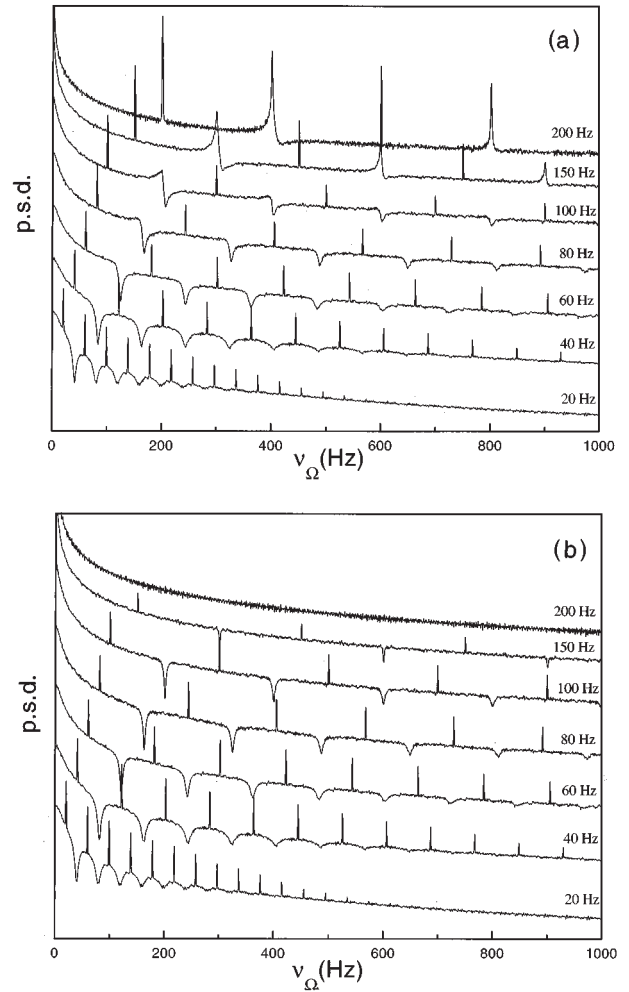


FIG. 14. Unsubtracted power spectral density (p.s.d.) (arbitrary units) of: (a) the full signal; (b) the filtered signal, for different values of the forcing frequency $\nu_{\Omega} = \Omega/2\pi$. The circuit parameters are $\Delta V/D = 1.6 \times 10^3$, $A_0 x_m/D = 2.3 \times 10^3$, and $a = 3.2 \times 10^4 \text{ s}^{-1}$. After Gammaitoni, Marchesoni, Menichella-Saetta, and Santucci (1995).

In a recent paper, Gammaitoni, Marchesoni, Menichella-Saetta, and Santucci (1995) used analog simulations for the periodically driven weak-noise limit of the quartic double-well dynamics to verify the theoretical predictions of Eqs. (4.75)–(4.78). In order to build up significant statistics, these authors increased the amplitude A_0 of the input signal close to, but smaller than, the critical value A_c above which bistability is lost when the maximal tilt is assumed. In this way, the ratios of $A_0 x_m / D$ became of the order 10^3 , and could easily be simulated (strong-forcing regime). As a matter of fact, the condition $A_0 x_m \ll \Delta V$ does not enter explicitly in the discrete-switching approximation. It was originally introduced to simplify $r_{\pm}(\theta)$ to $r_K \exp[\mp (A_0 x_m / D) \cos \theta]$ in Eq. (4.61) and to determine the probability α in Eq. (4.72). Therefore, we expect that in the strong-forcing regime r_{\max} does differ substantially from $r_K \exp(A_0 x_m / D)$ in Eq. (4.72), but its physical role remains unchanged. Gammaitoni, Marchesoni, Menichella-Saetta, and Santucci (1995) verified that in

the limit of weak noise and strong forcing the shape of the n th $N(T)$ peak is approximated well by a Gaussian function with standard deviation $\sigma_T = (D/A_0 x_m)^{1/2} \Omega^{-1}$ independent of n . Furthermore, the peak height $N(T_n)$ turned out to decay according to the exponential law (4.78), whence the estimate of the parameter α . Mante-gna and Spagnolo (1996) reached the same conclusion in their experimental investigation of the switching-time distributions in a periodically driven tunnel diode.

Gammaitoni, Marchesoni, Menichella-Saetta, and Santucci (1995) experimentally confirmed the predictions of Shneidman, Jung, and Hänggi (1994a, 1994b) in great detail. In particular, the following.

(a) The interwell and intrawell dynamics were separated by filtering the output signal $x(t)$ through a two-state filter ($x = \pm x_m$) and contrasting the statistics of the filtered signal with that of the full signal. The *nonsubtracted* power spectral densities for both output signals are displayed in Fig. 14. In both cases the number of resolved dips m is bounded from above by the inequality in Eq. (4.76).

(b) The dip structure of the power spectral densities becomes more apparent by filtering $x(t)$. Most notably, the spectral dips tend to disappear with increasing Ω , and their shape is not as sharp as that predicted by Eq. (4.75).

(c) No peak structure due to the intrawell modulation is observable in the power spectral density of the filtered signal. In contrast, for the full signal, broad peaks located in the vicinity around $2n\Omega$ can be resolved at relatively high values of Ω , namely for $\alpha \ll 1$, in agreement with Eq. (4.77) for $A_0 x_m \approx \Delta V$. Moreover, such peaks get sharper on further increasing Ω , and their height decreases with increasing peak index.

(d) Dips and peaks of the power spectral density may coexist for the full signal, as suggested by items (a) and (b). In such a case, their position deviates slightly from the predicted value $2n\Omega$, by shifting to the right and, possibly, to the left. At very low Ω values, no intrawell modulation peak is detectable, whereas at high Ω values, peaks dominate over dips which, in turn, tend to vanish. We additionally remark that these intrawell modulation peaks should not be mistaken for the delta-like spikes at $(2n+1)\Omega$.

Finally, we mention that in addition to the detailed experimental analog simulations (Gammaitoni, Marchesoni, Menichella-Saetta, and Santucci, 1995), the structure of the characteristic dips in the time-averaged power spectrum $S(\omega)$ at even-numbered harmonics have also been observed in computer simulations of a neural network using a model describing the perceptual interpretation of ambiguous figures (Riani and Simonotto, 1994, 1995) and even in *in situ* experiments with human observers (Riani and Simonotto, 1995). These latter experiments yielded results that are in good qualitative agreement with the neural model predictions for the stochastic resonance power spectra that characterize the perceptual bistability in the presence of noise and weak periodic perturbations.

V. APPLICATIONS

A. Optical systems

1. Bistable ring laser

A ring laser (Sargent *et al.*, 1974) consists of a ring interferometer formed by three or more mirrors and a laser medium inside the cavity. In two-mode ring lasers, the light can travel in a clockwise or counterclockwise direction. Bistability with respect to the direction has been discussed in large detail—see, for example, Mandel, Roy, and Singh (1981). Random switching of the beam intensities, initiated by spontaneous emission in the laser medium and fluctuations in the pump mechanism, indicates bistable operation of the ring laser. To demonstrate stochastic resonance, the symmetry between the two modes has to be broken by applying a periodic modulation that favors one of the modes. In the pioneering experiment by McNamara, Wiesenfeld, and Roy (1988), an acousto-optic modulator (Roy *et al.*, 1987) has been used to convert an acoustic frequency to a modulation of the pump parameter. Gage and Mandel (1988) have alternatively used Faraday and quartz rotators to control the asymmetry of the modes. Before we discuss the experiment by McNamara, Wiesenfeld, and Roy (1988), we briefly review the theoretical description of bistable ring lasers.

The semiclassical equations for a two-mode ring laser (Sargent *et al.*, 1974) augmented by noise sources $\Gamma_{1,2}(t)$ to account for the effect of spontaneous emission and pump noise $p(t)$ assume in dimensionless units the form

$$\begin{aligned}\dot{I}_1 &= 2I_1(a_1 + p(t) - I_1 - \xi I_2) + \sqrt{2I_1}\Gamma_1(t), \\ \dot{I}_2 &= 2I_2(a_2 + p(t) - I_2 - \xi I_1) + \sqrt{2I_2}\Gamma_2(t),\end{aligned}\quad (5.1)$$

where I_1 and I_2 are the scaled dimensionless intensities of the modes. The terms linear in $I_{1,2}$ describe the net gain of the laser modes, those with $I_{1,2}^2$ describe saturation and the mixed terms $\xi I_1 I_2$ represent coupling between the laser modes. The fluctuations of the pumping mechanism are described by the exponentially correlated Gaussian noise $p(t)$, i.e.,

$$\begin{aligned}\langle p(t)p(t') \rangle &= \frac{P}{\tau_c} \exp\left(-\frac{1}{\tau_c}|t-t'|\right), \\ \langle p(t) \rangle &= 0\end{aligned}\quad (5.2)$$

with correlation time τ_c and intensity P , while the spontaneous-emission noise terms are assumed to be uncorrelated:

$$\begin{aligned}\langle \Gamma_i(t)\Gamma_j(t') \rangle &= \delta_{ij}\delta(t-t'), \\ \langle \Gamma_{i,j}(t) \rangle &= 0.\end{aligned}\quad (5.3)$$

The pump parameters a_1 and a_2 are modulated antisymmetrically by noise and a periodic signal (Vemuri and Roy, 1989)

$$\begin{aligned}a_1(t) &= \bar{a} + \Delta a(t) + r(t), \\ a_2(t) &= \bar{a} - \Delta a(t) - r(t),\end{aligned}\quad (5.4)$$

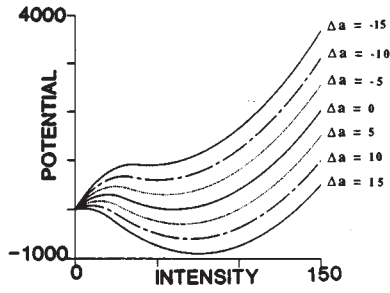


FIG. 15. The effective potential of Eq. (5.8) for the intensity of a single laser mode in a bistable ring laser is plotted as a function of the intensity at the pump strength $\bar{a}=60$ and the coupling $\xi=2$ for different amplitudes $\Delta a=A_0 \cos(\Omega t)$ of the external field (i.e., during one period of the external field). It is seen that the effective barrier height for a transition between the “ON” state and the “OFF” state becomes periodically modulated. After Vemuri and Roy (1989).

with $r(t)$ being white Gaussian (injected) noise

$$\begin{aligned} \langle r(t)r(t') \rangle &= 2R \delta(t-t'), \\ \langle r(t) \rangle &= 0, \end{aligned} \quad (5.5)$$

and

$$\Delta a(t) = A_0 \cos(\Omega t). \quad (5.6)$$

In the absence of the periodic modulation ($\Delta a = \text{const.}$), pump noise, and injected noise, the stationary probability $P_{st}(I_1, I_2)$ can be obtained analytically (Sargent *et al.*, 1974). Integration over the intensity I_2 yields the stationary probability density of I_1 , which can be written as

$$P_{st}(I_1) = \int_0^\infty P_{st}(I_1, I_2) dI_2 = \frac{1}{Z} \exp[-V(I_1)], \quad (5.7)$$

with the effective potential

$$\begin{aligned} V(I_1) = & -\frac{1}{4}(\xi^2 - 1)I_1^2 + \left[\frac{1}{2} \bar{a}(\xi - 1) - \frac{1}{4} \Delta a(\xi + 1) \right] I_1 \\ & - \ln \left\{ \text{erfc} \left[\frac{1}{2} \xi I_1 - \frac{1}{2} \bar{a} + \frac{1}{4} \Delta a \right] \right\}, \end{aligned} \quad (5.8)$$

and erfc being the complementary error function (Abramowitz and Stegun, 1965). For slow and weak periodic driving, the potential $V(I_1)$ undergoes a periodic change obtained by substituting Δa by $A_0 \cos(\Omega t)$, but it remains bistable. The minimum corresponding to the ON state (high intensity) rocks up and down—see Fig. 15. This situation looks similar to the quartic double-well potential, discussed in Sec. IV. There are, however, some differences:

- (1) The potential $V(I_1)$ is only an effective potential. For its construction, the pump noise and injected noise had been neglected. The intuitive picture of a particle (here the intensity) moving in the effective potential can be used only as a heuristic guideline.
- (2) The original Langevin equations (5.1) exhibit no inversion symmetry.

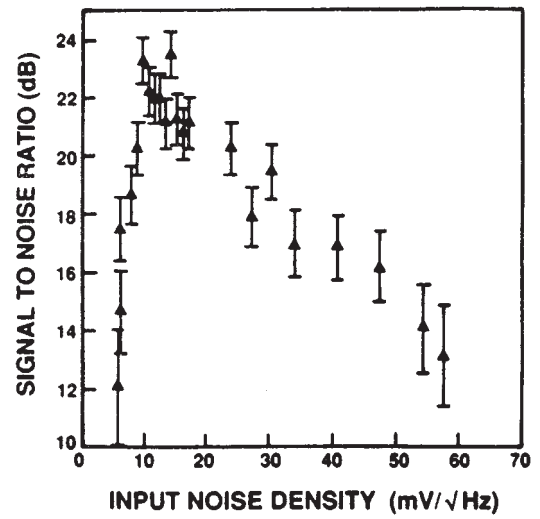


FIG. 16. The signal-to-noise ratio, obtained from the time dependence of the intensity of one laser mode, shown as a function of the injected noise strength. After McNamara, Wiesenfeld, and Roy (1988).

In the experiment by McNamara, Wiesenfeld, and Roy (1988), the intensity of one of the two modes has been extracted from the ring laser. The time series for the intensity has subsequently been compressed into a binary pulse train that contains only information on whether the mode was “ON” or “OFF.” This procedure has been repeated for a large number of samples to obtain the sample-averaged power spectrum. The power spectrum consists of a smooth background, a sharp peak at the frequency of the modulating field Ω , and smaller peaks at multiples of Ω . Most significantly, the intensity of the peaks at the driving frequency first increases with increasing injected noise strength, passes through a maximum and then decreases again. The signal-to-noise ratio, shown in Fig. 16, reflects this characteristic bell-shaped behavior.

For the archetypal quartic bistable potential in Sec. IV, the power spectrum contains peaks only at odd multiples of the driving frequency. In the power spectrum

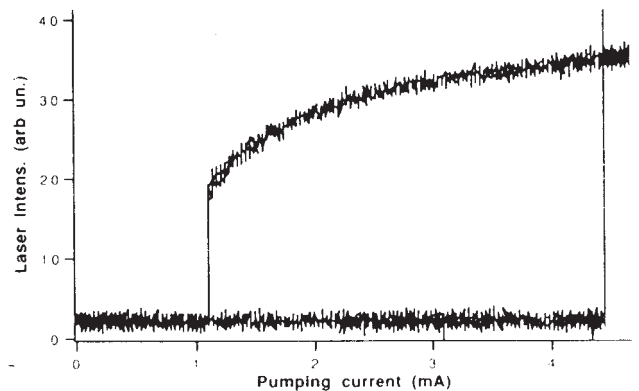


FIG. 17. The light intensity of a single-mode laser with a saturable absorber shown as a function of the pump current. The hysteresis loop indicates optical bistability.

here, however, a peak at twice the driving frequency can be observed, reflecting the lack of inversion symmetry of the two-mode laser Langevin equations (5.1), or the effective potential (5.8).

Digital simulations of Eq. (5.1) in Vemuri and Roy (1989) are in qualitative agreement with the experimental results.

2. Lasers with saturable absorbers

A laser with a saturable absorber is a quantum device consisting of a laser cavity where an amplifying as well as an absorbing medium are placed. Bistability has been observed in these devices by Arimondo and Dinelli (1983) and Arimondo *et al.* (1987). Within a certain range of the pump intensity, the output of the laser can be in two different modes, depending on its previous history.

In Fig. 17, the laser intensity (Fioretti *et al.*, 1993) is shown as a function of the pump current, which was slowly increased until the output switched to the high intensity level, and then decreased again. The observed hysteresis loop is an indication of bistability.

In the experiment by Fioretti *et al.* (1993), the dc pump current is chosen in order that the laser operates in the bistable regime. The pump current is modulated around this dc value by a small periodic signal plus an external Gaussian noise source whose intensity can be controlled.

The periodic signal is too small to cause switching by itself. In the presence of pump noise, however, switching takes place. The intensity of the laser light has been recorded as a function of time and subsequently filtered by a two-bit filter, which only detects “ON” and “OFF” information. The extracted signal-to-noise ratios exhibit maxima as a function of noise strength Q (see below).

The laser with a saturated absorber is described on a semiclassical level by a system of three ordinary differential equations (Zambon *et al.*, 1989)

$$\begin{aligned}\dot{E} &= -\frac{1}{2} \left(D + \frac{\bar{A}}{1+a|E|^2} + 1 \right) E + \xi(t), \\ \dot{D} &= -\gamma(D+A+D|E|^2) - c_1(S-D), \\ \dot{S} &= -\gamma_1(S-D),\end{aligned}\quad (5.9)$$

where E is the complex field amplitude and D is the difference between the population of the upper and lower energy level of the amplifier medium relevant for lasing. The quantity S describes the influence of other energy levels coupled to the populations of the lasing levels; A and \bar{A} describe the strength of the amplifying and absorbing medium, respectively. Quantum fluctuations are modeled by zero-mean white Gaussian noise $\xi(t)$ with $\langle \xi^*(t)\xi(t') \rangle = 4q\delta(t-t')$. In order to simplify the laser equations, S and D are adiabatically eliminated by assuming the field strength E to be small and much slower varying in time than S and D . Periodic modulation and (Gaussian) fluctuations $\zeta(t)$ of the

pump current are taken into account by modulating the amplifier strength A , i.e., $A \rightarrow A(t) = A + F \cos(\Omega t) + \zeta(t)$, with

$$\begin{aligned}\langle \zeta(t)\zeta(t') \rangle &= 2q\delta(t-t'), \\ \langle \zeta(t) \rangle &= 0.\end{aligned}\quad (5.10)$$

One finally arrives at the equation of motion for the field intensity $I \equiv E^*E$

$$\begin{aligned}\dot{I} &= -I \left(1 + \frac{\bar{A}}{1+aI} - \frac{A}{1+I} - \frac{\zeta(t) + F \cos(\Omega t)}{1+I} \right) \\ &\quad + \sqrt{I}\Gamma(t),\end{aligned}\quad (5.11)$$

with the real-valued zero-mean Gaussian noise $\Gamma(t)$, characterized by the correlation function

$$\langle \Gamma(t)\Gamma(t') \rangle = 2Q\delta(t-t'). \quad (5.12)$$

Again, in contrast to the quartic double-well system in Sec. IV.A, the equation of motion (5.11) does not have inversion symmetry. The stationary intensities in the absence of fluctuations and modulation, are determined by the zeros of $F(I) = I(1 + \bar{A}/(1+aI) - A/(1+I))$. For $a > A/(A-1)$, the dynamical system exhibits a subcritical pitchfork bifurcation at $A = \bar{A} + 1$, which explains the hysteretic behavior (see Fig. 17), bistability, and stochastic resonance observed in the experiment when the pump current A is ramped slowly up and down.

3. Model for absorptive optical bistability

Consider a ring interferometer with a passive medium placed in it. Light is coupled into the interferometer through a semipermeable mirror and, likewise, light is transmitted at another mirror. Measuring the intensity of the transmitted wave against the intensity of the incident wave, one finds an S-shaped curve; e.g., for some values of the intensity of the incident beam the intensity of the transmitted wave can have a small and a large intensity. There are different mechanisms that can be responsible for this behavior. One of them is due to nonlinear absorption in the passive medium.

A model for purely absorptive optical bistability in a cavity was introduced by Bonifacio and Lugiato (1978). For the scaled dimensionless amplitude y of the input light and the scaled dimensionless amplitude of the transmitted light x , they have derived the equation of motion

$$\dot{x} = y - x - \frac{2cx}{1+x^2} + \frac{x}{1+x^2}\Gamma(t), \quad (5.13)$$

with $\Gamma(t)$ denoting Gaussian fluctuations of the inversion

$$\begin{aligned}\langle \Gamma(t)\Gamma(t') \rangle &= 2D\delta(t-t'), \\ \langle \Gamma(t) \rangle &= 0.\end{aligned}\quad (5.14)$$

The dimensionless parameter c is proportional to the population difference in the two relevant atomic levels.

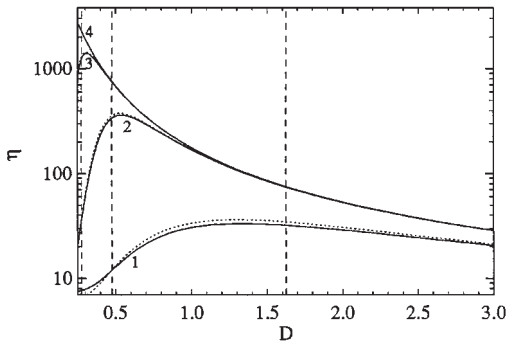


FIG. 18. The numerical results for the spectral amplification η are shown by the solid lines for the “symmetric case” $y_0 = 6.72584$ at $c = 6$ and $A_0 = 10^{-4}$. Different lines labeled according to “ n ” correspond to the external angular frequency $\Omega = 10^{-n}$. The dotted lines correspond to results within the linear-response approximation. They can be distinguished from the numerical results only for frequencies larger than about 10^{-2} . The vertical dashed lines indicate the position D_{SR} of the maxima determined by the argument of matching time scales discussed in Sec. II, and also in Bartussek, Hänggi, and Jung (1994).

For a large enough population difference c , the stationary transmitted light amplitude is an S-shaped function of the amplitude of the injected light, thus indicating bistability.

In the presence of a small, periodic perturbation of the incident light, the parameter y in Eq. (5.13) has to be modified by $y \rightarrow y(t) = y + A_0 \cos(\Omega t)$. The Fokker-

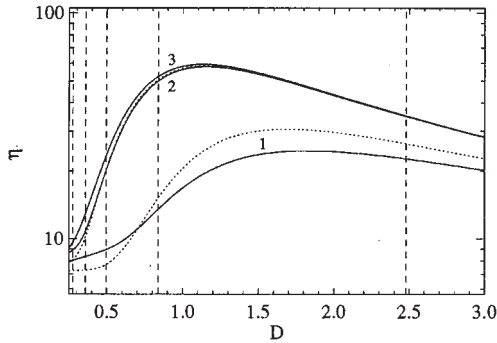


FIG. 19. The numerical results for the spectral amplification η are shown by the solid lines for the “asymmetric case” $y_0 = 6.8$ at $c = 6$ and $A_0 = 10^{-4}$. Different lines labeled according to “ n ” correspond to the external frequency $\Omega = 10^{-n}$. The curves for $\Omega < 10^{-3}$ are not distinguishable from the curve for $\Omega = 10^{-3}$. The dotted lines correspond to results within the linear-response approximation. They can be distinguished from the numerical results only for frequencies larger than about 10^{-2} . The rightmost dashed vertical line indicates the position D_{SR} of the maximum determined by the argument of matching time scales between the period of driving T_Ω and the sum of corresponding two escape times at $\Omega = 10^{-1}$; see Bartussek, Hänggi, and Jung (1994). The angular driving frequencies corresponding to the vertical dashed lines from right to left are $\Omega = 10^{-1}$, $\Omega = 10^{-2}$, $\Omega = 10^{-3}$, $\Omega = 10^{-4}$, and $\Omega = 10^{-7}$.

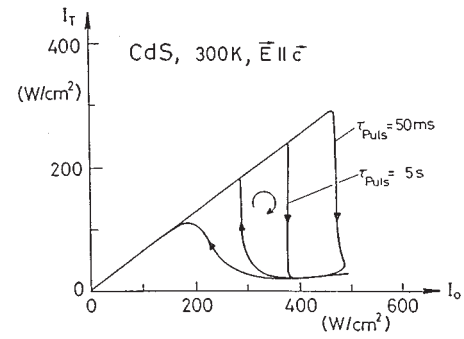


FIG. 20. Transmitted intensity I_t vs incident light intensity I_0 for a laser-illuminated sample of the optical bistable semiconductor CdS. The pulse length of the incident light intensity determines the rate at which the hysteresis loop is run through.

Planck equation corresponding to the Langevin equation (5.13) with the modulated parameter y has been solved numerically by Bartussek, Hänggi, and Jung (1994) by using the matrix-continued fraction technique, and alternatively with linear-response theory for weak modulation A_0 . According to Sec. IV, the spectral amplification η [see Eq. (4.21)] of the periodic modulation has been constructed from the asymptotic long-time solution of the Fokker-Planck equation.

The following discussion is restricted to the bistable regime. Here, it is important to distinguish a symmetric case from an asymmetric case. In the symmetric case the stationary probability density in the absence of periodic driving has two peaks with the same heights in the zero-noise limit $D \rightarrow 0$. For all other cases (asymmetric cases), the peaks have different probabilistic weights at weak noise.

In the symmetric case, one observes stochastic resonance very much like in the quartic bistable double-well potential (see Fig. 18), i.e., a peak in the amplification of the modulation when the sum of the mean escape times out of both stable states equals the period of the driving (the values of the noise strength D where we have such a time-scale matching are indicated by dashed lines in Fig. 18).

In the asymmetric case (Fig. 19), the spectral amplification is suppressed at weak noise, because—in contrast to the symmetric case—the contribution of the hopping motion to the response of the system disappears exponentially for small noise; cf. Sec. IV.B.2. As a consequence, the maximum of the spectral amplification does not indicate a time-scale matching as in the symmetric case. For small driving frequencies, the maximum of the spectral amplification becomes independent of the driving frequency.

4. Thermally induced optical bistability in semiconductors

It has been shown that semiconductors exhibit a thermally induced optical bistability facilitated by the thermal shift of the fundamental band edge (Lambsdorff *et al.*, 1986; Grohs *et al.*, 1989). The semiconductor is almost transparent at low intensities of the incident light.

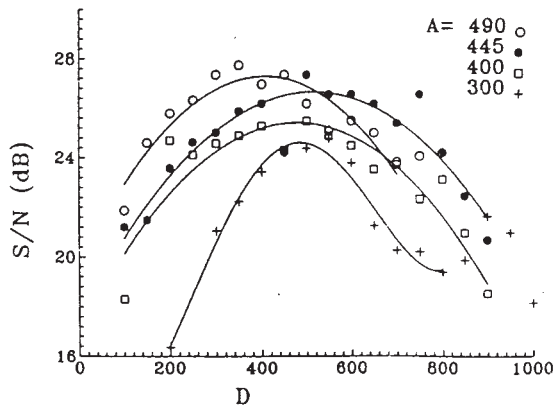


FIG. 21. The signal-to-noise ratio of the transmitted light intensity as a function of the noise strength (in arbitrary units) for different amplitudes A (in arbitrary units) of the modulation of the incident light intensity. The lines are to guide the eye.

Increasing the intensity, the absorbed fraction of light heats up the probe, which in turn induces a stronger light absorption—a nonlinear effect. For large intensities of the incoming light, the transmitted intensity is therefore small. Ramping down the intensity of the incoming light, the transmitted intensity becomes larger again, but describes a hysteresis loop (see Fig. 20) if the absorbed fraction of light is a steep function of the temperature. For increasing ramping speed, the hysteresis loop smears out, but its area increases (a quantitative study of

the overshoot has been done by Grohs *et al.*, 1991, and Jung *et al.*, 1990). Thermally optically bistable semiconductors have been discussed for the design of optical parallel computers (Grohs *et al.*, 1989).

Experiments on stochastic resonance have been performed (Grohs *et al.*, 1994) with a semiconductor (CdS). The CdS crystal had a thickness of about $6\text{ }\mu\text{m}$ and shows thermally induced optical bistability (Lambsdorff *et al.*, 1986; Grohs *et al.*, 1989) under illumination with an Ar^+ laser ($\lambda = 514.5\text{ nm}$). A two-beam setup has been used for the experiments, with both beams incident on the same spot of the crystal and having an equal diameter of approximately $100\text{ }\mu\text{m}$. The transmission state of the crystal is read by a constant beam with an intensity that is too small to induce nonlinear behavior by itself. The intensity of the second beam consists of a constant part to hold the system at the working point (i.e., in the bistable regime), and of a weak periodic signal. Moreover, the weak periodic signal is perturbed by additional injected noise ξ with a controllable amplitude. The transmitted intensity has been measured as a function of time and the power spectrum has been measured. In Fig. 21, the signal-to-noise ratio is shown as a function of the strength of the injected noise. It shows the bell-shaped curve, which is characteristic of stochastic resonance.

5. Optical trap

In the experiment by Simon and Libchaber (1992) a spatial bistable potential was generated by optical means. The experimental setup was based on the fact

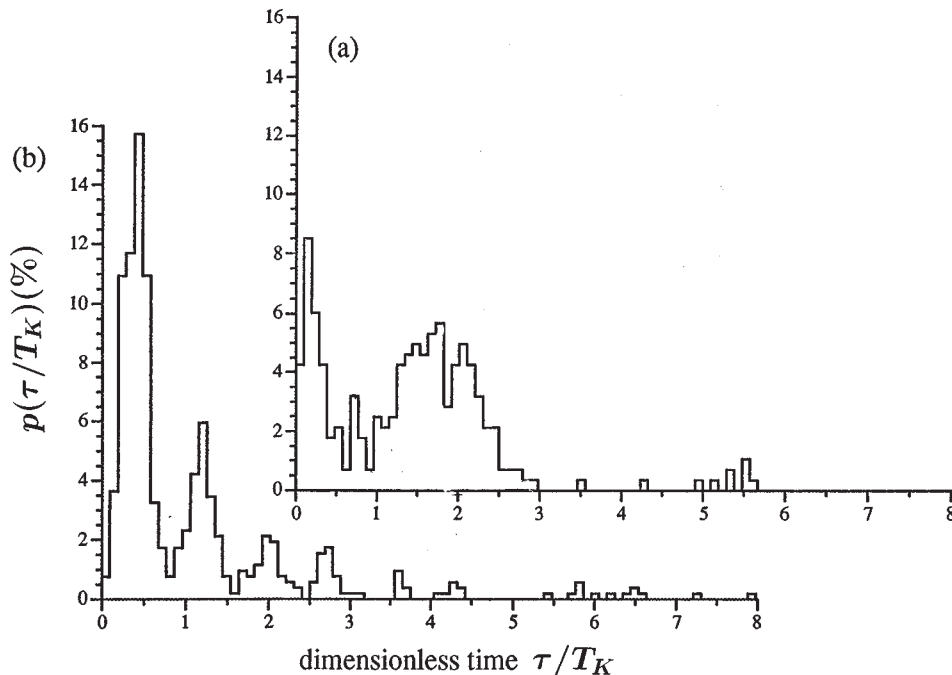


FIG. 22. Escape-time distributions of a particle in an optical trap. The time is measured in units of the mean escape time T_K (from one potential minimum to the other one). (a) The period of the forcing in units of the mean escape time was chosen as $T/T_K = 3.08 > 1$; (b) the period is given by $T/T_K = 0.76 < 1$. While in (b), the peaks are clearly located at odd multiples of half the forcing period, the second peak in (a) is shifted to the left (as far as the accuracy allows for such an interpretation).

that in the presence of an electrical field gradient, a dielectric object (here a $1\ \mu\text{m}$ glass sphere) moves towards the region of highest field strength. The electric field of a laser beam has typically a transverse Gaussian intensity profile, so that dielectric objects are pulled into the beam axis. The two double wells are created by two Gaussian beams obtained from one beam of an Ar laser by utilizing beam splitters. In order to observe stochastic resonance, the depths of the potential wells have to be modulated periodically, which is achieved by modulating the intensity of the two partial beams. The experimental setup was positioned under a microscope so that the motion of the glass sphere in water was directly observable. The pictures were recorded on video and electronically analyzed.

Simon and Libchaber measured the distribution of times the dielectric sphere stayed in one well before it was kicked into the other one. In the absence of modulations, they obtained an exponentially decaying distribution. In the presence of the periodic forcing, they observed a sequence of peaks at odd multiples of the half driving period, with exponentially decaying peak height as shown in Fig. 22. The time is measured in units of the mean escape time T_K out of a potential well in the absence of periodic modulation, i.e., $\tilde{t} = t/T_K$. The period of the modulation in Fig. 22 is measured in the same units. They are given by $T/T_K = 3.08$ (a) and $T/T_K = 0.76$ (b). The first peaks are observed at $\tilde{t} \approx 1.54$ (a) and $\tilde{t} \approx 0.38$ (b), i.e., at the half period of the driving. The other peaks are located at odd multiples of the half period of the driving as predicted by the theory in Sec. IV.C.

In the case that the dwell time equals half the period T an optimal synchronization occurs, leading to the concentration of the escape within the first period; thus anticipating the notion of stochastic resonance in symmetric, bistable systems as a “resonant” synchronization phenomenon (see Sec. IV).

B. Electronic and magnetic systems

In this subsection we review some applications of stochastic resonance to electronic and magnetic systems. We recall that the very first experimental verification of stochastic resonance was realized in an electronic circuit, a simple Schmitt trigger (Fauve and Heslot, 1983). Since then, stochastic resonance has been observed in a variety of more or less complicated electronic devices, mostly constructed with the purpose of building flexible and inexpensive simulation tools. A rather peculiar electronic device that exhibits stochastic resonance is the tunnel diode, a semiconductor device with a bistable characteristic I-V curve, whose dynamics can be controlled by tuning the operating voltage. Due to the fast switching dynamics between stable states (a few tenths of a ns), stochastic resonance in a tunnel diode has been observed for forcing frequencies as high as 10 KHz (Mantegna and Spagnolo, 1994, 1995, 1996). More recently, stochastic resonance has been reported also in a

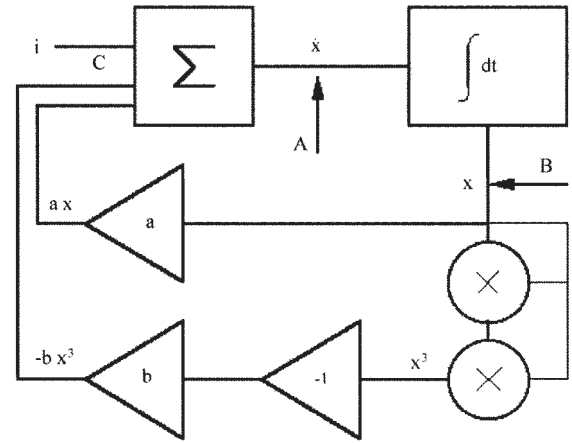


FIG. 23. Functional block scheme for simulating a heavily damped particle moving in a quartic double-well potential; see Eq. (2.1). The triangles denote operational amplifiers.

non-bistable standard np semiconductor diode (Jung and Wiesenfeld, 1997). Further experimental evidence of stochastic resonance driven by externally time-modulated magnetic fields (magnetic stochastic resonance) was reported by Spano, Wun-Fogle, and Ditto (1992) in magnetoelastic ribbons, and by I and Liu (1995) in weakly ionized magnetoplasmas. Finally, magnetic stochastic resonance was predicted theoretically by Grigorenko, Konov, and Nikitin (1990) and Grigorenko and Nikitin (1995) for the interdomain magnetization tunneling in uniaxial ferromagnets, by Raikher and Stepanov (1994) for single-domain uniaxial superparamagnetic particles, and by Pérez-Madrid and Rubí (1995) in an assembly of single-domain ferromagnetic particles dispersed in a low-concentration solid phase. In recent experiments, stochastic resonance has been demonstrated in Bi-substituted ferrite-garnet films (Grigorenko *et al.*, 1994) and also in yttrium-iron garnet spheres, where the noise-free, chaotic spin-wave dynamics alone induces stochastic resonance in the presence of an external modulation (Reibold *et al.*, 1997).

1. Analog electronic simulators

As mentioned in Sec. II.C, electronic circuits have been widely employed in the study of nonlinear stochastic equations (for reviews, see McClintock and Moss, 1989; Fronzoni *et al.*, 1989). Initially (Debnath *et al.*, 1989; Gammaitoni, Marchesoni, *et al.*, 1989; Gammaitoni, Menichella-Saetta, Santucci, Marchesoni, and Presilla, 1989; Zhou and Moss, 1990; Hu *et al.*, 1991), the stochastic resonance studies via analog simulations concentrated on the simulation of the standard quartic double-well system of Eq. (2.1). The temporal behavior of the stochastic process $x(t)$ can be reproduced by using a voltage source $v(t)$ that obeys Eq. (2.1): the coefficients for the dynamical flow are realized by means of a suitable combination of passive analog components, like resistors and capacitors. Active elements (transistors, signal generators, etc.) are employed to simulate more

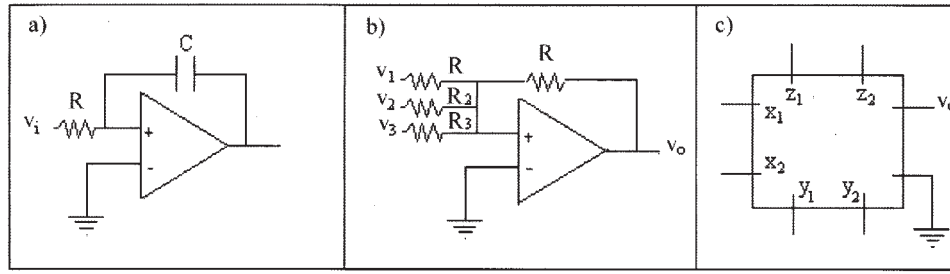


FIG. 24. Circuits for the functional blocks. (a) Miller integrator; (b) adder-amplifier; (c) multiplier.

complicated potential functions. The block scheme of the circuit corresponding to Eq. (2.1) is drawn in Fig. 23. The signal in **A** is assumed to represent the derivative $\dot{x}(t)$ of the variable of interest at time t . In **B**, after the integration block, the signal $x(t)$ is fed into three input terminals: the terminal of the amplifier block **a**, to obtain the signal ax , and the terminals of two multiplier blocks **X**, which yield x^2 after the first block, and x^3 after the second block. The latter signal is then inverted and amplified through two cascaded amplifying blocks. Finally, the three signals in **C** are added together through block Σ to give the starting signal $\dot{x}(t)$.

The realization of an electronic simulation circuit requires the design of specific electronic devices which operate as the single components of the block scheme depicted in Fig. 23. Nowadays, such a purpose is best served by the use of operational amplifiers (Millman, 1983): versatile electronic analog devices characterized by high input impedance, low output impedance, and wide frequency-response range. In Fig. 24 we show the main devices employed to simulate Eq. (2.1). The relevant output voltages v_o are

$$v_o(t) = \frac{1}{RC} \int_0^t v_i(\tau) d\tau \quad (5.15)$$

for a Miller integrator,

$$v_o(t) = \frac{R}{R_1} v_1 + \frac{R}{R_2} v_2 + \frac{R}{R_3} v_3 \quad (5.16)$$

for an adder amplifier, and

$$v_o(t) = \frac{(x_1 - x_2)(y_1 - y_2)}{10} + (z_1 - z_2) \quad (5.17)$$

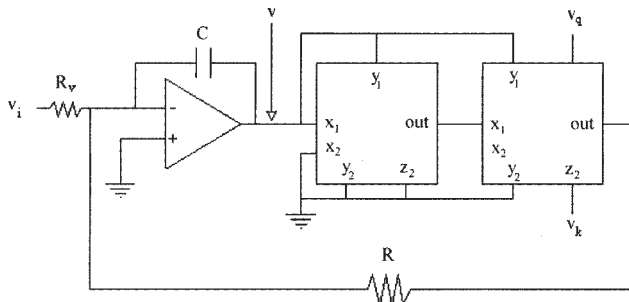


FIG. 25. Simulator circuit for the overdamped dynamics of Eq. (4.3).

for a multiplier. The block scheme of Fig. 23 can thus be translated into the electronic circuit scheme of Fig. 25. The underdamped dynamics of Eq. (4.1), with $V(x)$ defined in Eq. (4.7), can be simulated by a modified version of the previous circuit—see Fig. 26. Usually both the stochastic force $\xi(t)$ and the periodic modulation $A_0 \cos(\Omega t)$ are fed into the simulation circuits by means of suitable voltage generators. The output voltage $v_o(t)$ is sampled at regularly spaced times t_m , with $t_{m+1} - t_m = \Delta t$, digitized as $v_{0,m} = v_o(t_m)$ and then stored into a digital memory unit. The sequence (t_m, v_m) is finally analyzed, mostly off line, by means of standard data-analysis algorithms.

The first, and probably the simplest circuit of this type employed to investigate stochastic resonance dates back to Fauve and Heslot (1983). It consists of a Schmitt trigger, a two-state hysteretic device, that can be easily realized by means of an operational amplifier with positive feedback (Fig. 27). If the input voltage v_i is lower than a threshold value V_+ , the output voltage v_o takes on the constant value $v_o = V_-$. On increasing v_i through V_+ , the output voltage v_o switches to $v_o = V_+$ and stays positive as long as v_i is larger than a second threshold value V_- with $V_- < V_+$. Vice versa, the output transition from V_+ to V_- occurs at $v_i = V_-$. The amplitude of the hysteresis cycle is given by $V_+ - V_-$. The stochastic resonance phenomenon follows from the periodic modulation of the position of the center of the hysteresis cycle around the mean value $(V_+ + V_-)/2$. Such a modulation is reminiscent of the periodic tilting of the wells of a bistable potential model. As a matter of fact, Eqs. (2.6) and (2.7) apply to the output $x(t) \equiv v_o(t)$ of a symmetric Schmitt trigger (where, say, $V_+ = -V_- = V_b$) with minor

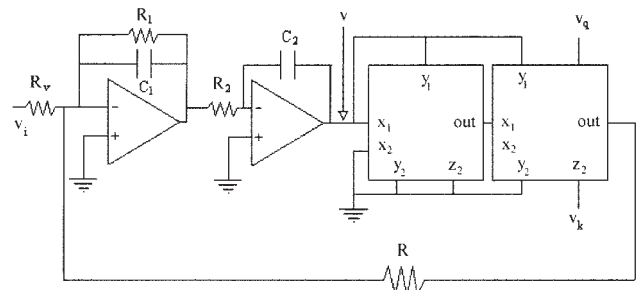


FIG. 26. Simulator circuit for the underdamped dynamics of Eq. (4.1).

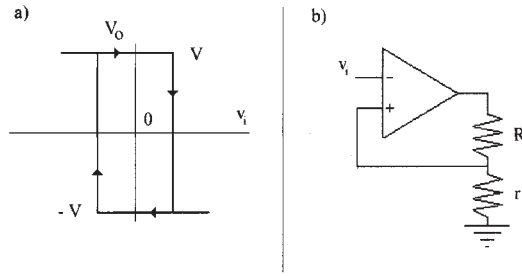


FIG. 27. Operational amplifier in the Schmitt trigger configuration: (a) transfer characteristic; (b) electronic scheme.

changes. For a Gaussian, exponentially autocorrelated input signal $v_i(t)$ with zero mean, variance σ_i^2 , and correlation time τ_i , the parameters in Eqs. (2.7) read $x_m = V$, $D = \sigma_i^2 \tau_i$, and $r_K = [T_K]^{-1}$. The time $T_0(V_b)$ denotes the mean-first-passage time for $v_i(t)$ to diffuse from $-V_b$ up to V_b ; in the subthreshold limit $\sigma_i \ll V_b$ it may be approximated to $\tau_i \sqrt{2\pi} (\sigma_i/V_b) \exp(V_b^2/2\sigma_i^2)$ (Melnikov, 1993; Shulgin *et al.*, 1995). Note that when the amplitude of the hysteresis cycle is forced to zero, the two thresholds overlap and we end up with a single threshold system (see Sec. VII.A.1).

By now, stochastic resonance has been studied in a variety of analog circuits: symmetric and asymmetric quartic double-well potentials (Debnath *et al.*, 1989; Dykman, Luchinsky, *et al.*, 1992; Gammaitoni, Marchesoni, and Santucci, 1994; Gammaitoni, Marchesoni, *et al.*, 1994); the Hopfield neuron potential (Bulsara, Jacobs, Zhou, Moss, and Kiss, 1991); the Fitzhugh-Nagumo neuron model (Wiesenfeld *et al.*, 1994), to mention only a few.

2. Electron paramagnetic resonance

An EPR system consists of a paramagnetic sample placed in a microwave cavity. A microwave generator irradiates the sample while a feedback electronic circuit locks the oscillator frequency to the resonant frequency ν_c of the cavity. A static magnetic field H_0 is applied to the cavity and slowly modulated in order to vary the Larmor frequency $\nu_0 = \gamma H_0/2\pi$ of the sample. Here γ is the gyromagnetic factor. The cavity response, determined by measuring the reflected microwave power, usually exhibits a single minimum at $\nu = \nu_c$; however, in the presence of a strong coupling between the cavity and the spin system (a high number of paramagnetic centers), one observes a splitting of the resonance frequency into two frequencies and the cavity response exhibits two distinct minima separated by a local maximum. The block scheme of an EPR experiment is shown in Fig. 28: an electronic adder (block Σ); a standard microwave spectrometer (block **S**) made of a microwave generator, a resonant cavity, and the relevant circuitry; and the feedback system (block **F**) that locks the microwave source frequency to the maximum absorption of the cavity. Block **C** contains the measurement instrumentation, mainly a frequency counter, to monitor the working frequency, and a power meter to

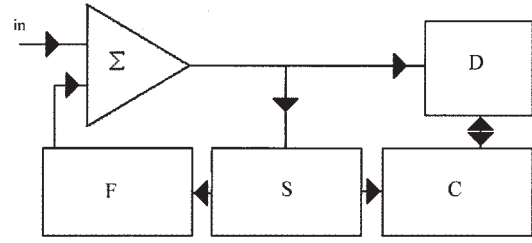


FIG. 28. Block scheme of an EPR system. Block **S**: EPR spectrometer; block **C**: measurement devices; block **D**: data-acquisition system; block **F**: feedback system; block Σ : adder.

measure the power reflected from the cavity. Block **D** represents the data acquisition system.

In such a device, stochastic resonance was observed first by monitoring the working frequency ν (Gammaitoni, Marchesoni, *et al.*, 1991; Gammaitoni, Martinelli, *et al.*, 1991, 1992). A polypyrrole paramagnetic sample was placed in the resonant cavity. The reflection coefficient of the cavity, measured as the ratio between the reflected to the incoming power, was monitored by varying the working frequency. Under proper conditions (strong coupling) the reflection coefficient, as a function of the frequency, showed two separated minima, both stable working points for the spectrometer. The dynamical behavior of the frequency ν is driven by the feedback system. Under stationary conditions, ν fluctuates around the frequency of one of the two minima of the reflection coefficient due to the action of the internal noise of the system. When such noise intensity grows appreciable compared to the height of the barrier that separates the two minima, random switches are observed. Under such circumstances the EPR system shows a noise-driven operating frequency. Its dynamics can be described by the approximate stochastic differential equation:

$$\dot{\nu} = -\gamma \dot{\nu} - V'(\nu) + \xi(t), \quad (5.18)$$

where $V(\nu)$ denotes the effective bistable potential related to the reflection coefficient. The same dynamics can be observed by measuring the error voltage signal generated by the feedback system. On inserting in the feedback loop an external signal made of a periodic and a random component and varying the intensity of the injected noise, behavior typical of stochastic resonance was detected and measured.

3. Superconducting quantum interference devices

The basic components of a SQUID are a superconducting loop and a Josephson junction. For practical purposes, a SQUID can be envisioned as an electromagnetic device that converts a magnetic flux variation into a voltage variation and, as such, it has been successfully employed in monitoring small magnetic field fluctuations. One of the major limitations to a wide use of these devices is their extreme sensitivity to environmental noise. Recently, two different groups (Hibbs *et al.*, 1995; Rouse *et al.*, 1995) succeeded in operating SQUIDS under stochastic resonance conditions with the aim of in-

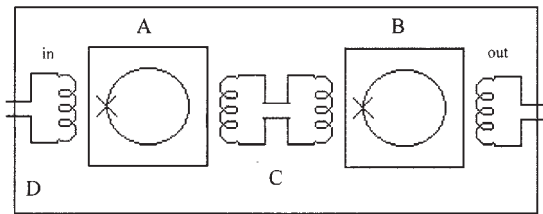


FIG. 29. Block scheme of the SQUID system. Block A: test SQUID, driven by the external signal; block B: measurement SQUID; block C: inductive coupling between the SQUIDs and with the outside circuitry; block D: low-temperature shield.

creasing low magnetic-field detection performance. The dynamics of the magnetic flux Φ trapped in a standard rf. SQUID can be generally described in terms of a second-order differential equation:

$$LC\ddot{\Phi} = -\tau_L\dot{\Phi} - V'(\Phi) + \Phi_e(t), \quad (5.19)$$

where

$$V(\Phi) = \Phi + (\beta/2\pi)\sin(2\pi\Phi) \quad (5.20)$$

plays the role of an effective potential and $\Phi_e(t)$ represents an externally induced flux variation. Here, the flux Φ is measured in units of the fundamental flux quantum $\Phi_0 = h/2e$; L is the inductance of the loop, while C is the junction capacitance and $\tau_L = L/R_j$, where R_j is the normal-state resistance of the junction. The quantity $\beta = 2\pi Li_c/\Phi_0$, with i_c being the loop critical current, is the adjustable parameter that allows us to set the shape of the effective potential. With a proper choice of β and for small Φ values, the system undergoes bistable dynamics. The SQUID loop can be shunted by a low resistance in order to reduce the capacitance and discard inertial effects in the flux [Eq. (5.19)].

The block scheme of a generic SQUID system is shown in Fig. 29. Blocks A and B represent two SQUIDs. The first one is driven with the external signal and the second one is used to pick up the signal output. The inductive coupling between the two loops and with the outside circuitry is also shown (block C). The system is shielded by a low-temperature cage (block D). Stochastic resonance was observed both in the external noise-injected configuration (Hibbs *et al.*, 1995) and in the thermally driven configuration (Rouse *et al.*, 1995), where the switching of the magnetic flux was driven by the internal noise inherent in the SQUID. In both cases a low-frequency periodic signal was injected externally.

C. Neuronal systems

The development of stochastic resonance took a large leap forward when its potential relevance for neurophysiological processes had been recognized. Longtin, Bulsara, and Moss (1991) observed that interspike interval histograms of periodically stimulated neurons exhibit a remarkable resemblance to residence-time distributions of periodically driven bistable systems (Gammaitoni, Marchesoni, Menichella-Saetta, and Santucci, 1989; Zhou *et al.*, 1990).

In this section, we report on the relevant neurophysiological experiments and describe how stochastic resonance enters naturally into standard models for neuronal dynamics. By now, stochastic resonance is a well accepted paradigm in the biological and neurophysiological sciences, and several recent reviews on neurophysiological applications of stochastic resonance are available (Moss, 1994; Moss *et al.*, 1994; Moss and Wiesenfeld, 1995a, 1995b; Wiesenfeld and Moss, 1995). Thus we keep this subsection—though there is a prodigious potential for future applications—somewhat tight.

1. Neurophysiological background

There is a large variety of types of neurons in the nervous system of animals and humans with variations in structure, function, and size. Let us restrict ourselves here to a canonical neuron (Amit, 1989), which presents the underlying functional skeleton for all neurons. The canonical neuron is divided into three parts, an input part (the *dendritic arbor*), a processing part (the *soma*), and a signal transmission part (the *axon*).

A neuron communicates via *synapses*, which are the interfaces between its dendrites and axons of presynaptic neurons, i.e., neurons that talk to the considered neuron. There are a number of dendritic trees entering the soma of the neuron. Usually, one axon leaves the soma and then, downstream, it branches repeatedly to communicate with many postsynaptic neurons, i.e., the neurons the specified neuron is talking to. The fundamental process of neural communication is based upon the following sequence:

- (1) The neural axon is in one of two possible states. In the first state, it propagates an action potential based on the result of the processing in the soma. The shape and amplitude of the propagating action potential—the potential difference across the cell membrane—is very stable, and is replicated at the branching points in the axon. The amplitude is of the order of 10^{-1} mV. In the second state of the axon, i.e., the resting state, no action potential is propagated along the axons.
- (2) When the propagating action potential reaches the endings of the axons it triggers the secretion of *neurotransmitters* into the synaptic cleft.
- (3) The neurotransmitters travel across the synapse and reach the membrane of the postsynaptic neuron. The neurotransmitters bind to receptors that cause the opening of channels, allowing the penetration of ionic currents into the postsynaptic neuron. The amount of penetrating current per presynaptic spike is a parameter that specifies the efficiency of the synapse. There are different ion channels for different ions. To open, say, a potassium channel, a specific neurotransmitter substance is required.
- (4) In the absence of a neurotransmitter the resting potential of the membrane of the postsynaptic neuron is determined by the balance of the resting fluxes of ions such as sodium and potassium. The resting

membrane voltage is typically slightly above the low-lying potassium voltage. The opening of say a sodium channel disturbs the equilibrium and triggers a postsynaptic potential close to the high sodium voltage. The membrane voltage is bound between the lower potassium voltage and the higher sodium voltage. The fact that the rest state is very close to the lower limit leads to a rectification of external stimulus in sensory neurons and is of particular importance for the study of stochastic resonance in neurons.

- (5) The postsynaptic potential diffuses in a graded manner towards the soma. It loses thereby around 80% of its amplitude. Here, in the processing unit of the neuron, the inputs from all presynaptic neurons (of the order 10^4) are summed. The individual postsynaptic potentials are about 1 mV in amplitude. These inputs may be *excitatory*—depolarizing the membrane of the postsynaptic neuron, increasing the probability of a neuronal discharge event (spike), or they may be *inhibitory*—hyperpolarizing the postsynaptic membrane, thereby reducing the probability of a spike. The high connectivity allows for two kinds of summation processes, temporal and spatial summation. Both summation processes are used in nature. Having a serial input of a train of incoming pulses at **one** synapse, local summation can take place. The typical separation of the pulses is, however, of the same order as the typical decay time of a postsynaptic potential (leakage rate). Spatial summation of incoming events from **many** different synapses does not suffer from the leakage rate, but requires a spatial distribution of (even very little) information throughout a local neural network.
- (6) If the sum of postsynaptic potentials arriving within a short period of time exceeds a certain threshold, the probability for the emission of a spike becomes very large. This threshold is of the order of tens of milliseconds and it therefore requires quite a number of inputs to produce a spike.

2. Stochastic resonance, interspike interval histograms, and neural response to periodic stimuli

Over the last 50 years a large body of research has been carried out to understand the encoding of acoustic information on the primary auditory nerve of mammals (see, for example, Teich *et al.*, 1993, and references therein). Rose, Brugge, Anderson, and Hind (1967) measured the interspike interval histogram of sinusoidally stimulated auditory nerve fibers. A few typical examples from a squirrel monkey are shown in Fig. 30. On the vertical axes, the numbers of intervals of length τ (horizontal axes) between two subsequent spikes (taken from a long spike train) are shown. The first peak is located at the period T_Ω of the stimulus and the following peaks are located at multiples of T_Ω . In contrast to the conventional theory of auditory information encoding, the results above indicate that the information of

the period of the stimulus is also encoded in the temporal sequence of the action potentials (spikes). In other words, there is a correlation between the temporal sequence of neuronal discharge and the time dependence of the stimulus. In the neurophysiological literature this correlation is termed *phase locking*. Earlier work reporting the limitation of the phase locking of the neural discharge to the stimulus to small frequencies (<6 kHz) (Rose *et al.*, 1967) has been suggested to be too pessimistic by Teich, Khanna, and Guiney (1993). These authors argue that the synchronization holds up to 18 kHz. They further argue that information about the period of the stimulus is encoded in the temporal sequence of the action potentials over virtually the complete band of acoustical perception.

The resemblance of interspike interval histograms and residence-time distributions of noisy driven bistable systems—see Sec. IV.D—has connected stochastic resonance research with neuronal processes. In Longtin, Bulsara, and Moss (1991), Longtin (1993), and Longtin *et al.* (1994), the interspike interval histograms of a sinusoidally stimulated auditory nerve from a cat have been compared with return-time distributions of the periodically driven quartic bistable potential, i.e.,

$$\dot{x} = x - x^3 + \xi(t) + A_0 \cos(\Omega t), \quad (5.21)$$

and a soft bistable potential

$$\dot{x} = -x + b \tanh(x) + \xi(t) + A_0 \cos(\Omega t), \quad (5.22)$$

with the Gaussian noise

$$\begin{aligned} \langle \xi(t) \xi(t') \rangle &= \frac{D}{\tau_c} \exp\left(-\frac{|t-t'|}{\tau_c}\right), \\ \langle \xi(t) \rangle &= 0. \end{aligned} \quad (5.23)$$

The left potential well corresponds to a neuron that is quiescent; the right potential well corresponds to the firing state. The noise correlation time τ_c has been adjusted to a typical value of the decay of a membrane potential, i.e., $\tau_c = 10^{-4}$ in the same dimensionless units as in Eqs. (5.21) and (5.22). In Fig. 31, the return-time distributions, i.e., the density of time intervals T it takes for the system to be first kicked from one stable state to the other and back again (this second process simulates a reset mechanism), are compared with interspike interval histograms taken from the cat's primary auditory nerve. With only one fitting parameter, it was possible to achieve excellent agreement. In particular, the sequence of peaks in the return-time distributions as well as those in the interspike interval histograms decay exponentially for large return times.

The key question is whether neurons also exhibit stochastic resonance. To this end, Moss and collaborators set up an experiment to study the neural response of sinusoidally stimulated mechanoreceptor cells of crayfish (Douglass, Wilkens, Pantazelou, and Moss, 1993). This experiment has allowed detailed studies of interspike interval histograms to be carried out for a wide range of values of the amplitude and frequency of the stimulation. As in Longtin, Bulsara, and Moss (1991),

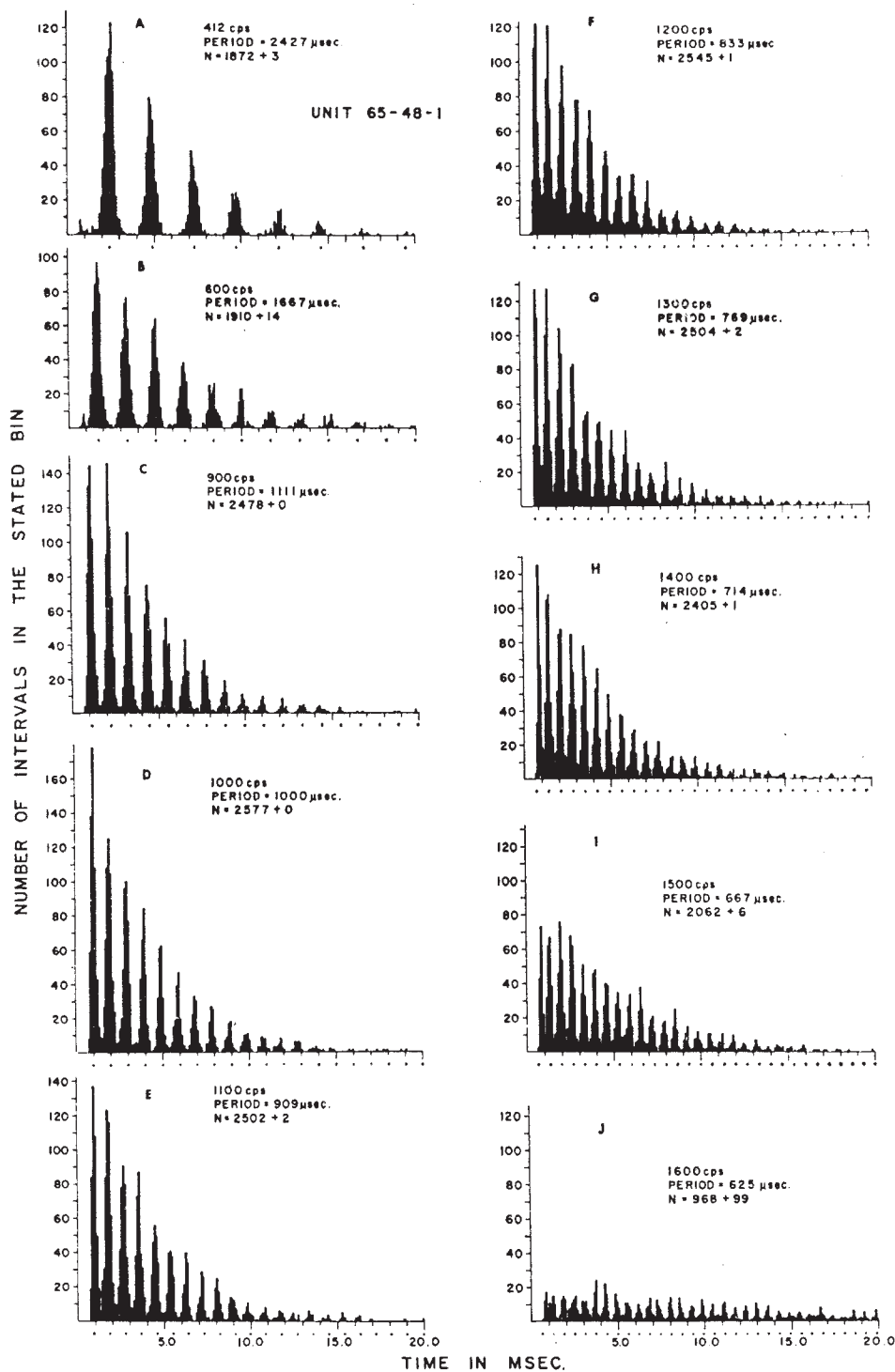


FIG. 30. Distributions of interspike intervals when pure tones of different frequencies activate the neuron. Stimulus frequency in cycles per second (cps) is indicated in each graph. Time on the abscissa is in milliseconds. The dots below the abscissa indicate multiples of the period for each frequency employed. On the ordinate, the number of intervals in one bin is plotted (1 bin = 100 μ s). N is the total number of interspike intervals in the sample. N is given as two numbers: the first indicates the number of intervals plotted; the second is the number of intervals whose values exceeded the maximal value of the abscissa. After Rose *et al.* (1967).

the interspike interval histograms have been reproduced by return-time distributions of periodically driven bistable systems. Without a stimulus, the interspike interval histograms decay exponentially for large return

times. In the presence of periodic driving, Moss *et al.* (1993) found—as in the earlier experiments—a multi-peaked structure with exponentially decaying peak heights. In order to identify stochastic resonance, the

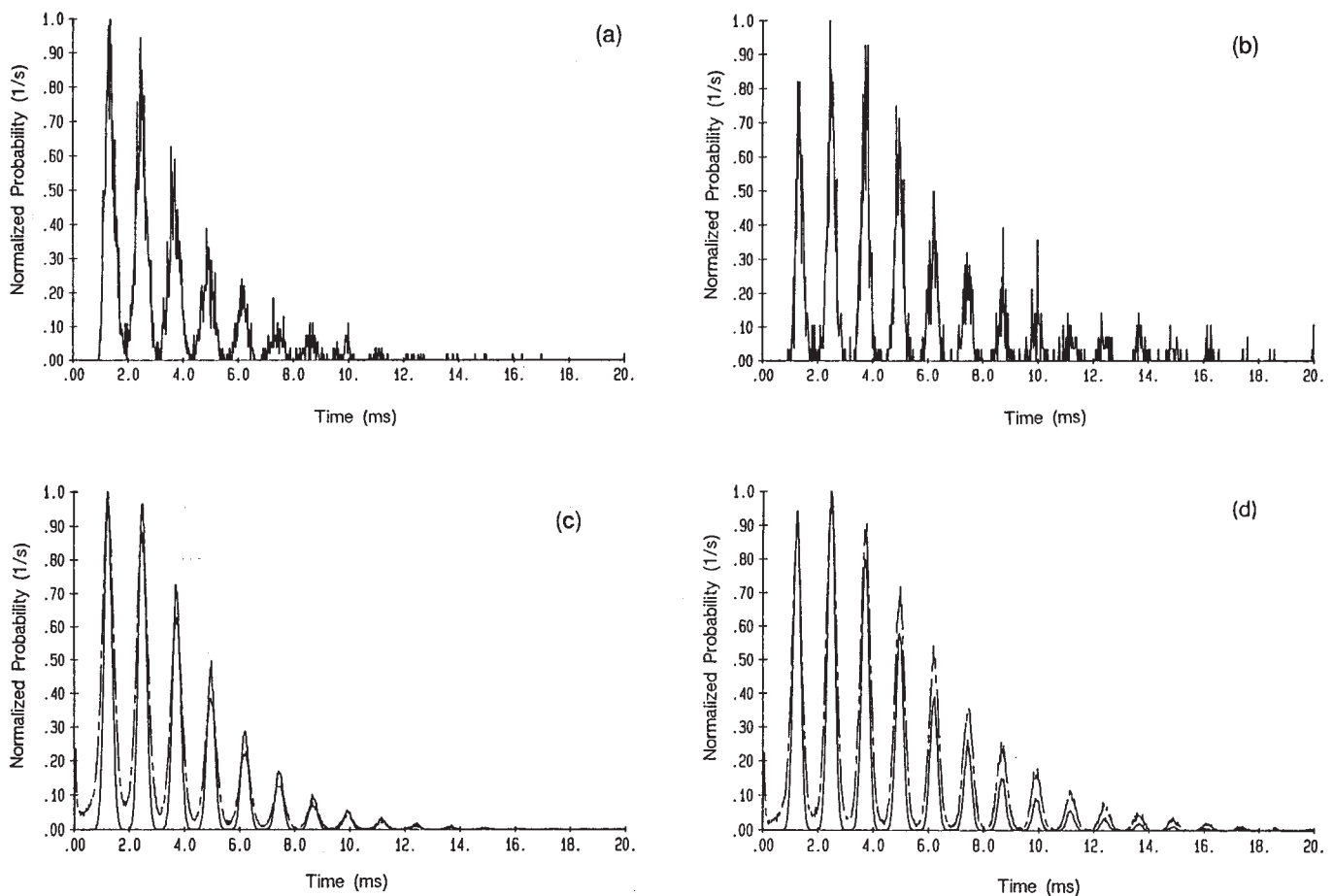


FIG. 31. Return-time distributions for a sinusoidally stimulated auditory nerve: (a), (b) experimental interspike-interval histogram data from a cat primary auditory nerve with an 800-Hz stimulus. The amplitude of the stimulus is 60 dB in (a) and 30 dB in (b). The full curves in (c) and (d) are results obtained from analog simulation of the standard quartic model of Eqs. (5.21)–(5.23). The parameters A_0 and τ_c have been chosen fixed, while D has been fitted to yield best agreement with the interspike-interval histograms in (a) and (b). The best fit in (c) is obtained at a smaller noise level D than in (d). The ratio of driving and noise level in (c) is thus higher than in (d). This is in agreement with the higher stimulus in (a) than in (b). After Longtin, Bulsara, Pierson, and Moss (1994).

noise level had to be changed. To change the intrinsic noise level is not straightforward and requires some more involved experimental procedures (Pei *et al.*, 1996). Douglass, Wilkens, Pantazelou, and Moss (1993) chose to change the noise level by adding noise externally, i.e., by applying to the neuron a sum of a single tone and some noise. The spectral properties of the series of action potentials were analyzed, yielding a power spectrum typified by background noise plus sharp peaks at multiples of the frequency of the stimulus. The signal-to-noise ratio is shown in Fig. 32. The shape of the curve indicates stochastic resonance in a living neuron. An alternative way of describing stochastic resonance is the dependence of the peak height of the peaks in the interspike interval histograms on the noise level (Zhou *et al.*, 1990). The height of the first histogram peak at the period of the stimulus runs as a function of the noise strength through a maximum at a value of the noise that is very close to the peak of the signal-to-noise ratio (see Fig. 32).

The experiments provide evidence that the firing of

periodically stimulated neurons actually exhibits stochastic resonance. These results, obtained as *in situ* experimental results, are quite satisfactory. The present theory, however, based on bistable dynamics [Eqs. (5.21) and (5.22)], does not describe neuronal dynamics very well. This is because the firing state of a neuron is not a stable state. After a neuron has fired, it becomes automatically quiescent after a refractory time. For a more realistic modeling of stochastic resonance in neuronal processes it is therefore necessary to study different, nonlinear (nonbistable) systems.

3. Neuron firing and Poissonian spike trains

Wiesenfeld, Pierson, Pantazelou, Dames, and Moss (1994) proposed a very elegant approximate theory for modeling neuron firing in the presence of noise and a periodic stimulus. The neuron emits uncorrelated, sharp spikes (δ spikes with weights normalized to unity) at random times t_n . The spiking rate, however, is inhomogeneous, i.e., sinusoidally modulated. This sort of pro-

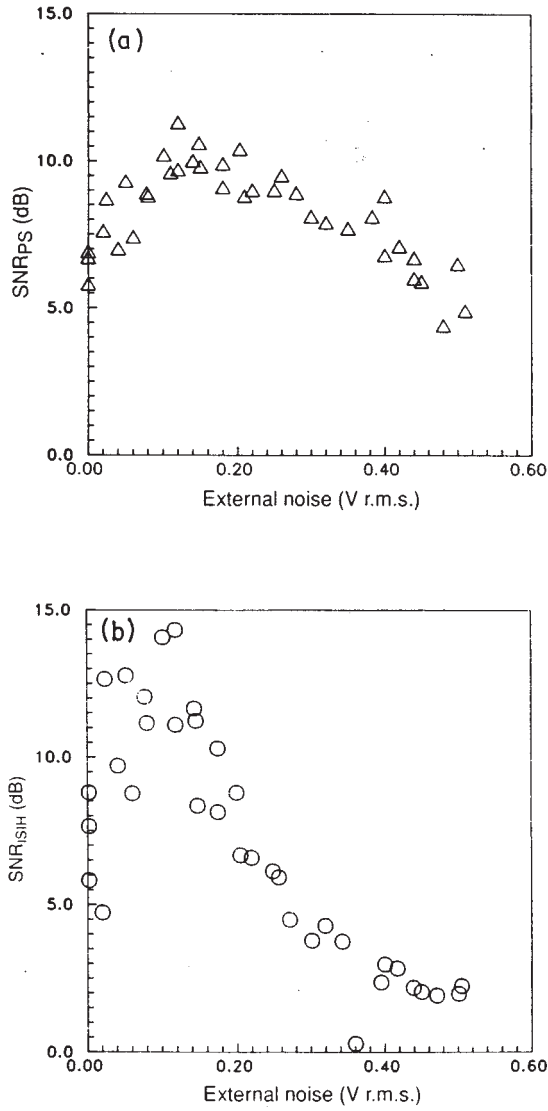


FIG. 32. Stochastic resonance in mechanoreceptor cells of the tail fan of a crayfish: (a) The signal-to-noise ratio obtained from power spectra and (b) from the peak height of the first peak in the interspike-interval histogram as a function of the externally applied noise level.

cess is described by the theory of inhomogeneous Poissonian point processes—see, for example, chapter 6 of Stratonovich (1963). Given the time-dependent spiking rate $r(t)$, the probability to find s events in the time interval T is given by [cf. Eq. (6.33) in Stratonovich (1963)]

$$P_s = \frac{1}{s!} \left[\int_0^T r(t) dt \right]^s \exp \left\{ - \int_0^T r(t) dt \right\}. \quad (5.24)$$

The (phase-averaged) spectral density of the spike train consists of a frequency-independent term (white shot noise), given by the time-averaged spiking rate \bar{r} , and a sum of δ peaks at multiples of the stimulus frequency, where the intensities of the peaks are given by the Fourier coefficients α_n of the periodic spiking rate $r(t)$, i.e.,

$$S(\omega) = \bar{r} + 2\pi \sum_{n=1}^{\infty} |r_n|^2 \delta(\omega - n\Omega),$$

$$r_n = \frac{1}{T} \int_0^T r(t) \exp(-in\Omega t) dt. \quad (5.25)$$

At this point, the only assumption made is that there are no correlations between the spikes. It is remarkable that spike-spike correlations yield only an additional term to the spectral density (Jung, 1995), but otherwise leave the result [Eq. (5.25)] invariant. To analyze further the expression (5.25), some approximations need to be made. For $r(t)$, an expression has been chosen, motivated by the rate theory for noise-induced barrier crossing in the presence of periodic external forces, i.e. [cf. Eq. (4.60)],

$$r(t) = \nu \exp \left[-\frac{\Delta V}{D} - \frac{A_0 x_m}{D} \cos(\Omega t) \right], \quad (5.26)$$

where ΔV is the barrier height in the absence of the forcing, D is the noise strength, A_0 is the amplitude and Ω is the frequency of the periodic forcing, x_m is a scale factor, and the prefactor ν depends on details of the rate process. The expression (5.26) is limited to slow and weak periodic forcing, i.e., Ω is small compared to the local (intrawell) relaxation rate and A_0 is small enough that its effect on the rate can be treated as a perturbation, i.e., $A_0 x_m / D$ has to be small. In leading order $(A_0 x_m / D)^2$, the signal-to-noise ratio (SNR) is given by the ratio of the intensity of the δ peak of $S(\omega)$ at Ω , i.e.,

$$r_1 = \nu I_1(A_0 x_m / D) \exp \left(-\frac{\Delta V}{D} \right) \approx \nu \frac{A_0 x_m}{2D} \times \exp \left(-\frac{\Delta V}{D} \right) \quad \text{for } A_0 x_m / D \ll 1, \quad (5.27)$$

to the noise background in the absence of the periodic driving

$$S_N^0(\Omega) = \bar{r} = \nu \exp \left(-\frac{\Delta V}{D} \right), \quad (5.28)$$

yielding

$$SNR = \frac{4\pi |r_1|^2}{\bar{r}} \approx \frac{\pi x_m^2 A_0^2}{D^2} \exp \left(-\frac{\Delta V}{D} \right). \quad (5.29)$$

Note that corresponding expression in Wiesenfeld, Pierson, Pantazelou, Dames, and Moss (1994) differs in the prefactor, because they used a different definition of the spectral density.

The signal-to-noise ratio shows the characteristic feature of stochastic resonance, i.e., a peak as a function of the noise strength D . Comparison with data obtained from spike sequences of a mechanically modulated mechanoreceptor of a crayfish (see Fig. 33) shows qualitative agreement. The decrease of the SNR at large noise levels, however, is overestimated by this theory. A nonadiabatic theory based on threshold crossing dynamics (see Sec. V.C.5) predicts a slower decrease. Stochastic resonance was also demonstrated in an experiment that uses a sensory hair cell of a cricket (Levin and Miller, 1996).

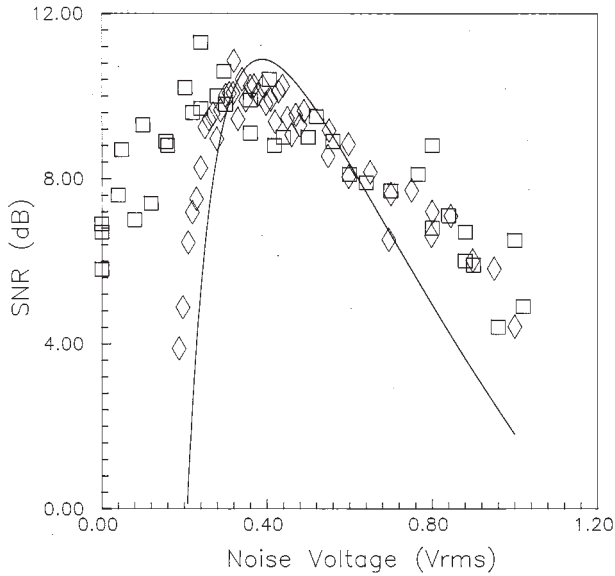


FIG. 33. Signal-to-noise ratio in crayfish mechanoreceptors (squares) compared to the electronic Fitzhugh-Nagumo simulation (diamonds) (Wiesenfeld *et al.*, 1994) and the theory of Sec. V.C.3 (solid curve). The horizontal axis represents the intensity of externally applied noise: hydrodynamic noise in the case of the mechanoreceptors, and electronic noise in the case of neuron models. The crayfish data do not decrease rapidly for small noise because of the residual internal noise of the neuron. Figure provided by Professor Moss.

4. Integrate-and-fire models

A very common model for neuronal dynamics is the so-called integrate-and-fire model. The model works as follows: the input $i(t)$ of the neuron consists of a spike train (cortical neurons) or a continuous signal (sensory neurons). The membrane voltage $u(t)$ is obtained by integrating the input $i(t)$. A capacitance C together with an ohmic resistance R across the membrane leads to an exponential decay of the membrane voltage. With additional noise $\xi(t)$, the equation of motion for the membrane voltage u reads

$$\dot{u} = -\frac{1}{\tau_{RC}}u + i(t) + \xi(t), \quad (5.30)$$

where $\tau_{RC} = RC$. Due to the linearity of Eq. (5.30), the noise $\xi(t)$ can consist of a sum of two contributions stemming from inherent (correlated) fluctuations of the membrane potential and noise in the input. Here, we consider only noisy input and assume that $\xi(t)$ is Gaussian white noise, i.e.,

$$\begin{aligned} \langle \xi(t)\xi(t') \rangle &= 2D\delta(t-t'), \\ \langle \xi(t) \rangle &= 0. \end{aligned} \quad (5.31)$$

When the membrane voltage reaches a critical value u_0 (the threshold), the neuron fires, then is reset, and the whole procedure can start all over again.

Denoting by t_n the times at which the neuron fires, the neuron exhibits a spike train of the form

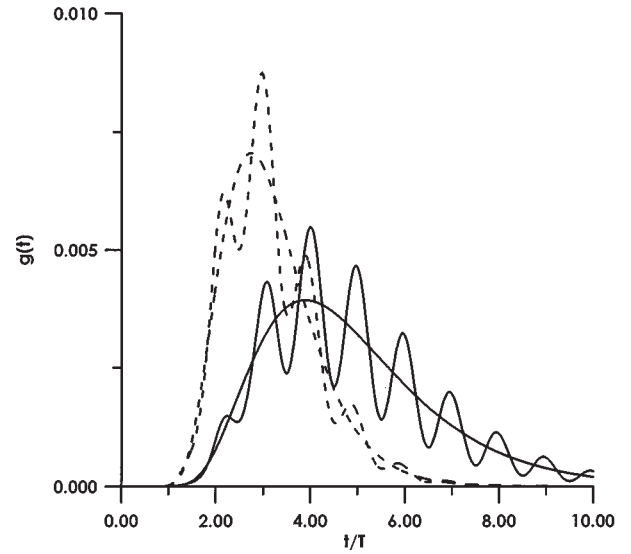


FIG. 34. The first-passage time distribution function $g(t)$ vs time measured in units of the normalized time t/T_Ω (note that in the figure t denotes the first passage time, and $T = T_\Omega$ is the forcing period). The figure contains two sets of curves, one with solid lines and another with dashed lines. The set with solid lines corresponds to a current $i_0 = 0.065 A_0$. The smooth solid curve (without multiple peaks) represents the distribution function without periodic stimulus ($A_0 = 0$) while the solid curve with the peaks corresponds to $A_0 = 0.03$. The smooth dashed curve corresponds to $A_0 = 0$ and the multi-peaked dashed curve corresponds to $A_0 = 0.03$. The other parameters (for all curves) are $b = 20$, $\Omega = 0.1$, $D = 0.2$.

$$y(t) = \sum_n h(t - t_n) = \int_{-\infty}^{\infty} y_d(t - s)h(s)ds,$$

$$y_d(t) = \sum_n \delta(t - t_n), \quad (5.32)$$

where $h(t)$ describes the (fixed) shape of a neuronal spike. Given the distribution function of the times t_n , or equivalently that of the interspike intervals $\Delta_n = t_n - t_{n-1}$, the spectral properties of the spike train can be computed by means of the theory of random point processes [see, for example, Stratonovich (1963)].

In the case of a “perfect” integrator [$1/(RC) = 0$], the Fokker-Planck equation for the membrane voltage, equivalent to Eq. (5.30), reads for a sinusoidal input $i(t) = i_0 + A_0 \cos(\Omega t)$

$$\begin{aligned} \frac{\partial P(u, t)}{\partial t} &= -[i_0 + A_0 \cos(\Omega t)] \frac{\partial P(u, t)}{\partial u} \\ &+ D \frac{\partial^2 P(u, t)}{\partial u^2}. \end{aligned} \quad (5.33)$$

The distribution function of the interspike intervals $t = \Delta$ is given by the mean-first-passage time distribution $\rho_{\text{MFPT}}(t)$ [see, for example, Hänggi *et al.* (1990)], which is obtained by solving Eq. (5.33) with absorbing boundary conditions at the threshold b , i.e., $P(u = b, t) = 0$, and

subsequent differentiation with respect to time. This task was carried out first analytically in the absence of the periodic stimulus $A_0=0$ [see also the corresponding solution for a finite leakage rate $1/(RC)$ in Goel and Richter-Dyn (1974)], and then numerically in the presence of the stimulus by Gerstein and Mandelbrot (1964). More recently, Bulsara, Lowen, and Rees (1994) and Bulsara, Elston, Doering, Lowen, and Lindenberg (1996) used an approximate image-source method to solve the Fokker-Planck equation in the presence of a weak and slowly varying sinusoidal stimulus. The features of their result follow:

- (1) The first-passage time distribution shows in the presence of the periodic stimulus a multi-peaked structure (see Fig. 34). For sufficiently large stimulus, the peaks are located at $t_{\max}^n = n T_\Omega$, with $T_\Omega = 2\pi/\Omega$ being the period of the stimulus [see also Gerstein and Mandelbrot (1964)].
- (2) The peak heights decay exponentially for increasing intervals t .
- (3) The peak heights run through a maximum as a function of the noise strength D .

This behavior resembles very closely the behavior of return-time distribution of the bistable models described in Sec. V.C.2. As yet, the theory above is based on a number of unrealistic assumptions; moreover, it contains technical difficulties that have yet to be overcome:

- (1) The phase of the sinusoidal stimulus has been reset after each firing event to the same initial value. This approximation is unrealistic from a physiological point of view, since a large amount of information about the coherence of the stimulus is eliminated. A theory of first-passage time distributions in the presence of a periodic forcing that explicitly avoids this assumption has not yet been put forward.
- (2) Since the resting voltage of the membrane of a neuron is very close to the potassium voltage, being a lower bound for the variation of the membrane voltage, an originally sinusoidal stimulus becomes strongly rectified. It is therefore not realistic simply to add the sinusoidal stimulus to the membrane voltage in the integrate-and-fire model without taking into account rectification.
- (3) Strictly speaking, the method of image sources is applicable only to diffusion processes that are homogeneous in space and time variables. The error made by using this method (as an approximation) in time-inhomogeneous equations such as Eq. (5.33) has not been estimated mathematically.

5. Neuron firing and threshold crossing

The threshold-crossing model for neuronal spiking is motivated from the *leaky* integrate-and-fire model as follows: the input $i(t)$ consists of a constant i_0 and a sinusoidal modulation $A_0 \sin(\Omega t)$. In the absence of noise, the solution of Eq. (5.30) reads for large times

$$u_\infty(t) = i_0 \tau_{RC} + \frac{A_0 \tau_{RC}}{\sqrt{1 + \Omega^2 \tau_{RC}^2}} \sin(\Omega t - \varphi_{RC}), \quad (5.34)$$

where $\tan(\varphi_{RC}) = \Omega/\tau_{RC}$. The threshold is larger than the maximum of $u_\infty(t)$, i.e., we assume a subthreshold stimulus. In the presence of the noise $\xi(t)$, the membrane voltage $u(t)$ will randomly cross the threshold. In contrast to the integrate-and-fire model, the membrane voltage is *not* reset after a threshold crossing in the models being discussed here. The threshold-crossing models are relying on a stochastic self-resetting due to the noise itself.

The simple picture we are drawing is the following: the input of our threshold trigger consists of the sum of a sinusoidal signal with amplitude $A_0 \rightarrow A_0 \tau_{RC} / \sqrt{1 + \Omega^2 \tau_{RC}^2}$ and random noise $\xi(t)$ which occasionally crosses the reduced threshold $b \rightarrow b - i_0 \tau_{RC}$. Whenever the threshold b is being crossed (at times t_n), a spike is created. This yields a stochastic spike train given in Eq. (5.32). To keep things simple, we assume a δ -shaped spike with area normalized to unity, i.e., $h(t) = \delta(t)$ in Eq. (5.32). In specific terms, we assume Gaussian-colored noise $\xi(t)$ with a zero mean and the correlation function

$$\begin{aligned} \langle \xi(t) \xi(0) \rangle = & \frac{D}{\tau_2^2 - \tau_1^2} [\tau_2 \exp(-t/\tau_2) \\ & - \tau_1 \exp(-t/\tau_1)], \end{aligned} \quad (5.35)$$

where τ_1 and τ_2 are characteristic time scales. Using the fundamental work of Rice (1948), one finds for the threshold crossing rate the central result (Jung, 1995)

$$\begin{aligned} r(t) = & \frac{1}{2\pi\sqrt{\tau_1\tau_2}} \exp\left(-\frac{(1 - \bar{A} \sin(\Omega t))^2}{2\bar{\sigma}^2}\right) \\ & \times \left[\exp\left(-\frac{\bar{A}^2 \epsilon^2 \cos^2(\Omega t)}{2\bar{\sigma}^2}\right) \right. \\ & \left. + \frac{1}{2} \bar{A} \epsilon \sqrt{\frac{2\pi}{\bar{\sigma}^2}} \cos(\Omega t) \operatorname{erfc}\left(-\frac{\bar{A} \epsilon \cos(\Omega t)}{\sqrt{2\bar{\sigma}^2}}\right) \right] \\ \equiv & \frac{1}{\sqrt{\tau_1\tau_2}} f(\bar{A}, \bar{\sigma}, \epsilon), \end{aligned} \quad (5.36)$$

with the scaled parameters $\bar{A} = A_0/b$, $\bar{\sigma}^2 = \sigma^2/b^2$, $\epsilon^2 = \Omega^2 \tau_1 \tau_2$. The periodicity reflects the encoding of the periodic input signal in the spike train. The power spectrum of the spike train has the same form as in Eq. (5.25) but has an additional term describing the spike-spike correlation function. Computing the Fourier coefficient r_1 of the periodic threshold crossing rate $r(t)$, one finds for the spectral amplification

$$\eta = \frac{4|r_1|^2}{A_0^2} = \frac{1}{\tau_1 \tau_2} \bar{\eta}(\bar{A}, \bar{\sigma}^2, \epsilon). \quad (5.37)$$

The spectral amplification η describes the encoding gain of the periodicity of the input signal in the stochastic

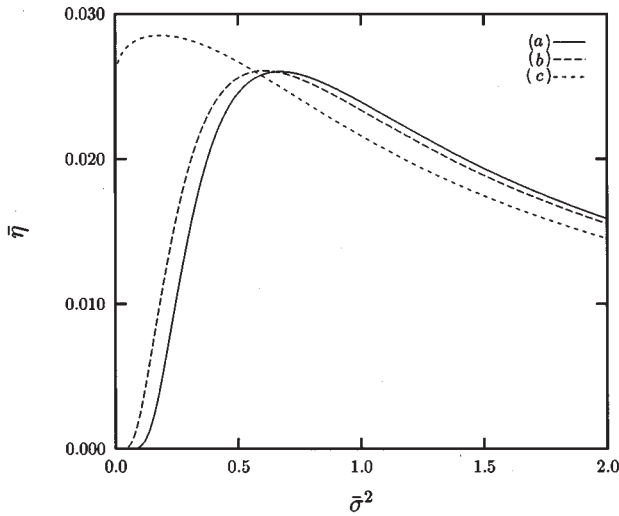


FIG. 35. The scaled spectral amplification $\bar{\eta}$, of Eq. (5.37), is shown as a function of the variance of the noise (a) for $\bar{A}=0.1$, (b) $\bar{A}=0.5$, and (c) $\bar{A}=1$. The time-scale ratio ϵ was chosen as $\epsilon=1$. Note that due to the scaling relation for η in Eq. (5.37), the actual spectral amplification can become much larger than unity (real spectral amplification) if the cutoff frequencies $1/\tau_1$ and/or $1/\tau_2$ are chosen large enough.

spike train. The scaled spectral amplification $\bar{\eta}$ is shown as a function of the variance of the noise in Fig. 35. For $A_0/b < 1$, i.e., in the subthreshold regime, the spectral amplification first increases with increasing noise, reaches a maximum at $\bar{\sigma}_{\max}^2 \approx 1/2$, and then decreases again. For small time scales $\tau_{1,2}$, the spectral amplification becomes large and describes a real encoding gain *facilitated by random noise*. The position of the maximum is obtained from Eq. (5.36) by expanding for $A_0 \tau_{RC}/D \ll 1$, i.e.,

$$\eta = \frac{1}{2} r^2 (A_0 = 0) \left(\frac{1}{\bar{\sigma}^4} + \epsilon^2 \frac{\pi}{2 \bar{\sigma}^2} \right). \quad (5.38)$$

The first term on the right-hand side of Eq. (5.38) has been obtained within the adiabatic theory of Gingl, Kiss, and Moss (1995), and also by the approach of Wiesenfeld *et al.* (1994) (see Sec. V.C.3), while the other term represents nonadiabatic corrections (Jung, 1995). For large variances of the noise, the nonadiabatic corrections become very important; they yield a decrease of the spectral amplification proportional to $1/\bar{\sigma}^2$ instead of $1/\bar{\sigma}^4$ in the adiabatic limit. The position of the peak deviates significantly from that in a driven symmetric bistable system. In the limit of vanishing frequency Ω , the spectral amplification approaches a limit curve with the maximum at approximately $\bar{\sigma}_{\max}^2 = \frac{1}{2}$. Increasing the frequency, the peak increases due to periodic jittering back and forth across the threshold, and shifts towards larger values of the variance $\bar{\sigma}^2$.

In Fig. 36, the scaled spectral amplification $\bar{\eta}$ is shown as a function of the amplitude \bar{A} . For $\bar{\sigma}^2 < \bar{\sigma}_{\max}^2$, the spectral amplification shows a maximum as a function of

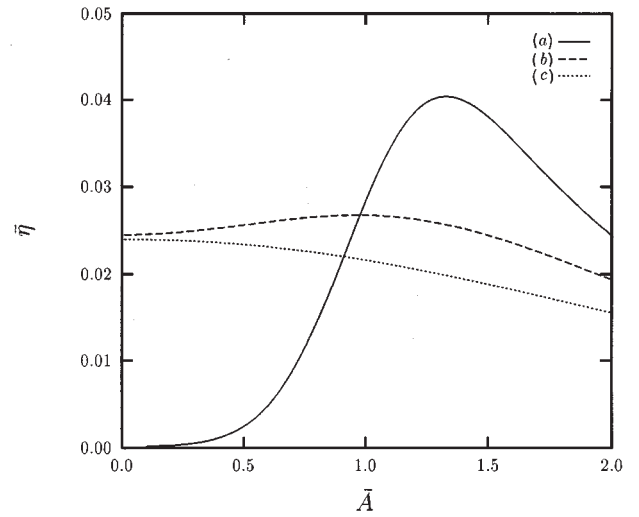


FIG. 36. The scaled spectral amplification $\bar{\eta}$ shown as a function of the amplitude of the sinusoidal signal \bar{A} (a) for $\bar{\sigma}^2=0.1$, (b) $\bar{\sigma}^2=0.5$, and (c) $\bar{\sigma}^2=1$. The time-scale ratio ϵ was chosen as $\epsilon=1$.

the amplitude, as in the quartic double-well system (Sec. IV).

In conclusion, stochastic resonance has been demonstrated experimentally in neuronal systems although these systems are not bistable. In the course of developing a theoretical understanding of these experimental results, the notion of stochastic resonance has been generalized to include excitable systems with threshold dynamics. In this latter context we refer also to related work of Collins and collaborators on aperiodic stochastic resonance (Collins *et al.*, 1995a, 1995b, Collins, Chow, *et al.*, 1996; Collins, Imhoff, and Grigg, 1996; Heneghan *et al.*, 1996), and the recent developments aimed at detecting stochastic resonance in nondynamical systems with no intrinsic sharp thresholds (Fuliński, 1995; Barzykin and Seki, 1997; Bezrukov and Vodyanoy, 1997; Jung, 1997; Jung and Wiesenfeld, 1997).

VI. STOCHASTIC RESONANCE—CARRIED ON

A. Quantum stochastic resonance

Recently, the question has been posed whether stochastic resonance manifests itself on a quantum scale. In particular, recent experiments in a macroscopic quantum system, such as a superconducting quantum interference device (SQUID), established the mechanism of stochastic resonance in the classical regime of thermal activation (Rouse *et al.*, 1995; Hibbs *et al.*, 1995). The experimental work of Rouse, Han, and Lukens (1995) also addressed nonlinear stochastic resonance, such as the noise-induced resonances, which are elucidated in Sec. VII.D.1 below. Because quantum noise persists even at absolute zero temperature, the transport of quantum information should naturally be aided by quantum fluctuations as well. Indeed, quantum mechanics

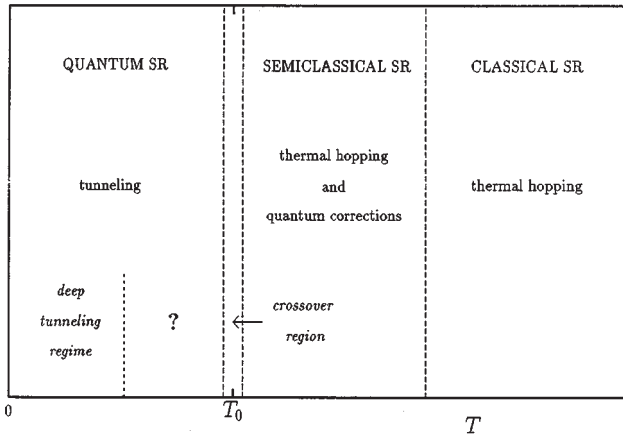


FIG. 37. The dominant escape mechanisms out of a metastable potential, and corresponding regimes for stochastic resonance depicted vs temperature T . T_0 denotes the crossover temperature below which quantum tunneling dominates over thermally activated hopping events. We note that T_0 depends on the potential shape and also on the dissipative mechanism (Hänggi *et al.*, 1985). The relative size of the corresponding stochastic resonance regions hence vary with the dissipation strength. In the region marked by a question mark, quantum stochastic resonance has presently not yet been investigated analytically; a two-level approximation is no longer adequate in that temperature regime.

provides the nonlinear system with an additional channel to overcome a threshold. This additional channel is provided by quantum tunneling, i.e., a particle can tunnel through a barrier without ever going over it. As a matter of fact, we shall see that the classical stochastic resonance effect can be assisted by quantum tunneling contributions even at finite temperatures. For strongly damped systems, such contributions can enhance the classical stochastic resonance effect up to two orders of magnitude. With decreasing temperature, quantum transitions thus start to dominate over thermally activated transitions below a crossover temperature T_0 (Hänggi *et al.*, 1985), which characterizes the temperature at which activated hopping and quantum tunneling become equally important. Its value depends both on the potential barrier shape and the dissipative mechanism. It is interesting to note that this crossover temperature can be quite large, reaching in some physical and chemical systems values larger than 100 K (Hänggi *et al.*, 1990; Hänggi, 1993). On the other hand, in Josephson systems (Schwartz *et al.*, 1985; Clarke *et al.*, 1988; Hänggi, 1993) and in mesoscopic, disordered metals (Golding *et al.*, 1992; Chun and Birge, 1993) tunneling dominates in the cold mK region only. The various escape mechanisms that predominate the physics of stochastic resonance as a function of temperature T are depicted in Fig. 37.

1. Quantum corrections to stochastic resonance

Let us first focus on the regime $T \geq T_0$, where quantum tunneling is not the dominant escape path, but nevertheless leads to significant quantum corrections. The

role of quantum tunneling in this regime has been investigated only recently by Grifoni *et al.* (1996). These authors investigated the dissipative inertial bistable quantum dynamics $x(t)$ at thermal temperature T in a double-well configuration which is modulated by the periodic force $A_0 \cos(\Omega t)$. The asymptotic power spectrum $S_{as}(\omega) = \int_{-\infty}^{\infty} \exp(-i\omega\tau) \bar{K}_{as}(\tau) d\tau$, cf. Eq. (4.24), of the time-averaged, symmetrized autocorrelation function of $x(t)$ is given by

$$S_{as}(\omega) = 2\pi \sum_{n=-\infty}^{\infty} |M_n(\Omega, A_0)|^2 \delta(\omega - n\Omega). \quad (6.1)$$

Hereby, we introduced the notation $M_n(\Omega, A_0)$ for the complex-valued Fourier amplitude to explicitly indicate the dependence on the relevant parameters. The two quantities to exhibit quantum stochastic resonance are the power amplitude η in the first frequency component of $S_{as}(\omega)$ and the ratio of η to the unperturbed, equilibrium power spectrum $S_N^0(\omega)$ of $x(t)$ in the absence of driving, evaluated at the external modulation frequency Ω , i.e. the so-called signal-to-noise ratio (SNR):

$$\eta = 4\pi |M_1(\Omega, A_0)|^2, \quad (6.2)$$

$$SNR = 4\pi |M_1(\Omega, A_0)|^2 / S_N^0(\Omega).$$

By definition, η has the dimension of a length squared, while SNR has the dimension of a frequency. Thus to investigate the interplay between noise and the coherent driving input, giving rise to the phenomenon of stochastic resonance, we shall consider two dimensionless quantities: The scaled spectral amplification $\tilde{\eta}$, and the scaled signal-to-noise ratio \widetilde{SNR} . They read

$$\tilde{\eta} = \frac{4\pi |M_1(\Omega, A_0)|^2}{(A_0 x_m^2 / V_b)^2}, \quad (6.3)$$

$$\widetilde{SNR} = \frac{[4\pi |M_1(\Omega, A_0)|^2 / S_N^0(\Omega)] / \omega_b}{(A_0 x_m / V_b)^2}.$$

Here, V_b is the barrier height at the barrier position $x_b = 0$, ω_b denotes the corresponding angular barrier frequency, and $\pm x_m$ are the positions of the two minima of the bistable potential. These two quantities that characterize stochastic resonance can be evaluated within quantum linear-response theory to give for the scaled spectral amplification

$$\tilde{\eta} = \pi \left(\frac{V_b}{k_B T} \right)^2 \frac{1}{\cosh^4(\epsilon_0 / 2k_B T)} \frac{\lambda^2}{\Omega^2 + \lambda^2}. \quad (6.4)$$

Here, k_B is the Boltzmann constant, and $\lambda = r_+ + r_-$ is the sum of the forward and backward quantum rates r_+ and r_- , respectively. These unperturbed quantum rates have been evaluated previously in the literature—see Sec. IX in the review of Hänggi, Talkner, and Borkovec (1990). A possible difference between the left and the right potential minimum is accounted for by the bias energy ϵ_0 . The backward and forward rates are related by the detailed-balance condition $r_- = r_+ \exp(-\epsilon_0 / k_B T)$. Note that information about the detailed form of the potential, and the dissipative mechanism as well, is still

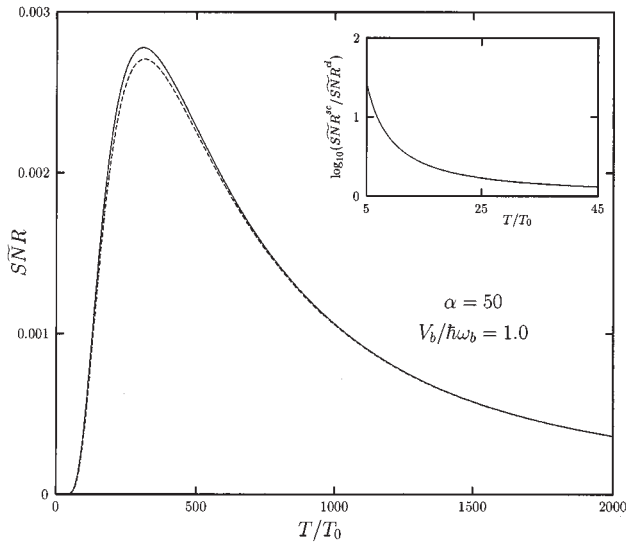


FIG. 38. The scaled signal-to-noise ratio [within linear response; see Eq. (6.5)] \widehat{SNR} of semiclassical, inertial dynamics in a symmetric double well vs the dimensionless temperature T/T_0 , with T_0 denoting the crossover temperature (see text), for ohmic friction γ of strength $\alpha = \gamma/2\omega_b = 50$. The solid line gives the dimensionless semiclassical signal-to-noise ratio with tunneling corrections. The dashed line gives the corresponding classical result without quantum corrections. The ratio between semiclassical quantum signal-to-noise ratio, \widehat{SNR} , and the corresponding classical result, \widehat{SNR}^{cl} , is depicted in the inset. Note that the tunneling contribution can enhance stochastic resonance up to two orders of magnitude. From Grifoni *et al.* (1996).

contained in the total quantum rate λ . Likewise, one finds the result for the scaled \widehat{SNR} , i.e.,

$$\widehat{SNR} = \frac{\pi}{2} \left(\frac{V_b}{k_B T} \right)^2 \frac{\lambda/\omega_b}{\cosh^2(\epsilon_0/2k_B T)}. \quad (6.5)$$

Both the zero-point energy fluctuations and the dissipative tunneling across the barrier near the barrier top result in a characteristic enhancement of the \widehat{SNR} , or the scaled spectral amplification $\tilde{\eta}$. The enhancement can reach values up to two orders of magnitude as compared to a prediction based solely on a classical analysis. In Fig. 38 we depict the quantum-tunneling-corrected, scaled \widehat{SNR} for zero bias, i.e., $\epsilon_0 = 0$, together with the enhancement over the corresponding classical study (see the inset).

The temperature dependence at different driving frequencies Ω of the scaled spectral amplification is depicted in Fig. 39 for quantum stochastic resonance in a symmetric ($\epsilon_0 = 0$) double well subjected to ohmic quantum friction. In presence of tunneling, the role of both temperature and dissipation must be treated simultaneously in a manner consistent with the fluctuation-dissipation theorem (Grifoni *et al.*, 1996). In particular, with strong damping the effects of quantum fluctuations on stochastic resonance can extend well above the crossover temperature T_0 . As depicted in the inset of Fig. 39, the stochastic resonance peak is dominated by the two

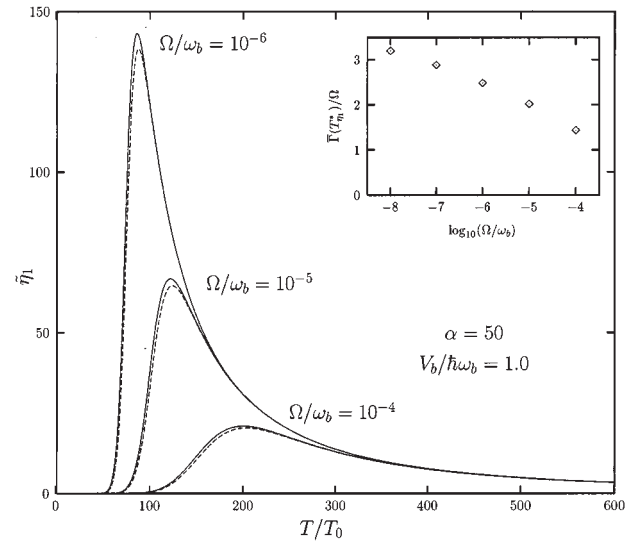


FIG. 39. Scaled spectral amplification $\tilde{\eta}$ [see Eq. (6.4)] in a symmetric double well vs dimensionless temperature (cf. Fig. 38) for different driving frequencies Ω (solid lines). For comparison, the dashed lines give the results for the classical stochastic resonance spectral amplification. The dimensionless ohmic friction strength is $\alpha = \gamma/2\omega_b = 50$. The inset depicts the ratio between the total (forward and backward) rate $\Gamma \equiv \lambda$ and Ω at the temperature T_η^* where $\tilde{\eta}$ assumes its maximum. The stochastic resonance maximum is thus approximately determined by the condition that twice the escape rate, i.e. $\Gamma(T_\eta^*)$, approximately equals the external driving frequency Ω . After Grifoni *et al.* (1996).

competing effects of an increasing Arrhenius factor and a decreasing factor $(k_B T)^2$ with increasing noise temperature. These two quantities characteristically rule the stochastic resonance effects—see Eqs. (4.51) and (4.54).

2. Quantum stochastic resonance in the deep cold

The situation changes drastically when we proceed towards the extreme cold. Here, we shall focus on the deep quantum regime, where tunneling is the only channel for barrier crossing. In this low-temperature regime, periodic driving induces several new interesting, counterintuitive physical phenomena, such as “coherent destruction of tunneling” (Grossmann *et al.*, 1991), the “stabilization of dissipative coherence” with increasing temperature (Dittrich *et al.*, 1993; Oelschlägel *et al.*, 1993), or the effect of driving-induced quantum coherence (Grifoni *et al.*, 1995). Quantum stochastic resonance within the regime of incoherent tunneling transitions at adiabatic driving frequencies has been investigated first by Löfstedt and Coppersmith (1994a, 1994b) in the context of impurity tunneling in ac-driven mesoscopic metals (Golding *et al.*, 1992; Chun and Birge, 1993; Coppinger *et al.*, 1995). Linear response for quantum stochastic resonance as well as nonlinear quantum stochastic resonance has been investigated in the whole parameter range by Grifoni and Hänggi (1996a, 1996b); their studies encompass adiabatic and nonadiabatic driving frequencies, as well as the role of both an

incoherent (i.e., rate-dominated relaxation) and coherent (i.e., oscillatory-dominated damped relaxation) tunneling dynamics.

In the regime below the crossover regime, i.e., the regime marked by a question mark (?) in Fig. 37, there exist at present no analytical studies of quantum stochastic resonance. This is due mainly to the fact that in this regime the dissipation-driven tunneling dynamics involve many states. At very low temperatures, however, the dynamics are ruled mainly by two tunnel-split levels only. Thus in the deep quantum regime the investigation of quantum stochastic resonance reduces to the study of the dynamics of the spin-boson system in the presence of ohmic dissipation which, in addition, is subjected to periodic driving (Grifoni *et al.*, 1993, 1995; Dakhnovskii and Coalson, 1995; Makarov and Makri, 1995; Goychuk, Petrov, and May, 1996; Makri, 1997). More explicitly, let us consider the driven spin-boson Hamiltonian $H = H_{\text{TLS}}(t) + H_B$, i.e.,

$$H_{\text{TLS}}(t) = -\frac{\hbar}{2}(\Delta\sigma_x + \epsilon_0\sigma_z) - \frac{\hbar\hat{\epsilon}}{2}\cos(\Omega t)\sigma_z \quad (6.6)$$

represents the driven bistable system in a two-level-system approximation with $(\hbar\hat{\epsilon}/a)\cos(\Omega t)$ being the applied harmonic force. The σ 's are Pauli matrices, and the eigenstates of σ_z are the basis states in a localized representation, while $a \equiv 2x_m$ is the tunneling distance. The tunneling splitting energy of the symmetric two-level system is given by $\hbar\Delta$ while the bias energy is again $\hbar\epsilon_0$. Within the spin-boson model (Leggett *et al.*, 1987; Weiss, 1993), the environment is modeled by a term H_B describing an ensemble at thermal temperature T of harmonic oscillators. The term H_B in addition includes the interaction between the two-level system and the bath via a bilinear coupling in the two-level system-bath coordinates. The effects of the bath are captured by the spectral density $J(\omega)$ of the environment coupling. We make the specific choice of ohmic dissipation $J(\omega) = (2\pi\hbar/a^2)\alpha\omega e^{-\omega/\omega_c}$, where α denotes the dimensionless ohmic coupling strength and $\omega_c \gg \omega_0$ is a cutoff frequency. Insightful exact numerical path-integral studies of this driven, dissipative spin-boson system have been carried out recently by Makri (1997).

The relevant theoretical quantity describing the dissipative dynamics under the external perturbation is the expectation value $P(t) = \langle \sigma_z(t) \rangle$. On the other hand, the quantity of experimental interest for quantum stochastic resonance is the time-averaged power spectrum $S(\omega)$, defined as the Fourier transform of the correlation function

$$\bar{K}(\tau) = \frac{\Omega}{2\pi} \int_0^{2\pi/\Omega} dt \frac{1}{2} \langle \sigma_z(t+\tau)\sigma_z(t) + \sigma_z(t)\sigma_z(t+\tau) \rangle.$$

The combined influence of dissipative and driving forces renders an evaluation of the full correlation function $\bar{K}(\tau)$ extremely difficult (and hence of the power spectrum). Matters simplify for times t, τ large compared to the time scale of the transient dynamics, where $P(t)$ and

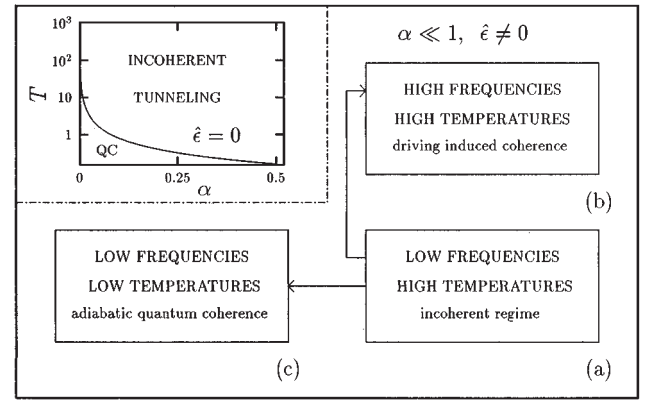


FIG. 40. Sketch of the different regimes for the driven $[(\hbar\hat{\epsilon}/a)\cos\Omega t]$ dissipative tunneling dynamics for a two-level system subject to weak ohmic coupling $\alpha \equiv \gamma(a^2/2\pi\hbar)$ (a denotes the tunneling distance and γ denotes the Ohmic viscous strength). As the temperature T or the driving angular frequency Ω are varied in the parameter space of “TEMPERATURE” and “FREQUENCY” [see the boxes labeled (a), (b), (c)], different novel tunneling regimes are encountered. For strong dimensionless ohmic coupling α , the regimes (a) and (b) extend down to the lowest temperatures. For comparison we also depict the static situation, i.e., $\hat{\epsilon} = 0$, in the upper left panel: in it, we sketch the dissipative tunneling behavior in the (T, α) parameter space for quantum *incoherent* and quantum *coherent* (QC) tunneling.

$\bar{K}(\tau)$ acquire in the asymptotic regime the periodicity of the external perturbation. Upon expanding the asymptotic expectation $P_{as}(t) = \lim_{t \rightarrow \infty} P(t)$ into a Fourier series, i.e.,

$$P_{as}(t) = \sum_{n=-\infty}^{\infty} M_n(\Omega, \hat{\epsilon}) \exp(-in\Omega t), \quad (6.7)$$

it is readily seen that the amplitudes $|M_n(\Omega, \hat{\epsilon})|$ determine the weights of the δ spikes of the power spectrum in the asymptotic state $S_{as}(\omega)$ via the relation $S_{as}(\omega) = 2\pi \sum_{n=-\infty}^{\infty} |M_n(\Omega, \hat{\epsilon})|^2 \delta(\omega - n\Omega)$. In particular, to investigate nonlinear quantum stochastic resonance, we shall examine the newly scaled power amplitude η_n in the n th frequency component of $S_{as}(\omega)$, i.e.,

$$\eta_n = 4\pi |M_n(\Omega, \hat{\epsilon})/\hbar\hat{\epsilon}|^2. \quad (6.8)$$

For a quantitative study of quantum stochastic resonance, it is necessary to solve the asymptotic dynamics of the nonlinearly driven dissipative bistable system. In doing so, we shall take advantage of novel results for the driven dynamics obtained by use of a real-time path-integral approach (Grifoni *et al.*, 1993; 1995). At weak and strong ohmic coupling, driving distinctly alters the qualitative tunneling dynamics: see Fig. 40. The incoherent dynamics can still be modeled by rate equations, however. At low driving frequencies, these rate equations are intrinsically Markovian—see region (a) in Fig. 40. As the external frequency Ω is increased and/or when the temperature is lowered, quantum coherence and/or driving-induced correlations render the

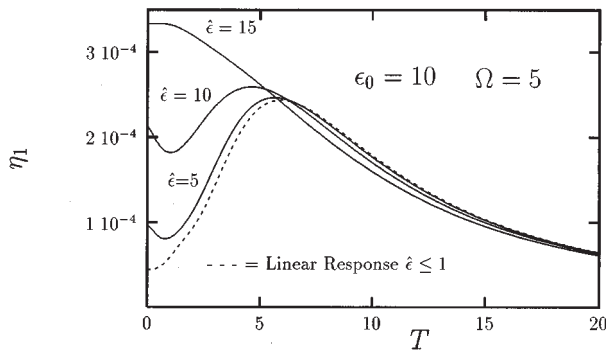


FIG. 41. The spectral amplification η_1 , Eq. (6.8), for the periodically driven (driving strength $\hat{\epsilon}$), ohmic damped spin-boson system depicted vs temperatures T at a fixed bias energy $\hbar\epsilon_0=10$ and at an angular driving frequency $\Omega=5$. The dimensionless parameters are defined in the text. For the smallest driving strength $\hat{\epsilon}=5$ we additionally depict the linear-response theory approximation (dashed line). The solid lines give the nonlinear quantum stochastic resonance for the exactly solvable case of ohmic coupling strength at $\alpha=1/2$ (cf. Fig. 40). The data are taken from Grifoni and Hänggi (1996b).

asymptotic dynamics intrinsically non-Markovian—note regions (b) and (c) in Fig. 40.

Let us focus first on some characteristics of quantum stochastic resonance as they emerge from the study of the case $\alpha=1/2$ of the ohmic strength. For the special value $\alpha=1/2$, exact analytical solutions for quantum stochastic resonance are possible (Grifoni and Hänggi, 1996a, 1996b). The resulting fundamental power amplitude $\eta_1 \equiv \eta$ is plotted in Fig. 41 as a function of the temperature for different driving strengths $\hat{\epsilon}$. There, and in Fig. 42, frequencies are in units of the bath-renormalized tunneling splitting $\Delta_e = \Delta(\Delta/\omega_c)^{\alpha/(1-\alpha)}[\cos(\pi\alpha)\Gamma(1-2\alpha)]^{1/(2-2\alpha)}$, $\alpha < 1$. For $\alpha=1/2$, Δ_e reduces to $\pi\Delta^2/2\omega_c$. For $\alpha > 1$, relevant energy scale is set by Δ_{ren} , which equals the previous expression Δ_e without the term in the square brackets (Leggett *et al.*, 1987; Weiss, 1993; Löfstedt and Coppersmith, 1994a). The temperatures are in units of $\hbar\Delta_e/k_B$. Note that the spectral amplification is measured in units of $(\hbar\Delta_e)^{-2}$. For highly nonlinear driving fields $\hat{\epsilon} > \epsilon_0$, the power amplitude decreases monotonically as the temperature increases (uppermost curve). As the driving strength $\hat{\epsilon}$ of the periodic signal is decreased, a shallow minimum, followed by a maximum, appears when the static asymmetry ϵ_0 equals, or slightly overcomes, the strength $\hat{\epsilon}$ (intermediate curves). For even smaller external amplitudes, quantum stochastic resonance can be studied within the quantum linear-response theory (dashed curve in Fig. 41). In the linear-response region the shallow minimum is washed out, and only the principal maximum survives. It is now interesting to observe that—because the undriven two-level system dynamics (which comprises linear-response theory) for $\alpha=1/2$ is always incoherent down to $T=0$ —the principal maximum arises at the temperature T at which the relaxation pro-

cess towards thermal equilibrium is maximal. On the other hand, the minimum in η appears in the temperature region where driving-induced coherent processes are of importance. This latter feature is a nonlinear quantum stochastic resonance effect, which linear-response theory clearly cannot describe. In addition, this implies that the power amplitude η plotted versus frequency shows resonances when $\Omega \approx \epsilon_0/n$ ($n=1,2,\dots$); correspondingly, the driven dissipative dynamics are intrinsically non-Markovian! For arbitrary values of the ohmic coupling strength, one has to resort to approximate solutions of the dissipative dynamics. For strong coupling $\alpha > 1$, or weak coupling $\alpha < 1$ and high enough temperatures, the bath-induced correlations between tunneling transitions may be treated within the noninteracting blip approximation (Leggett *et al.*, 1987; Weiss, 1993). A set of coupled equations for the Fourier coefficients M_n can then be derived for any strength and frequency of the driving force. In particular, driving-induced correlations may result in an highly coherent dynamics, leading to resonances in the power spectrum. In this coherent regime, quantum stochastic resonance never occurs: the power amplitudes η_n always show a monotonic decay as the temperature is increased (Grifoni and Hänggi, 1996a, 1996b). It is only in the low-frequency regime $\hbar\Omega \ll \alpha k_B T$ that—to leading order—driving-induced non-Markovian correlations do not contribute. The asymptotic dynamics, within the noninteracting blip approximation, are intrinsically incoherent and governed by the rate equation $\dot{P}_{as}(t) = -\lambda(t) \times [P_{as}(t) - P_{eq}(t)]$, with time-dependent rate $\lambda(t) = \text{Re} \Sigma[\epsilon(t)]$ and equilibrium value $P_{eq}(t) = \tanh[\hbar\epsilon(t)/2k_B T]$. Here, $\epsilon(t) = \epsilon_0 + \hat{\epsilon} \cos \Omega t$ plays the role of a time-dependent adiabatic asymmetry, and the rate $\lambda(t)$ ($\alpha < 1$) is obtained as the real part of

$$\Sigma[\epsilon(t)] = \frac{\Delta_e}{\pi} \left(\frac{\beta\Delta_e}{2\pi} \right)^{1-2\alpha} \frac{\Gamma(\alpha + i\hbar\beta\epsilon(t)/2\pi)}{\Gamma(1-\alpha + i\hbar\beta\epsilon(t)/2\pi)}, \quad (6.9)$$

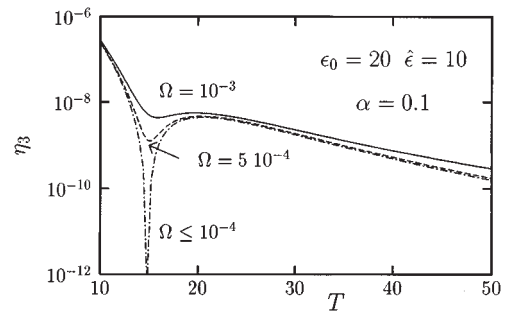


FIG. 42. Quantum noise-induced resonances for the third-order amplitude $\eta_3 = 4\pi|M_3(\Omega, \hat{\epsilon})/\hbar\hat{\epsilon}|^2$ vs temperature T in the regime of adiabatic incoherent tunneling. The different lines are for three different driving frequencies Ω . The noise-induced suppression which characterizes the “resonance” sharpens with decreasing driving frequency Ω . The results are for a bias energy $\epsilon_0=20$, driving strength $\hat{\epsilon}=10$, and a coupling strength of $\alpha=0.1$. Data are from Grifoni and Hänggi (1996b).

where $\Gamma(z)$ denotes the gamma function, and β is the inverse temperature ($1/k_B T$).

The rate equation can then be solved in terms of quadratures, and the nonlinear low-frequency power spectrum can be investigated. Quantum stochastic resonance indeed occurs in this incoherent tunneling regime. As for the case $\alpha = \frac{1}{2}$, the quantum stochastic resonance maximum appears only when the static asymmetry ϵ_0 overcomes the external strength $\hat{\epsilon}$. Moreover, Fig. 42, which shows the behavior of the third scaled spectral amplification η_3 versus temperature, reveals another striking effect: As the driving frequency is decreased, a noise-induced suppression of higher harmonics occurs in correspondence to the stochastic resonance maximum in the fundamental harmonic. In this regime one finds the quantum analogue of noise-induced resonances that characterize classical nonlinear stochastic resonance (see Sec. IV.A). A numerical evaluation shows that the noise-induced suppression indeed appears when $\Omega \ll \min \lambda(t)$, so that the quasistatic expression holds, i.e.,

$$M_n = \frac{1}{2\pi} \int_0^{2\pi} dx \tanh[\hbar \beta (\epsilon_0 + \hat{\epsilon} \cos x)/2] \cos(mx). \quad (6.10)$$

In contrast to classical stochastic resonance, where the enhancement is maximal for symmetric bistable systems, we find that a nonzero bias is necessary for quantum stochastic resonance. To understand this behavior, one can investigate the predictions for quantum stochastic resonance within a linear-response approach. In this case, only the harmonics $0, \pm 1$ of $P_{as}(t)$ in Eq. (6.7) are different from zero. In particular, P_0 becomes the thermal equilibrium value P_{eq} of the operator σ_z in the absence of driving, and $M_{\pm 1} = \hbar \hat{\epsilon} \chi(\pm \Omega)$ is related by Kubo's formula to the linear susceptibility $\tilde{\chi}_{xx}(\Omega) = a^2 \chi(\Omega)$ for the particle position $x = (a/2)\sigma_z$, where

$$\tilde{\chi}_{xx}(\Omega) = \frac{i}{\hbar} \int_{-\infty}^{+\infty} d\tau \exp(-i\Omega\tau) H(\tau) \times \langle [x(\tau), x(0)] \rangle_\beta. \quad (6.11)$$

Here, $H(\tau)$ is the Heaviside function, $[\dots, \dots]$ denotes the commutator, and $\langle \dots \rangle_\beta$ the thermal statistical average of the full system in the absence of the external periodic force ($\hat{\epsilon} = 0$). In the regime where incoherent transitions dominate, the dynamical susceptibility is explicitly obtained in the form

$$\chi(\Omega) = \frac{1}{4k_B T} \frac{1}{\cosh^2(\epsilon_0/2k_B T)} \frac{\lambda}{\lambda - i\Omega}. \quad (6.12)$$

The quantity $\lambda = \text{Re} \Sigma(\epsilon_0)$, $\alpha < 1$ [for $\alpha > 1$, see Hänggi, Talkner, and Borkovec (1990), or Weiss (1993)] is the sum of the forward and backward (static) quantum rates out of the metastable states, r_+ and r_- , respectively. The factor $1/\cosh^2(\hbar\epsilon_0/2k_B T)$ expresses the fact that the two rates are related by the detailed balance condition

$r_+ = e^{\beta \hbar \epsilon_0} r_-$. It is now interesting to note that formally the *same* expression for the incoherent susceptibility (and hence for η) holds true for the classical case, with r_+ and r_- denoting the corresponding classical forward and backward Kramers rates. Thus in classical stochastic resonance, the maximum arises because of competition between the thermal Arrhenius dependence of these rates and the algebraic factor $(k_B T)^{-1}$ that enters the linear susceptibility, and this maximum occurs (roughly) at a temperature that follows from the matching between the frequency scales of the thermal hopping rate and the driving frequency. Detailed investigation reveals that quantum stochastic resonance characteristically occurs when incoherent tunneling contributions dominate over coherent tunneling transitions. Moreover, in clear contrast to classical stochastic resonance, and also to semiclassical stochastic resonance near and above T_0 (see above), quantum stochastic resonance in the deep cold occurs only in the presence of a finite asymmetry $\epsilon_0 \neq 0$ between forward and backward escape paths. Thus while classical and semiclassical stochastic resonance is maximal for zero bias [see Eqs. (4.55), (4.56)] the quantum stochastic resonance phenomenon vanishes in the deep quantum regime when the symmetry between forward and backward dissipative tunneling transitions is approached. What is the physics that governs this behavior? Clearly, with decreasing temperature the thermal, exponential-like Arrhenius factor no longer dominates the escape rates; rather, its role is taken over by the action of the tunneling paths that govern adiabatic and nonadiabatic tunneling—see Sec. IX in Hänggi, Talkner, and Borkovec (1990). This non-Arrhenius action term possesses a rather weak temperature dependence as compared to the Arrhenius dependence. Hence the crucial factor in quantum stochastic resonance is not this exponential action part governing the quantum rate behavior but rather the Arrhenius-like detailed balance factor relating the forward rate to the backward rate [see below Eq. (6.12)]. This exponential detailed balance factor contains the energy scale $(\hbar\epsilon_0/k_B T)$; thus it is this exponential dependence that crucially competes with the algebraic factor $(k_B T)^{-1}$ that enters the linear susceptibility. Whenever $\hbar\epsilon_0 \ll k_B T$, the energy levels are essentially equally occupied; hence with $\epsilon_0 = 0$ no quantum stochastic resonance peaks occur!

The second consequence is that over a wide range of driving frequencies the stochastic resonance maximum arises at a temperature obeying $k_B T \approx \hbar\epsilon_0$. Similar qualitative results, together with the occurrence of noise-induced suppression are obtained also in the parameter region of low temperatures $kT \leq \hbar\Delta$ and weak coupling $\alpha \ll 1$, where overdamped quantum coherence occurs. In this regime, the noninteracting blip approximation fails to predict the correct long-time behavior. This is so because the neglected bath-induced correlations contribute to the dissipative effects to first order in the coupling strength. Nevertheless, a perturbative treatment allows an investigation of quantum stochastic resonance even in this regime.

In conclusion, quantum noise in the presence of periodic driving can substantially enhance or suppress quantum stochastic resonance. Note also that all of this discussion of quantum stochastic resonance constitutes a situation where inertial effects in stochastic resonance (i.e., finite ohmic friction strengths α) are always implicitly accounted for. Of particular relevance is the occurrence of noise-induced suppression within nonlinear quantum stochastic resonance. This phenomenon of signal suppression at higher harmonics can be used for a distortion-free spectral amplification of information in quantum systems. Experimental candidates to observe these novel quantum stochastic resonance effects are the above-mentioned mesoscopic metals (Golding *et al.*, 1992; Chun and Birge, 1993; Löfstedt and Coppersmith, 1994a, 1994b; Coppinger *et al.*, 1995), ac-driven SQUID systems (Hibbs *et al.*, 1995; Rouse *et al.*, 1995), as well as ac-driven atomic force microscopy (Eigler and Schweitzer, 1990; Louis and Sethna, 1995), or ac-modulated proton tunneling (Grabert and Wipf, 1990; Benderskii *et al.*, 1994).

B. Stochastic resonance in spatially extended systems

So far we mainly investigated stochastic resonance in systems with only one degree of freedom, such as a particle moving in a potential under the influence of an external driving force and noise. In this section we describe how stochastic resonance manifests itself in spatially extended systems such as a string moving in a bistable potential under the influence of noise and external forcing, or in a two-dimensional medium forming spatiotemporal patterns in the presence of noise.

1. Global synchronization of a bistable string

In this section, we consider a one-dimensional bistable medium in the presence of noise and isotropic periodic forcing. The model is described by the one-dimensional Ginzburg-Landau equation (Benzi *et al.*, 1985):

$$\frac{\partial \Phi(x, t)}{\partial t} = m\Phi(x, t) - \Phi^3(x, t) + \kappa \frac{\partial^2 \Phi(x, t)}{\partial x^2} + A_0 \cos(\Omega t) + \xi(x, t), \quad (6.13)$$

where $\xi(x, t)$ is white Gaussian noise in both time and space, i.e.,

$$\begin{aligned} \langle \xi(x, t) \xi(x', t') \rangle &= 2D \delta(t - t') \delta(x - x') \\ \langle \xi(x, t) \rangle &= 0. \end{aligned} \quad (6.14)$$

In the absence of driving, Eq. (6.13) can be cast in the form

$$\frac{\partial \Phi(x, t)}{\partial t} = - \frac{\delta V[\Phi]}{\delta \Phi} + \xi(x, t), \quad (6.15)$$

with the functional $V[\Phi]$ (Rajaraman, 1982)

$$V[\Phi] = \int_0^L \left[\frac{1}{4} \Phi^4 - \frac{1}{2} m \Phi^2 + \frac{\kappa}{2} \left(\frac{\partial \Phi}{\partial x} \right)^2 \right] dx. \quad (6.16)$$

The stationary solutions $\Phi(x)$ in the absence of the noise, obeying

$$m\Phi - \Phi^3 + \kappa \frac{d^2 \Phi}{dx^2} = 0, \quad (6.17)$$

with von Neumann boundary conditions

$$\frac{d\Phi(0)}{dx} = \frac{d\Phi(L)}{dx} = 0, \quad (6.18)$$

extremalize the functional $V[\Phi]$. Equation (6.17) can be interpreted as a Newtonian equation of motion in the inverse double-well potential $U_N(\Phi) = -\frac{1}{4}\Phi^4 + \frac{1}{2}m\Phi^2$ with x as the time variable. The relevant homogeneous stationary solutions (stationary solution in the Newtonian picture) with boundary conditions (6.18) are given by

$$\begin{aligned} \Phi_{\pm} &= \pm \sqrt{m}, \\ \Phi_0 &= 0, \end{aligned} \quad (6.19)$$

with

$$\begin{aligned} V[\Phi_+] &= V[\Phi_-] = -\frac{1}{4}m^2L, \\ V[\Phi_0] &= 0. \end{aligned} \quad (6.20)$$

The first two solutions (the whole string is sitting in one of the potential minima) are stable and the third one (the whole string is sitting on the barrier top) is unstable. There is also a class of stable inhomogeneous solutions, the multi-instanton solutions $\Phi^{(k)}$, that obey the boundary conditions (6.18) and have k zeros in the interval $[0, L]$. For the potential energies $V[\Phi^{(k)}]$ one finds the inequalities

$$V[\Phi^{(0)}] < V[\Phi^{(1)}] < V[\Phi^{(2)}] \dots, \quad (6.21)$$

with $\Phi^{(0)} = \Phi_{\pm}$. In the presence of noise, the string can escape out of the stable homogeneous states Φ_{\pm} . A generalization of Ventsel and Freidlin's theory (Benzi *et al.*, 1985) yields for the mean exit times in the weak-noise limit $D \ll \Delta V$

$$T_{\pm} = C \exp\left(\frac{2\Delta V}{D}\right), \quad (6.22)$$

with C independent of D . Here, $\Delta V = V[\Phi^{(2)}] - V[\Phi_{\pm}]$ denotes the smallest $V[\Phi]$ barrier that separates the two stable states Φ_{\pm} . This conclusion was verified numerically by solving the discretized version of Eq. (6.13) (see below). In the presence of a weak and slow homogeneous external forcing ($A = A_0 \ll m^{3/2}$), Benzi *et al.* (1985) derived analytical expressions for the mean exit times T_+^a and T_+^b , relevant to the transition $\Phi_+ \rightarrow \Phi_-$ in the potential configurations with $\cos(\Omega t) = \pm 1$, i.e.,

$$\begin{aligned} T_+^a &= C \exp[2(\Delta V - A_0 L \sqrt{m})/D], \\ T_+^b &= C \exp[2(\Delta V + A_0 L \sqrt{m})/D]. \end{aligned} \quad (6.23)$$

The system shows stochastic resonance if one of the exit times of Eq. (6.23) is shorter than the half driving

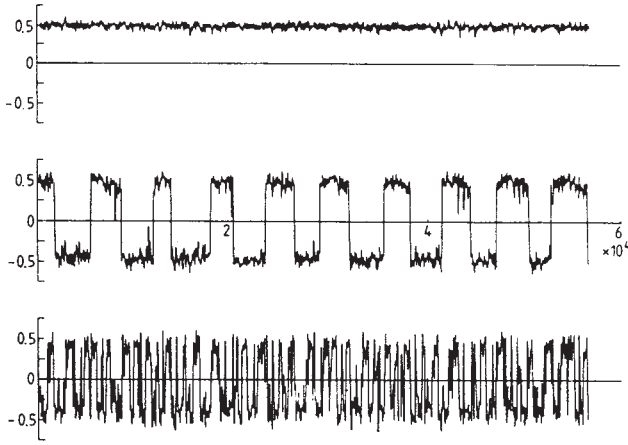


FIG. 43. The collective coordinate $u = \int \phi(x,t)dx/L$ plotted against time for $D/\Delta x = 0.002$ (upper plot), $D/\Delta x = 0.07$ (middle plot), and $D/\Delta x = 0.11$ (lower plot). The parameters are $L=1$, $N=20$, $m=0.25$, $A=0.0125$, $\omega=0.01\pi/3$, and $B=1/64$. These parameters imply a barrier height of $1/64$ and an Arrhenius factor of 4.457 .

period and the other is longer, so that on average the exit times are of the order the half driving period (the factor of 2 stems from the fact that the string has to escape twice within one period of the driving). This yields an upper and a lower bound for the optimal noise strength with arithmetic mean

$$D_{SR} = \frac{2\Delta V}{\ln(\pi/C\Omega)}. \quad (6.24)$$

A more refined analysis of stochastic resonance in a modulated string has been derived recently by Marchesoni, Gammaitoni, and Bulsara (1996) within the framework of the theory of thermal nucleation in one-dimensional chains (Hänggi, Marchesoni, and Sodano, 1988). The ensuing theoretical predictions have been verified experimentally by Löcher, Johnson, and Hunt (1996).

A full numerical investigation of the discretized Ginzburg-Landau equation has been carried out by Benzi *et al.* (1985). The discretized Ginzburg-Landau equation [Eq. (6.13)] reads

$$\begin{aligned} \dot{\psi}_n(t) = & m\psi_n - \psi_n^3 + \frac{\kappa}{(\Delta x)^2} [\psi_{n+1}(t) + \psi_{n-1}(t) \\ & - 2\psi_n(t)] + A_0 \cos(\Omega t) + \sqrt{D/\Delta x} \xi_n(t), \end{aligned} \quad (6.25)$$

with discretization step Δx , string sites $\psi_n(t) = \Phi(x_n, t)$, and

$$\begin{aligned} \langle \xi_n(t) \xi_m(t') \rangle &= 2\delta_{nm} \delta(t-t'), \\ \langle \xi_n(t) \rangle &= 0. \end{aligned} \quad (6.26)$$

In Fig. 43, the string collective coordinate $u = (1/L) \int_0^L \Phi(x) dx$ is shown as a function of time at three different levels of the homogeneous noise intensity. For a properly chosen value of the noise level D_{SR}

(second plot), the collective coordinate switches almost periodically between \sqrt{m} and $-\sqrt{m}$, i.e., between states where the entire string is either in the right or in the left potential well—the noise has globally synchronized the hopping along the bistable string. A similar conclusion has been reached recently by Lindner *et al.* (1995), who simulated numerically the same discretized Ginzburg-Landau equation [Eq. (6.13)]. Here, the different notation $\kappa/(\Delta x)^2 \rightarrow g$ and $D/\Delta x \rightarrow \epsilon$ hides the underlying Ginzburg-Landau equation. Using $L = N\Delta x$, the scaling relations $g \propto N^2$ and $\epsilon \propto N$, which were derived in Marchesoni, Gammaitoni, and Bulsara (1996), follow immediately. These authors also noticed that, while jumping back and forth between the stable configurations Φ_{\pm} , a long string develops a remarkable spatial periodicity, which attains its maximum at resonance. The relevant spatial correlation length can be easily estimated within the thermal nucleation theory. Related to these studies of stochastic resonance in extended systems is the recent study by Wio (1996) on stochastic resonance in a bistable reaction-diffusion system, or the investigation of stochastic resonance in weakly perturbed Ising models (Néda, 1995a, 1995b; Brey and Prados, 1996; Schimansky-Geier and Siewert, 1997).

2. Spatiotemporal stochastic resonance in excitable media

Pattern formation in excitable media is an important paradigm with many applications in biology and medicine such as contraction waves in cardiac muscle, slime mold aggregation patterns, and cortical depression waves, to name only a few (for an overview, see Murray, 1989). While most theoretical and experimental work on excitable media focuses on the propagation of spiral waves, the role of fluctuations for pattern selection and propagation has been studied only recently (Jung and Mayer-Kress, 1995) by using a stochastic cellular model. One of the many interesting features of noisy media is that spatiotemporal structures and coherence can strongly vary with the noise level. Spatiotemporal stochastic resonance describes the enhancement of a spatiotemporal pattern (externally applied or intrinsic) by an optimal dose of noise.

The model of Jung and Mayer-Kress consists of a square array of excitable threshold elements with lattice constant a . Each element e_{ij} can assume three states: the quiescent state, the excited state, and a subsequent refractory state. The state of each element e_{ij} is controlled by an input $x_{ij}(t)$. If the input $x_{ij}(t)$ is below a threshold b , the element is quiescent. If $x_{ij}(t)$ is crossing a threshold from below, the element switches into the excited state, i.e., it fires. The inputs $x_{ij}(t)$ are coupled to a homogeneous thermal environment, i.e., the time dependence is described by the Langevin equation

$$\dot{x}_{ij} = -\gamma x_{ij} + \sqrt{\gamma\sigma^2} \xi_{ij}(t), \quad (6.27)$$

with $\langle x_{ij}^2 \rangle = \sigma^2$ and zero-mean, uncorrelated noise in space and time $\langle \xi_{ij}(t) \xi_{kl}(t') \rangle = 2\delta_{(ij),(kl)} \delta(t-t')$. The excitable elements communicate via pulse coupling. When an element e_{kl} fires, it emits a spike that is re-

ceived by an element e_{ij} with an intensity depending on the distance $r_{(ij),(kl)}$ between e_{kl} and e_{ij} . Element e_{ij} integrates the incoming spikes from all firing elements yielding after one time step Δt (the smallest time scale of our system) the additional input f_{ij}

$$f_{ij} = K \sum_{kl} \exp\left(-\lambda \frac{r_{(ij),(kl)}^2}{a^2}\right), \quad (6.28)$$

which is added to x_{ij} . The parameter λ describes the inverse range of the interaction and the parameter K the coupling strength. The medium is updated synchronously in time steps of the smallest time scale Δt , the time interval of firing. All other time scales are measured in units of Δt . The proper normalization of this model is given by $\bar{\sigma}^2 = \sigma^2/b^2$, $\bar{\gamma} = \gamma\Delta t$, $\bar{K} = K/b$. The time step and the threshold are therefore normalized to unity. The dissipation constant $\bar{\gamma}$ defines the typical time scale of the temporal evolution of a single element. For large dissipation ($\bar{\gamma} > 1$), the element forgets its prehistory within one time step of temporal evolution, while for small dissipation $\bar{\gamma} \ll 1$, the system—as a whole—can build up a long memory.

It has been demonstrated that this model shows for large coupling \bar{K} (in the absence of noise) the typical excitation patterns of excitable media, i.e., rotating spiral waves or target waves, usually described in terms of reaction diffusion equations with two species (Murray, 1989). In the presence of noise, the typical excitation patterns can still be observed, but they exhibit rough wave fronts and—depending on the noise level—more serious imperfections such as breakup of wave fronts and collisions with noise-nucleated waves. The overall picture in the strong-coupling regime is the coexistence of multiple finite-sized cells with coherent patterns.

For weak coupling \bar{K} , however, the discrete nature of the model becomes important and different phenomena can be observed. To maintain a firing pattern, the coupling \bar{K} has to exceed a critical value \bar{K}_0 , which is estimated for small λ and negligible curvature effects as follows: an infinite front of firing elements reduces the firing threshold of an element next to the front by an amount S_0 which is the sum of the contributions from all firing elements along the front. The element, however, is precharged by the sum of the contributions \bar{S}_{pre} of firing elements of the front at earlier times (and larger spatial distances). At the critical coupling, the sum of the precharge \bar{S}_{pre} and \bar{S}_0 of an element right before the front is unity (the normalized threshold), i.e.,

$$\bar{K}_0(\gamma) \approx \sqrt{\frac{\lambda}{\pi}} \frac{1}{\exp(-\lambda) + \exp(\gamma) \sum_{n=2}^{\infty} \exp(-\lambda n^2 - n\gamma)}. \quad (6.29)$$

Spatiotemporal stochastic resonance can be observed for coupling strengths below the critical coupling, which we define as the subthreshold regime. The excitable medium (in the subthreshold regime) is driven by a single wave front (a line in the array) from bottom to top,

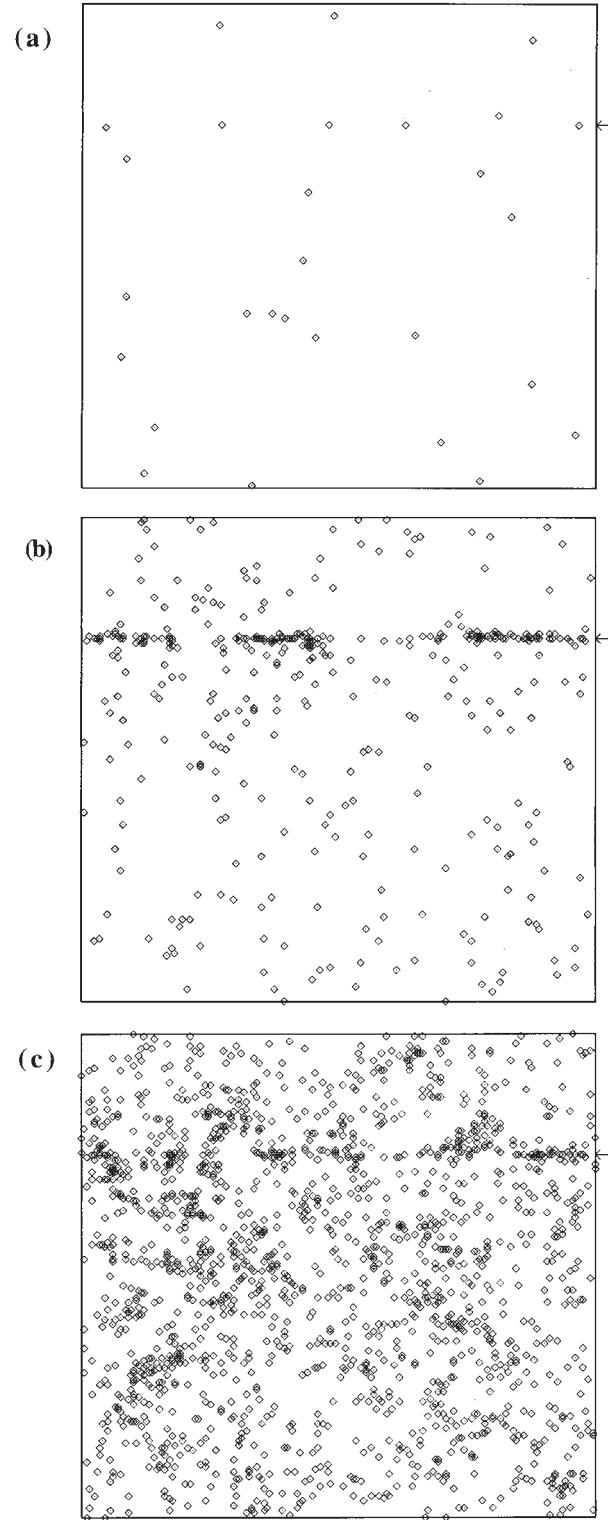


FIG. 44. The excitable medium (200×200 elements) driven by a single wave front of slightly increased excitability, $A_0 = 0.3$. The wave front is a single horizontal line in our array moving from bottom to top. The position of the wave front at the instant of time we took a snapshot is marked by a pointer on the right margin. The diamonds denote firing elements. We show three snapshots at three different noise levels; (a) $\bar{\sigma}^2 = 0.1$, (b) $\bar{\sigma}^2 = 0.16$, and (c) $\bar{\sigma}^2 = 0.2$. The other parameters are $\bar{K} = 0.121$, $\lambda = 0.1$, $\bar{\gamma} = 0.5$.

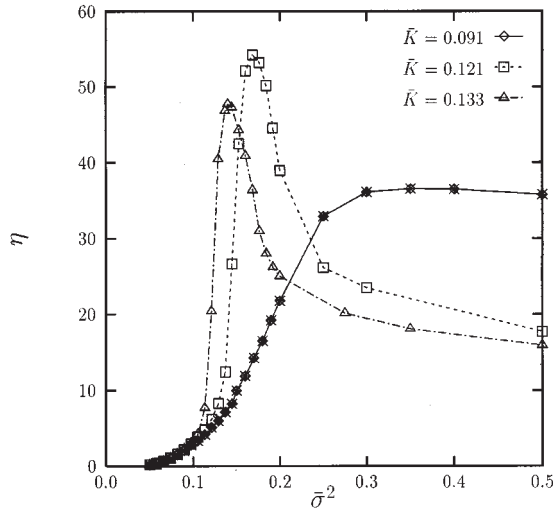


FIG. 45. The mean excess firing rate shown as a function of the variance of the noise σ^2 for $\lambda=0.1$, $\bar{\gamma}=0.5$, and $A_0=0.3$ at three different values of the coupling \bar{K} .

along which the excitability is slightly increased; i.e., the threshold b is reduced: $b \rightarrow b - A_0$. In Fig. 45, the response of the medium, characterized by the correlation between the driving pattern and the firing pattern (see Fig. 44), is shown. The correlation is defined here as the mean excess firing rate along the front of the driving wave, i.e., the number of excess firing events along the front in comparison to the average number of firing events along all other lines (a small layer around the front had been excluded). For vanishing coupling $\bar{K}=0$, one finds the maximum correlation according to the results for single thresholds at $\sigma^2=1/2$. For increasing coupling, the effective threshold that has to be overcome with the help of noise is reduced. Therefore, the maximum correlation is shifted to smaller values of the noise. The peak also becomes more pronounced since the firing activity is synchronized in an area determined by the interaction range $1/\lambda$.

A rough estimate of the optimal value of the variance for the enhancement of spatiotemporal patterns has been given in Jung and Mayer-Kress (1995) in terms of a mean-field type approximation.

The firing elements along the front generate a stochastic field acting on the elements the front is approaching. The main contributions to the field acting on the element e_{ij} stem from firing elements along the front close to e_{ij} . Assuming that all of these elements are actually firing, we approximate the sum of these contributions (for small λ) by a Gaussian integral and obtain for the mean field

$$\bar{f} = \bar{K} \sqrt{\pi/\lambda} \exp(-\lambda). \quad (6.30)$$

The firing threshold is reduced by the mean field, i.e., $\bar{b}_{eff} = 1 - \bar{f}$, leading to a renormalized condition for spatiotemporal stochastic resonance

$$\sigma_{opt}^2 = \frac{1}{2} \bar{b}_{eff}^2 = \frac{1}{2} [1 - \bar{K} \sqrt{\pi/\lambda} \exp(-\lambda)]^2. \quad (6.31)$$

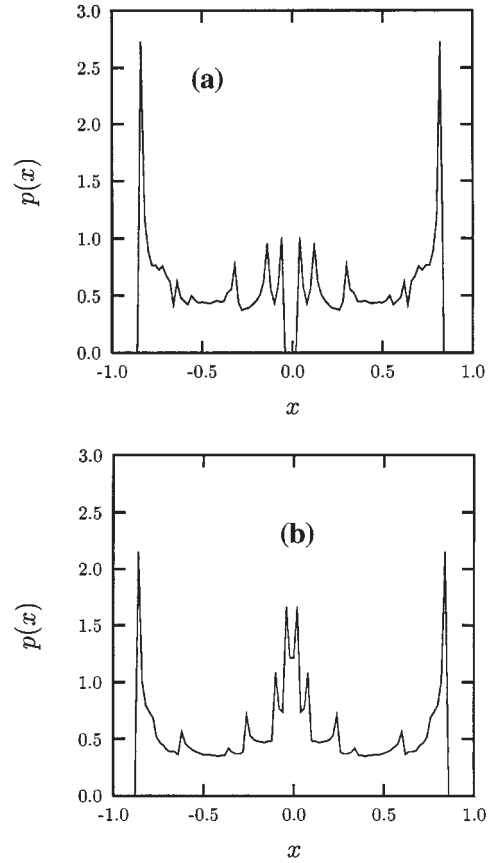


FIG. 46. The probability density $p(x)$ of 10^6 iterates x_n shown in the absence of modulation ($A_0=0$) (a) below the crisis at $a=3.57$, and (b) above the crisis at $a=3.62$.

The effect of improving an image by using stochastic resonance has been demonstrated very nicely in a recent work by Simonotto, Riani, Seife, Roberts, Twitty, and Moss (1997) for an array of uncoupled threshold detectors (in the model above, this corresponds to the case $\bar{K}=0$).

C. Stochastic resonance, chaos, and crisis

It is well known that deterministic chaos resembles the features of noise on a coarse-grained time scale. It is therefore a natural question to ask whether stochastic resonance can be observed in dynamical systems in the absence of noise. Two different approaches to this problem have been put forward in the recent literature. Carroll and Pecora (1993a, 1993b) substitute the stochastic noise by a chaotic source. The chaotic source is applied to a periodically driven Duffing oscillator in a regime where it produces a period-doubled periodic response. The chaotic source yields switching between the attractors corresponding to the two phase-shifted responses of the Duffing oscillator, separated by an unstable period-1 orbit. The switching happens at some preferred locations along the orbits, which are being visited periodically. It is therefore synchronized with the orbit. This situation resembles the conventional setup for stochastic

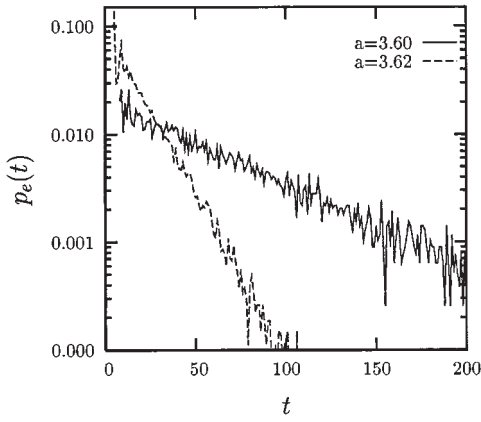


FIG. 47. The distribution of residence times t shown in the absence of forcing ($A_0=0$) at $a=3.6$ and $a=3.62$, both above the crisis $a_0=3.59$.

resonance and yields stochastic resonance as expected. A similar study has been published by Kapitaniak (1994) and by Yang, Ding, and Hu (1995). For the Lorentz system with a time-periodic variation of the control parameter operating near the threshold to chaos, stochastic resonance has been observed by Crisanti, Falcioni, Paladin, and Vulpiani (1994).

A conceptually different approach has been put forward by Anishchenko, Neiman, and Safanova (1993). They use the intrinsic chaotic dynamics of a nonlinear map in the vicinity of a band-merging crisis to generate a sort of activated hopping process which is then synchronized by a small periodic signal. No external source is necessary to provide the randomness. They use the nonlinear periodically driven map

$$x_{n+1} = (a-1)x_n - ax_n^3 + A_0 \sin(2\pi f_0 n). \quad (6.32)$$

The complete description of the period-doubling scenario towards chaos is described in Anishchenko, Neimann, and Safanova (1993). Most important for the following discussion is a crisis due to the merging of two chaotic bands ($x>0$ and $x<0$) at $a \approx 3.598 \equiv a_0$. This is demonstrated in Fig. 46 by the invariant measures of the undriven map $A_0=0$ at $a=3.57$ (a) and $a=3.62$ (b). The fixed point $x_1=0$ is stable for $0 < a < 2$ and unstable for $a > 2$. Two chaotic bands emerge out of two disjoint Feigenbaum-type period-doubling scenarios at $a \geq 3.3$. For $a < a_0$, these bands are separated by the unstable fixed point $x_1=0$. At $a=a_0$ the bands merge. The unstable fixed point $x_1=0$ acts in the vicinity of the band-merging point a_0 as a repeller allowing the trajectory to traverse between the formerly separated chaotic bands only very rarely, yielding activation-type behavior of the trajectory. The statistical distribution $p_e(t)$ of times between two exits, i.e., the residence-time distribution, is shown in Fig. 47. It shows for not too small times the typical exponential decay

$$p_e(t) = \frac{1}{T_e} \exp(-t/T_e), \quad (6.33)$$

with the mean residence time

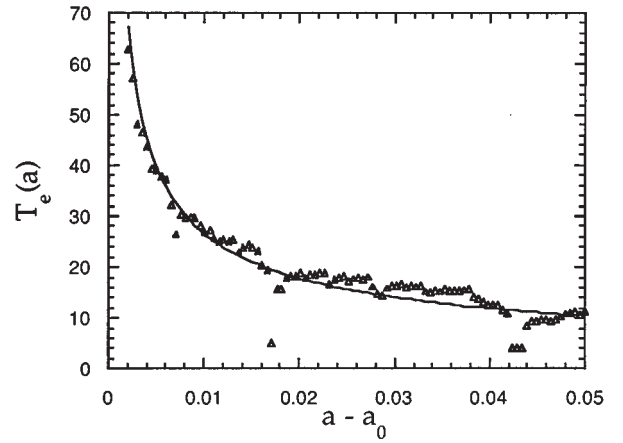


FIG. 48. The mean residence time shown in the absence of the driving as a function of the control parameter $a - a_0$, where a_0 is the value of the control parameter at which the crisis occurs. The triangles represent results from a numerical calculation. Apart from the resonances (the dips) the mean residence time can be fitted very well by a power law $T_e \propto (a - a_0)^\gamma$, with $\gamma=0.576$ (solid line).

$$T_e = \int_0^\infty t p_e(t) dt. \quad (6.34)$$

The power-law scaling of the mean residence time T_e with the distance to the crisis $a - a_0$, i.e.,

$$T_e(a) \propto (a - a_0)^{-\gamma}, \quad (6.35)$$

(see Fig. 48), is—according to Grebogi, Ott, and Yorke (1987)—characteristic of a *crisis*. There are some dips in $T_e(a)$, e.g., at $a=0.36405$, that correspond to periodic windows in the map.

The decrease (at least in the average) of the mean residence time for increasing $a - a_0$ implies an increasing level of stochasticity, i.e., the level of stochasticity can be controlled by varying $a - a_0$.

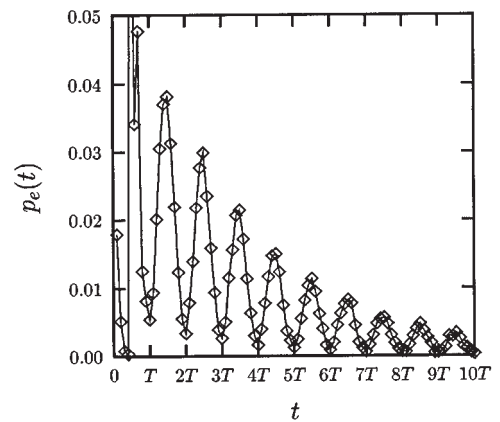


FIG. 49. The distribution of residence times t shown in the presence of the driving at $a=3.60$, $A_0=0.01$, and $f_0=0.1$. The carets show actual data points, while the solid line has been added to guide the eye. The locations of the sequence of exponentially decaying peaks are given by $t_n = (1/2)f_0^{-1}, (3/2)f_0^{-1}, (5/2)f_0^{-1}, \dots$.

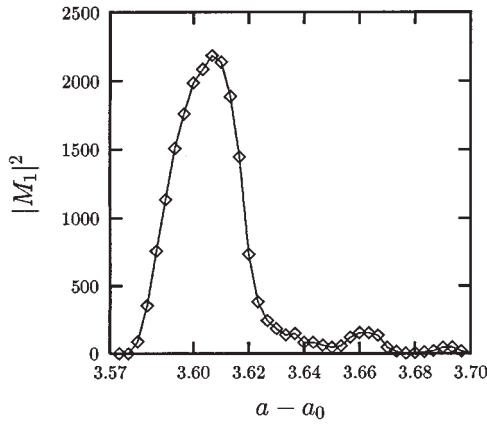


FIG. 50. The response to the periodic forcing (more precisely, the intensity of the line in the power spectrum at the driving frequency) shown as a function of the control parameter $a - a_0$ for $A_0 = 0.01$ and $f_0 = 0.01$.

In the presence of a periodic forcing, i.e., for $A_0 \neq 0$, the residence-time distribution exhibits a series of exponentially decaying peaks, located at odd multiples of the half period of the driving (see Fig. 49). In order to identify stochastic resonance, one has to vary the stochasticity in the presence of the periodic driving and compute the intensity of the spike (the signal strength) in the power spectrum of $x(t)$. To single out contributions to the response from forced periodic motion or resonances within the chaotic bands, we apply a binary filter process to monitor whether the system is in the right band (+1) or left band (-1). The response of the binary variable $y \equiv x/|x|$ measured in terms of the intensity of the line in its power spectrum at the driving frequency is shown in Fig. 50. Starting at the crisis, i.e., at $a = a_0$, one observes that the signal strength increases until it reaches a maximum, and then decreases again.

The differences to stochastic resonance in a noisy bistable system follow:

(i) There are several peaks plus additional resonances where the driving frequency f_0 and the switching frequency $f_s = 1/T_e$ are commensurable, i.e., $f_0/f_s = m/n$, with $m, n = 1, 2, 3, \dots$.

(ii) Changing $a - a_0$ does not change the stochasticity (noise strength) in a systematic way. To systematically compare with stochastic resonance in a noisy bistable system, one should first find a mapping between the noise strength and $a - a_0$.

(iii) The periodic windows of period M of the unperturbed map yield resonances with the external driving whenever their periods agree, i.e., $f_0 = 1/M$.

Similar results have been obtained by Nicolis *et al.* (1993) by studying a one-dimensional intermittent map.

D. Effects of noise color

In many practical situations the finite time scale τ_c characterizing the relaxation of the autocorrelation of the noise (i.e., colored noise) is much shorter than the characteristic time scale of the system. Hence as we did

in Sec. IV for the driven bistable dynamics, it often is appropriate to model the noise source by a (white) random force $\xi(t)$. In the physical world, however, such an idealization is never exactly realized. In order to investigate the importance of corrections to white noise, approximate techniques were introduced to compute the effects of small to moderate to arbitrarily large noise correlation times τ_c (Hänggi *et al.*, 1984, 1989; Hänggi and Jung, 1995). Strong color (i.e., a large τ_c value) is not unrealistic for many physical applications. Usually, a strongly correlated noise emerges as the result of coarse graining over a hidden set of slowly varying variables (Kubo *et al.*, 1985), or colored noise is simply applied and monitored externally by the experimenter.

The effect of color on stochastic resonance may be nontrivial, as suggested by the very characterization of stochastic resonance as a synchronization mechanism. The noise correlation time τ_c may compete with T_Ω and T_K to determine the realization and the magnitude of the resonance phenomenon. We anticipate that stochastic resonance in overdamped systems driven by an additive exponentially correlated Gaussian noise $\xi(t)$ is generally reduced compared to the case of white noise $\tau_c = 0$ of equal strength D . The stochastic resonance peak is shifted to larger noise intensities due to the fact that colored noise exponentially suppresses the switching rate with increasing τ_c (Gammaitoni, Menichella-Saetta, Santucci, Marchesoni, and Presilla, 1989; Hänggi *et al.*, 1993).

Following the approach developed by Hänggi *et al.* (1993), we treat here the archetypal case of a periodically perturbed double well in the presence of exponentially colored Gaussian noise (Ornstein-Uhlenbeck noise). In scaled, dimensionless variables, the dynamics reads explicitly

$$\dot{x} = -V'(x) + A_0 \cos(\Omega t) + \xi(t), \quad (6.36a)$$

$$\dot{\xi} = -\frac{1}{\tau_c} \xi + \frac{1}{\tau_c} \epsilon(t), \quad (6.36b)$$

where $V(x)$ is the standard quartic double-well potential of Sec. IV.A, i.e., $V(x) = -x^2/2 + x^4/4$, and $\xi(t)$ is an Ornstein-Uhlenbeck stochastic process driven by the Gaussian white noise $\epsilon(t)$ with $\langle \epsilon(t) \rangle = 0$ and $\langle \epsilon(t) \epsilon(0) \rangle = 2D \delta(t)$. The stationary autocorrelation function of $\epsilon(t)$ is an exponential function with time constant τ_c ,

$$\langle \xi(t) \xi(0) \rangle = (D/\tau_c) \exp(-|t|/\tau_c). \quad (6.37)$$

In the limit of zero correlation time $\tau_c \rightarrow 0$, Eq. (6.37) reproduces the white-noise source of Secs. II and IV. Within the framework of the linear-response theory of Sec. IV.B for small forcing amplitudes, the relevant response function [Eq. (4.30)] assumes the form of a fluctuation theorem; i.e., it is given in terms of a stationary correlation of two fluctuations of the unperturbed process

$$\chi(t) = -H(t) \frac{d}{dt} \langle x(t) \xi(x(0)) \rangle_0, \quad (6.38)$$

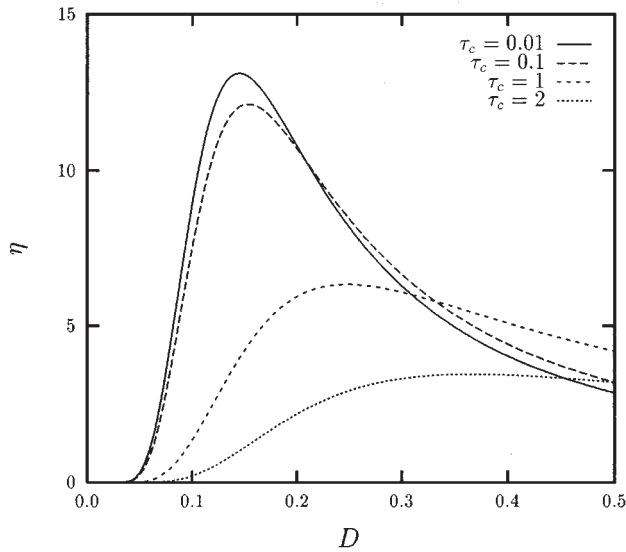


FIG. 51. The spectral amplification η (in linear response), as predicted by Eqs. (6.43) and (6.44) for the quartic double-well potential $V(x) = -x^2/2 + x^4/4$, shown for increasing values of the dimensionless noise correlation time τ_c at $\Omega = 0.1$.

where $H(t)$ denotes a Heaviside step function. In the opposite limits $\tau_c \ll 1$ (i.e., weak color) and $\tau_c \gg 1$ (i.e., strong color), $\zeta(x(t))$ may be approximated to read

$$\zeta(x(t)) = \frac{1}{D} [x + \tau_c V'(x)], \quad (6.39)$$

whereas the unperturbed averages $\langle \cdots \rangle_0$ in Eq. (6.38) must be taken over the relevant (approximate) probability density

$$p_{st}(x, \tau_c) = \frac{\mathcal{N}_1}{|1 - \tau_c V''(x)|} \exp \left[-\frac{V(x)}{D} - \frac{\tau_c}{2D} V'^2(x) \right] \quad (6.40a)$$

for $\tau_c \ll 1$, and

$$p_{st}(x, \tau_c) = \mathcal{N}_2 [1 + \tau_c V''(x)] \times \exp \left[-\frac{V(x)}{D} - \frac{\tau_c}{2D} V'^2(x) \right] \quad (6.40b)$$

for $\tau_c \gg 1$. Here, \mathcal{N}_1 and \mathcal{N}_2 denote the appropriate normalization constants.

Within the long-time approximation, the correlation function $\langle x(t)\zeta(x(0)) \rangle$ (see Sec. IV.B) can in leading order be estimated as

$$\langle x(t)\zeta(x(0)) \rangle_0 \sim \langle x\zeta \rangle_0 \exp[-2r_K(\tau_c)t] \quad (6.41)$$

with the colored noise-driven escape rate given as

$$r_K(\tau_c) = r_K(1 - 3\tau_c/2) \quad (6.42a)$$

for weakly colored noise $\tau_c \ll 1$, and

$$r_K(\tau_c) = r_K \exp[-(8/27)\tau_c(\Delta V/D)] \quad (6.42b)$$

for strongly colored noise, i.e., $\tau_c \gg 1$. Upon inserting Eqs. (6.41) and (6.39) into the expression for the susceptibility in Eq. (6.38) we finally obtain for the spectral amplification η

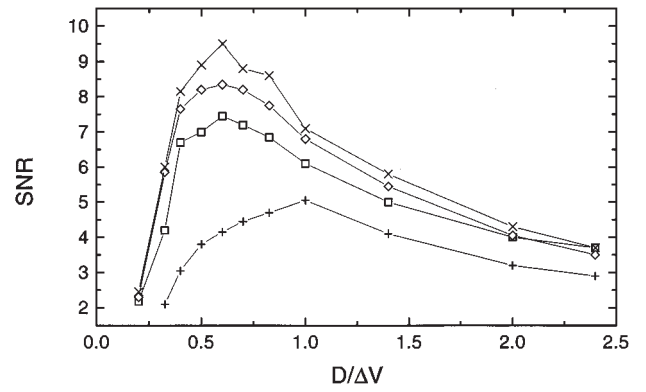


FIG. 52. Signal-to-noise ratio (SNR) vs D for different values of τ_c : $\tau_c = 30 \mu\text{s}$ (crosses); $\tau_c = 50 \mu\text{s}$ (diamonds); $\tau_c = 100 \mu\text{s}$ (squares); $\tau_c = 200 \mu\text{s}$ (pluses). Other simulation parameters are $\nu_\Omega = 30 \text{ Hz}$, $Ax_m = 0.5\Delta V$, $x_m = 7.3 \text{ V}$, and $a = 6850 \text{ Hz}$.

$$\eta = \frac{\langle x^2 \rangle_0 + \tau_c \langle x V'(x) \rangle_0}{D^2} \frac{4r_K^2(\tau_c)}{4r_K^2(\tau_c) + \Omega^2}, \quad (6.43)$$

with $r_K(\tau_c)$ and $\langle \cdots \rangle_0$ computed in the appropriate limits of Eqs. (6.42) and (6.40), respectively.

Prediction (6.43) for both color regimes suggests that noise color degrades the observability of stochastic resonance. Indeed, upon increasing τ_c , the relaxation rate $r_K(\tau_c)$ gets exponentially depressed with respect to $r_K(\tau_c = 0)$. For Ω fixed, we therefore must increase D to match the stochastic resonance condition, which consists of maximizing η (see Sec. IV.B). This results in a shift of the stochastic resonance peak towards higher D values and a corresponding reduction of the peak height.

We note that the limiting expressions (6.42) for $r_K(\tau_c)$ stem from one unified approximation scheme (Hänggi and Jung, 1995), that is

$$r_K(\tau_c) = \frac{1}{\sqrt{2}\pi} (1 + 3\tau_c)^{-1/2} \times \exp \left[-\frac{\Delta V}{D} \left(\frac{1 + \frac{27}{16}\tau_c + \frac{1}{2}(\tau_c)^2}{1 + \frac{27}{16}\tau_c} \right) \right]. \quad (6.44)$$

Correspondingly, the analytical expression obtained by replacing $r_K(\tau_c)$ of Eq. (6.43) with Eq. (6.44) bridges the two limiting expressions of η for $\tau_c \ll 1$ and $\tau_c \gg 1$. In Fig. 51 we display four such curves for η versus D , for increasing values of noise color τ . As expected, noise color suppresses stochastic resonance monotonically with τ_c . This feature is in accordance with the early analog simulations by Gammaitoni, Menichella-Saetta, Santucci, Marchesoni, and Presilla (1989), as depicted in Fig. 52. The suppression of stochastic resonance with increasing noise color has recently been demonstrated experimentally in a tunnel diode (Mantegna and Spagnolo, 1995), and by use of Monte Carlo simulations by Berghaus *et al.* (1996).

Inertial effects, which result in (non-Markovian) memory effects for the spatial coordinate $x(t)$ have been addressed theoretically for the spectral amplifica-

tion η by Hänggi *et al.* (1993), and experimentally for the signal-to-noise ratio by Gammaitoni, Menichella-Saetta, Santucci, Marchesoni, and Presilla (1989). Likewise, the role of memory friction, which relates—via the fluctuation-dissipation theorem—to internal colored noise, has been studied for overdamped dynamics by Neiman and Sung (1996). The result that inertial effects, or equivalently, decreasing finite friction strengths tend to enhance stochastic resonance can similarly be induced with strong color for the memory friction (Neiman and Sung, 1996).

VII. SUNDRY TOPICS

In this section we present several topics relating to the physics of stochastic resonance that have not been featured in detail in the previous sections. In doing so, we have confined ourselves to particular examples determined only by our knowledge and personal taste; thus this selection is necessarily incomplete.

A. Devices

1. Stochastic resonance and the dithering effect

A *Schmitt trigger*, see Sec. V.B.1, operating in the limiting case when its two threshold voltages coincide, provides an example of a two-state system, namely a *threshold device*. There are many examples of this kind of electronic device, the more common class being represented by the analog-to-digital converters (ADC). The basic (1-bit) ADC device consists of a signal comparator (an operational amplifier followed by a resistor and two back-to-back *Zener diodes*), the output voltage of which switches between V and $-V$ when the input v_i crosses a reference voltage. Multibit ADCs, realized by a proper combination of comparators (see, e.g., Millman, 1983), are of common use in digital signal processing (Oppenheim and Schaffer, 1975), where analog signals are sampled at discrete times and converted into a sequence of numbers. Since the register length is finite, the conversion procedure, termed signal *quantization*, results in distortion and loss of signal detail. In order to avoid distortion and recover the signal detail, it has become a common practice, since the 1960s, to add a small amount of noise to the analog signal prior to quantization—a technique termed *dithering* (Bennet, 1948). To understand how the addition of a proper quantity of noise can improve the performances of an ADC, we note that the conversion from a continuous (analog) to a digital signal consists of two different operations: time discretization and amplitude quantization. Time discretization, if properly applied, can be shown to be error free. The effects of amplitude quantization (finite word length) are instead always present and manifest themselves in a number of different ways. First, due to the presence of a nonlinear-response characteristic, signal quantization leads to an unavoidable distortion, i.e., the presence of spurious signals in a frequency band other than the original one. There is also a loss of signal detail that is small compared to the quantization step. The effects of

the amplitude quantization can be quantified by introducing a proper *quantization error*, $\zeta = y - x$, where x is the analog signal before quantization and y is the quantized signal. It is clear from this definition that if we had a linear-response characteristic (apart from amplification factors) ζ would be zero and there would be no distortion at all. A number of studies were performed over the last thirty years in order to find a way of reducing ζ . The main conclusions follow:

(1) The addition of a proper external signal (called *dither*) to the input x can statistically reduce ζ .

(2) The best choice for the dither signal is a random dither uniformly distributed.

(3) There exists an optimal value of the random dither amplitude, which coincides with the amplitude of the quantization step.

Hence the quantization error ζ is minimized and, correspondingly, the ADC performances maximized, when a noise of a proper intensity is added to the input signal. The similarity with stochastic resonance, where an optimal strength of the added noise maximizes the output signal-to-noise ratio, is apparent. As a matter of fact, stochastic resonance in this class of threshold systems is equivalent to the dithering effect, as demonstrated by Gammaitoni (1995a, 1995b).

B. Stochastic resonance in coupled systems

In this section we discuss the impact of noise and periodic forcing on an ensemble of coupled bistable systems. In view of a possible collective response of the system (especially close to a phase transition), one can expect that the stochastic resonance effect will be even more pronounced than in a single system (Jung *et al.*, 1992).

1. Two coupled bistable systems

The simplest way to study stochastic resonance in coupled systems is to consider two coupled overdamped bistable elements in the presence of noise and periodic forcing (Neiman and Schimansky-Geier, 1995):

$$\begin{aligned}\dot{x} &= \alpha x - x^3 + \gamma(y - x) + \xi_x(t) + A_0 \cos(\Omega t), \\ \dot{y} &= \beta y - y^3 + \gamma(x - y) + \xi_y(t) + A_0 \cos(\Omega t),\end{aligned}\quad (7.1)$$

with independent Gaussian white noise terms, but identical periodic forcing, i.e.,

$$\langle \xi_x(t) \xi_y(t') \rangle = 2D \delta_{xy} \delta(t - t'). \quad (7.2)$$

As in the bistable string (see Sec. VI.B.1), stochastic resonance in the coupled system has been quantified by the linear response for the sum $s(t)$ of the two degrees of freedom $s(t) = x(t) + y(t)$ due to small periodic modulations. With the help of digital simulations and approximation theory, the following results have been obtained:

(1) At a given coupling constant, the signal-to-noise ratio goes through a maximum as a function of the noise strength.

(2) Starting from zero coupling (which corresponds to two independent systems), the signal-to-noise ratio vs coupling first increases (i.e., the collective response is indeed higher than that of two uncoupled systems), runs through a maximum, and decreases again for large coupling towards an asymptotic (finite) value.

2. Collective response in globally coupled bistable systems

An early study dealing with stochastic resonance for systems with many degrees of freedom (Jung *et al.*, 1992) focused on a large number N of identical, linearly and homogeneously coupled bistable systems in the presence of periodic forcing. The coupled equations of motion are given by

$$\dot{x}_n(t) = x_n - x_n^3 + \frac{1}{N} \sum_{m=1}^N g(x_m - x_n) + \xi_n(t) + A_0 \cos(\Omega t), \quad (7.3)$$

with Gaussian, mutually independent and uncorrelated fluctuations

$$\begin{aligned} \langle \xi_n(t) \xi_m(t') \rangle &= 2D \delta_{nm} \delta(t - t') \\ \langle \xi_n(t) \rangle &= 0. \end{aligned} \quad (7.4)$$

The coupling constant is denoted by g . Systems such as this exhibit spontaneous-ordering transitions (Bruce, 1980; Amit, 1984; Dewel *et al.*, 1985; Valls and Mazenko, 1986). Analytical studies of these phase transitions are possible within a mean-field approximation (Mansour and Nicolis, 1975; Desai and Zwanzig, 1978; Bruce, 1980; Shiino, 1987; Van den Broeck *et al.*, 1994; Drozdov and Morillo, 1996; Hu, Haken, and Xie, 1996). The stationary solution of the Fokker-Planck equation in mean-field approximation is not unique below a critical noise strength. There are three solutions: two stable solutions with spontaneous symmetry breaking, which represent ferromagnetic ordered states, and an unstable one with zero magnetization m ; here the order parameter m is given by the averaged population difference in the potential wells. At the critical point $D = D_c$, the system undergoes a phase transition of second order.

Within the mean-field approximation and a two-state description, the response of the order parameter $\langle x \rangle \equiv m$ to the periodic forcing and thus the spectral amplification η of the order parameter has been obtained as

$$\eta = \left(\frac{2r_K}{D} \right)^2 \frac{1 - m^2}{\Omega^2 + \Lambda^2}, \quad (7.5)$$

with the collective relaxation rate given by

$$\Lambda = 2r_K \sqrt{1 - m^2} \left(\frac{1}{1 - m^2} - \frac{g}{D} \right). \quad (7.6)$$

The mean value m determined by the transcendental equation $m = \tanh[(g/D)m]$.

The spectral amplification strongly increases with the coupling to exhibit a peak at the critical point $D = D_c = g$. The maximum spectral amplification attains a maximum at $g = \Delta V$, a phenomenon that has been ob-

served for two coupled bistable systems (cf. Sec. VII.B.1) and for coupled-neuron models in Sec. VII.B.3. These results have also been confirmed in later studies by Morillo *et al.* (1995), and Hu, Haken, and Xie (1996).

3. Globally coupled neuron models

Another approach to describe the response of globally coupled bistable systems to periodic forcing is the application of adiabatic elimination of all but one degree of freedom (Bulsara and Schmeira, 1993; Inchiosa and Bulsara, 1995a, 1995b, 1995c; Inchiosa and Bulsara, 1996). The model used is motivated by the dynamics of artificial neural networks (Amit, 1989; Krogh and Palmer, 1991), namely

$$C_i \dot{u}_i = -\frac{u_i}{R_i} + \sum_{j=1}^N J_{ij} \tanh(u_j) + \xi(t) + A_0 \cos(\Omega t), \quad (7.7)$$

with C_i and R_i denoting capacitances and resistances of the membranes. The zero-mean Gaussian noise and the periodic signal are assumed to be identical for all elements. The coupling constants J_{ij} can be chosen arbitrarily. The correlation function of the noise is given by

$$\langle \xi(t) \xi(t') \rangle = 2D \delta(t - t'). \quad (7.8)$$

The globally coupled system of Eq. (7.7) has been solved numerically and analytically by assuming a separation of time scales of one neuron vs the rest of the neurons acting as a linearized bath, thus allowing for adiabatic elimination of the bath neurons. The most important result of these studies is that, as above, the maximal signal-to-noise ratio goes through a maximum as a function of the coupling strength (the J 's); moreover, the signal-to-noise ratio between the incoming periodic signal $A_0 \cos(\Omega t)$ and the noise strength D has been shown to provide an upper bound to the signal-to-noise ratio of the output $u_i(t)$.

C. Miscellaneous topics on stochastic resonance

1. Multiplicative stochastic resonance

There exist many cases of physical interest where the role of fluctuating control parameters is mimicked by *multiplicative* noise (Fox, 1978; Schenzle and Brand, 1979; Faetti *et al.*, 1982; Graham and Schenzle, 1982). Gammaitoni, Marchesoni, Menichella-Saetta, and Santucci (1994) have analogously simulated the phenomenon of stochastic resonance in the overdamped bistable system described by the stochastic differential equation

$$\dot{x} = -V'(x) + x \xi_M(t) + \xi_A(t) + A(t), \quad (7.9)$$

where $V(x)$ is the standard quartic double-well potential and $A(t) = A_0 \cos(\Omega t)$, with $A_0 x_m \ll \Delta V$. The fluctuating parameters $\xi_i(t)$, with $i = A, M$, are stationary zero-mean valued, Gaussian random processes with autocorrelation functions

$$\langle \xi_i(t) \xi_j(0) \rangle = 2Q_i \delta_{ij} \delta(t). \quad (7.10)$$

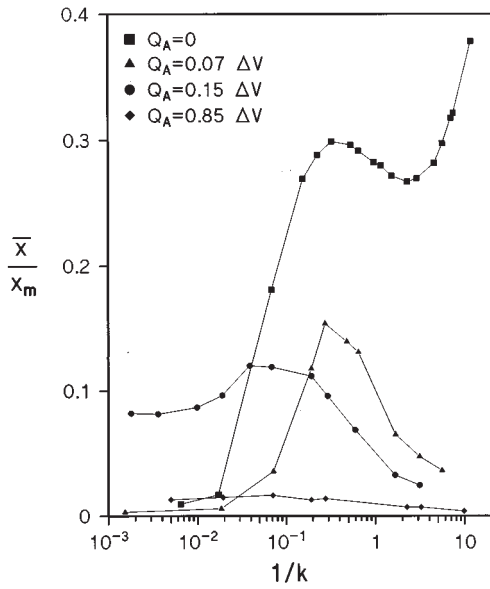


FIG. 53. The normalized response amplitude \bar{x}/x_m depicted vs the dimensionless multiplicative noise strength $k^{-1} \equiv 2Q_M/a$ for $A_0 = 0.1ax_m$ and different values of Q_A . The potential parameters are $x_m = 2.2$ V and $a = 10^4$ s $^{-1}$. After Gammaitoni, Marchesoni, Menichella-Saetta, and Santucci (1994).

The main conclusion of their investigation is that the process $x(t)$ develops a periodic component $\langle x(t) \rangle_{as}$ according to the approximate law (2.6); the amplitude \bar{x} and the phase $\bar{\phi}$ of $\langle x(t) \rangle_{as}$ depend on A_0 , Q_A , and Q_M . Most notably, \bar{x} shows typical stochastic resonance behavior with increasing Q_M while keeping Q_A fixed (multiplicative stochastic resonance). In Fig. 53, the dependence of \bar{x}_1 on Q_M is plotted for the most remarkable case of a purely multiplicative bistable process $Q_A = 0$.

To interpret the outcome of the analog simulation of Eq. (7.9), one should solve the relevant time-dependent Fokker-Planck equation

$$\frac{\partial}{\partial t} p = \frac{\partial}{\partial x} \left[V'(x) - Q_M x - A_0 \cos \Omega t \right] p + \frac{\partial}{\partial x} (Q_A + Q_M x^2) p \quad (7.11)$$

for the probability density $p = p(x, t; A_0)$. In the presence of a static tilting, i.e., for $A_0 \neq 0$ and $\Omega = 0$, the stationary solution of Eq. (7.11) reads

$$p_0(x; A_0) = \mathcal{N}_0(A_0) \left(x^2 + \frac{Q_A}{Q_M} \right)^{-1/2 + k[1 + (k/2)(Q_A/\Delta V)]} \times \exp \left(-k \frac{x^2}{x_m^2} - \frac{A_0}{Q_M |x|} \right), \quad (7.12)$$

where $k \equiv a/2Q_M$ and $\mathcal{N}_0(A_0)$ is a suitable normalization constant. In the presence of a periodic tilting; i.e., for $\Omega > 0$, the process $x(t)$ is no longer stationary and a time-dependent probability density $p_{as}(x, t; A_0)$ is required to describe its asymptotic state. However, in the

limit of low forcing frequency Ω , the adiabatic approximation $p_{as}(x, t; A_0) \approx p_0(x; A(t))$ suffices to shed light on the nonstationary dynamics underlying the phenomenon of multiplicative stochastic resonance.

In the purely multiplicative case $Q_A = 0$ (Fig. 53), the forcing term alone is responsible for $x(t)$ switching back and forth between the positive and the negative half axis. Should the adiabatic approximation hold true for any value of Q_M , the process $x(t)$ would approach instantaneously its most probable value in the vicinity of the peak of $p_0(x; A(t))$. Therefore, the amplitude \bar{x} of $\langle x(t) \rangle_{as}$ would be of the order of \bar{x}_m , which is a monotonic decreasing function of Q_M . However, Fig. 53 shows a dramatic drop of \bar{x} as Q_M tends to zero. Such a deviation from the prediction of the adiabatic approximation is due to the fact that, with decreasing Q_M , the switch time of $x(t)$ between positive and negative values, controlled by $A(t)$ with periodically reversing sign, grows much larger than the forcing period $T_\Omega = 2\pi/\Omega$. For instance, assuming that at $t = 0$ the variable x is confined to the unstable axis $x < x_A \approx A_0/a$ with $A_0 > 0$, the mean-first-passage time T_A required by x to escape through x_A onto the stable half axis $x > x_A$ diverges strongly for $x_A/x_m \rightarrow 0$. On increasing Q_M close to a , such a divergence is substantially weakened, so that the adiabatic approximation $T_A \ll T_\Omega$ applies. In this regime the relaxation process is controlled mainly by the modulated interwell dynamics of $x(t)$ described by $p_0(x; A(t))$ and, as stated above, \bar{x} approaches \bar{x}_m . In the opposite limit, $T_A \gg T_\Omega$ (i.e., $Q_M \ll a$), the steady-state distribution of $x(t)$ spreads over the entire x axis with oscillating local maxima at $\pm x_m + A(t)/2(a - Q_M)$ (modulated intrawell dynamics). It follows immediately that for $Q_M = 0+$ the amplitude \bar{x} is of the order $A_0/2a$, which is much smaller than the value of \bar{x}_m at $k = 1$, whence the appearance of the stochastic resonance peaks of Fig. 53 for $\Omega T_A \sim 1$. Accordingly, the stochastic resonance peaks shift to the left with increasing A_0 . In conclusion, the crossover from intrawell to interwell modulated dynamics is the basic mechanism responsible for multiplicative stochastic resonance.

2. Resonant crossing

In this section, we report on an intricate colored-noise effect for the residence-time distributions, which takes place when the correlation time of the noise is large (Gammaitoni, Marchesoni, *et al.*, 1993). In such a situation, the system dynamics are characterized by four time scales: the local relaxation rate a within a potential well, the correlation time of the noise τ_c , the forcing frequency Ω , and the transition rate r_K . The residence-time distribution (at small amplitudes of the driving A_0) consists—as shown repeatedly in figures throughout this review—of a series of peaks located at odd multiples of the half period of the driving $T_\Omega = 2\pi/\Omega$, superimposed on an exponential backbone. The periodic part can be extracted from the exponential backbone by a fitting procedure.

It has been demonstrated that the amplitude of the periodic part evolves as a function of the driving frequency Ω through a maximum if the correlation time τ_c of the noise is large, i.e., $a\tau_c \gg 1$. The location of the maximum has been estimated by $\Omega_{\max}^2 \sim a/\tau_c$. In contrast to the time-scale matching condition for stochastic resonance, i.e., $\Omega \sim \pi r_K$, this new condition describes a matching between the intrawell time scale and the driving frequency Ω .

3. Aperiodic stochastic resonance

Since most of the studies on stochastic resonance assume periodic external forcing, it is interesting to ask whether noise can also amplify small aperiodic signals. To this end several different studies (in scope and technique) have been put forward.

Jung and Hänggi (1991a) considered noise due to the phase diffusion of the external force. This is a realistic assumption, for instance, if the external field is provided by a laser where spontaneous emission generates phase diffusion (Haken, 1970). Instead of a deterministic phase Ωt , they proposed the use of a stochastic phase θ , i.e., $\dot{\theta} = \Omega + (1/\tau_d) \xi_\theta(t)$, with $\xi_\theta(t)$ the derivative of a Wiener process. The finite coherence time of the phase dynamics leads to a broadening of the peaks in the power spectrum and a suppression of the stochastic resonance effect.

Neiman and Schimansky-Geier (1994) considered the overdamped motion of a particle in a bistable potential $V(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2$, driven by white Gaussian noise and harmonic noise (Schimansky-Geier and Zülicke, 1990; Dykman, Mannella, McClintock, Stein, and Stocks, 1993b). Harmonic noise $y(t)$ is generated by applying white Gaussian noise on a second-order linear filter. The spectral density of the harmonic noise has a peak at a nonzero frequency ω_p and thus mimics a certain degree of periodicity. The power spectrum of $x(t)$ exhibits a maximum at ω_{\max} , which is located close to ω_p , but with a small variation as a function of the noise. As in the case of phase diffusion, the peaks have a finite width. The signal-to-noise ratio shows a relative maximum at a finite noise strength typical of stochastic resonance.

In recent years, we witness a prosperous period for aperiodic stochastic resonance, which was ushered in by addressing the problem of optimizing information transfer in excitable systems (Collins *et al.*, 1995a, 1995b; De Weese and Bialek, 1995; Collins, Chow, *et al.*, 1996; Collins, Imhoff, and Grigg, 1996; Heneghan *et al.*, 1996; Levin and Miller, 1996; Neiman *et al.*, 1997). The above-named authors considered the Fitzhugh-Nagumo equations driven by white noise and an arbitrary aperiodic signal. This system was operated below threshold and the aperiodic signal was not large enough to induce excitation. Together with the noise, however, excitations were possible. Stochastic resonance has been demonstrated for the correlation of the aperiodic signal with the excitation rate (the number of excitation events per unit time).

This area has stimulated an interesting ongoing discussion: Do there exist suitable measures quantifying stochastic resonance—and what are they—that can be based solely on information theory considerations? A promising approach has been put forward by Heneghan *et al.* (1996) who consider the so-termed *transinformation* that quantifies the rate of information transfer from stimulus to response. They demonstrated that the presence of noise optimizes, via aperiodic stochastic resonance, the information-transfer rate. An attempt to characterize conventional stochastic resonance by means of information theory tools has been put forward by Schimansky-Geier and co-workers (Neiman *et al.*, 1996; Schimansky-Geier *et al.*, 1998), by Bulsara and Zador (1996), and by Chapeau-Blondeau (1997). Considering conditional entropies and Kullback measures, Schimansky-Geier *et al.* (1998) demonstrated with a Schmitt trigger system, driven periodically at strong, but still subthreshold amplitude strengths, that information measures do exhibit characteristic extrema. These extrema, however, do *not* describe the conventional regime of stochastic resonance for the signal-to-noise ratio, but they rather seem to mimic the stochastic resonance behavior in a regime that is in accordance with stochastic resonance for the spectral amplification η .

D. Stochastic resonance—related topics

1. Noise-induced resonances

In studying stochastic resonance, one looks at the periodic contribution of the output at the same frequency as the input. More recently, the general question of the generation of higher harmonics in the presence of noise has been addressed in a number of studies (Bartussek *et al.*, 1993; Dykman, Mannella, McClintock, Stein, and Stocks, 1993a; Bartussek, Hänggi, and Jung, 1994; Dykman *et al.*, 1994; Jung and Talkner, 1995; Bulsara, Inchiosa, and Gammaitoni, 1996; Jung and Bartussek, 1996). In this section we focus on a novel effect (Bartussek, Hänggi, and Jung, 1994), namely the noise-selective resonance-like suppression of higher harmonics. These “noise-induced resonances” have been observed using numerical solutions of the Fokker-Planck equation in bistable systems (Bartussek, Hänggi, and Jung, 1994) as well as in monostable systems (Jung and Bartussek, 1996). Noise-induced resonances have already been observed in experiments with a periodically driven SQUID by Rouse, Han, and Lukens (1995). Recently, similar resonances have been predicted for quantum stochastic resonance by Grifoni and Hänggi (1996a, 1996b).

Apart from numerical, adiabatic studies (Bartussek, Hänggi, and Jung, 1994; Rouse *et al.*, 1995), an analytical theory allowing one to predict whether or not a particular system would exhibit noise-induced resonance has been put forward by Jung and Talkner (1995). Their approach is sketched as follows: we consider here a general overdamped system subject to additive white noise and periodic forcing, i.e.,

$$\dot{x} = f(x) + A_0 \cos(\Omega t) + \xi(t), \quad (7.13)$$

where $\xi(t)$ is as usual white Gaussian noise with zero mean and strength D , and $f(x)$ is the forcing function.

As shown in Sec. IV.A, the spectral density consists of a broad background and δ spikes at multiples of the driving frequency. The weights of the δ spikes are given by $g_n = 2\pi |M_n|^2$, where M_n are the complex Fourier coefficients of the asymptotic (large times) mean value $\langle x(t) \rangle_{as}$. Since the n th harmonic is in leading order proportional to A_0^n (see Jung and Bartussek, 1996), the intensities g_n are proportional to A_0^{2n} . We therefore define the characteristic coefficients

$$\gamma_n \equiv \frac{4\pi |M_n|^2}{n!^2 A_0^{2n}}, \quad (7.14)$$

to describe the intensities of the harmonics. In the adiabatic approximation, the characteristic coefficients γ_n have been obtained by Jung and Talkner (1995) in their respective leading order A_0^{2n} as

$$\gamma_n = \frac{4\pi |M_n|^2}{n!^2 A_0^{2n}} \approx \frac{4\pi}{(2D)^{2n}} \left(\frac{K_{n+1}}{n!} \right)^2, \quad (7.15)$$

where K_n are the *cumulants* of the stationary probability density of the unperturbed process ($A_0=0$). The intensity of the basic harmonic γ_1 cannot exhibit noise-induced resonance for any system, because the second order cumulant is strictly positive. The sign of the cumulants $K_{n>2}$ can change, for example, as a function of the noise strength D , giving rise to zeros of the intensities γ_n of the higher harmonics, i.e., to noise-induced resonances.

Decomposing the complex amplitude M_n into the product $M_n = |M_n| \sin \phi_n$, it can be seen that whenever a noise-induced resonance occurs, the phase ϕ_n exhibits a jump of magnitude π . Several concrete systems (single well, double well, two-state system, etc.) have been discussed by Jung and Talkner (1995).

2. Periodically rocked molecular motors

It is generally appreciated that useful work cannot be extracted from thermal equilibrium fluctuations. Such a device would violate the second law of thermodynamics. Feynman *et al.* (1966) discussed this issue by means of a model of a mechanical ratchet—a scheme that was originally devised and elucidated during the heyday of early Brownian motion by M. V. Smoluchowski (1912, 1914). In his articles, which these days still provide delightful reading, Smoluchowski (1912, 1914) shows that in the absence of an intelligent creature, such as a Maxwell demon, no net currents will occur. In the presence of nonequilibrium forces the situation changes drastically: now a thermal ratchet system, that is, a periodic structure with spatial asymmetry subjected to noise, can rectify symmetric, unbiased nonequilibrium fluctuations into a fluctuation-induced directed current (Ajdari and Prost, 1992; Magnasco, 1993; Astumian and Bier, 1994; Bartussek, Hänggi, and Kissner, 1994; Doering *et al.*, 1994; Leibler, 1994; for a comprehensive reviews see Hänggi and Bartussek, 1996; Astumian, 1997; Jülicher *et al.*, 1997). Put differently, by a ratchet we mean a system that is able to move particles with finite macroscopic velocity in the absence of any macroscopic bias forces

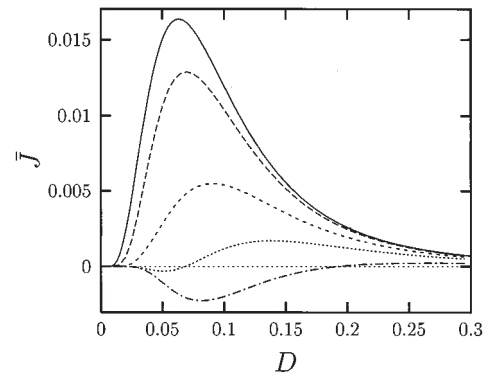


FIG. 54. Unidirectional probability current \bar{J} vs noise strength D at fixed driving amplitude $A_0=0.5$ in a rocking ratchet with asymmetric periodic potential $V_R(x) = -[\sin(2\pi x) + \frac{1}{4}\sin(4\pi x)]/2\pi$. The various lines correspond to different driving angular frequencies Ω : adiabatic driving: $\Omega=0.01$ (solid line); and nonadiabatic driving: $\Omega=1$ (dashed), $\Omega=2.5$ (short-dashed), $\Omega=4$ (dotted), and $\Omega=7$ (dashed-dotted). With the period L of the ratchet potential equal to unity, the average particle drift $\langle \dot{x} \rangle = L\bar{J}$ equals in this case the probability current \bar{J} . A characteristic current reversal (with \bar{J} passing through zero) occurs in the regime of nonadiabatic driving! The adiabatic theory, see Eq. (7.17), falls on to the line with $\Omega=0.01$. After Bartussek, Hänggi, and Kissner (1994).

such as static external force fields, field gradients of thermal, chemical, or other origin. Hence the acting non-equilibrium forces of zero ensemble average are spatially uniform, and generally are statistically symmetric. The same principle applies if the stationary nonequilibrium forces are substituted by a spatially uniform, coherent periodic signal $F(t)$ of zero temporal average (Magnasco, 1993; Ajdari *et al.*, 1994; Bartussek, Hänggi, and Kissner, 1994). These systems are thus closely related in spirit to the stochastic resonance phenomenon: Fluctuation-induced escape among neighboring states in a periodic, multistable potential supported by weak deterministic periodic signals is responsible for moving particles forward noisily. In short, a deterministic driving alone, which exceeds a lower threshold is sufficient to create an induced current. Increasing at fixed angular frequency Ω the driving strength in overdamped, deterministic ratchet dynamics reveals numerous interesting features such as a devil's staircase behavior of current vs driving strength, current-quantization phenomena, and further peculiar features. In the presence of thermal Nyquist noise $\zeta(t)$, with $\langle \zeta(t) \rangle = 0$ and correlation $\langle \zeta(t)\zeta(t') \rangle = 2D\delta(t-t')$ the noisy, periodically driven, overdamped rocking ratchet dynamics reads

$$\dot{x} = -\frac{d}{dx} V_R(x) + A_0 \cos(\Omega t) + \zeta(t). \quad (7.16)$$

Herein $V_R(x)$ is the periodic (period L) sawtooth-like ratchet potential $V_R(x) = V_R(x+L)$ possessing no reflection symmetry $V_R(x) \neq V_R(-x)$. In Fig. 54 we depict the noise-induced current versus the thermal intensity D . Noteworthy in Fig. 54 is the stochastic

resonance-like feature of the probability current \bar{J} vs D characteristics, as well as the phenomenon of *current reversal* (Bartussek, Hänggi, and Kissner, 1994) that occurs for nonadiabatic driving frequencies Ω . The mean velocity of the particle position is given by $\langle \dot{x} \rangle = L \bar{J}$, where \bar{J} is the time-averaged probability current. Within an adiabatic approximation, i.e., for slow driving ($\Omega \rightarrow 0$) the current \bar{J} reads explicitly

$$\bar{J} = \frac{D}{2\pi/\Omega} \int_0^{2\pi/\Omega} dt \left\{ \left[1 - \exp[\Phi(L, t)] \right]^{-1} \times \int_0^L dx \int_0^L dy \exp[\Phi(y, t) - \Phi(x, t)] - \int_0^L dx \int_0^x dy \exp[\Phi(y, t) - \Phi(x, t)] \right\}^{-1}, \quad (7.17)$$

where $\Phi(x, t) = [V_R(x) - xA_0 \cos(\Omega t)]/D$.

In Fig. 54 this approximation for the smallest frequency $\Omega = 0.01$ coincides within line thickness with the exact Floquet-theory result (solid line). The effects of inertia and/or weak friction are also intriguing: In the absence of thermal noise, the characteristic chaotic motion is sufficient to induce a directed current J , which exhibits multiple current reversals vs the driving amplitude A_0 (Jung *et al.*, 1996).

Another ratchet type that is related closely to the rocking ratchet in Eq. (7.16) is obtained if one substitutes the external coherent driving by an oscillating temperature, i.e.,

$$A_0 \cos(\Omega t) \rightarrow \zeta(t) [1 + A_0 \cos(\Omega t)]. \quad (7.18)$$

This defines a *diffusion ratchet* (Reimann *et al.*, 1996), which tends to resist carrying a finite current in the asymptotic limits of fast and slow driving. In this case the current starts only proportional to Ω^2 , as $\Omega \rightarrow 0$ (i.e., a zero net current in the leading-order adiabatic approximation) and vanishes again proportional to Ω^{-2} , as $\Omega \rightarrow \infty$.

3. Escape rates in periodically driven systems

The problem of activated rates in threshold systems that are exposed to noise and periodic perturbations is nontrivial. The phase $\Omega t \equiv \theta$ of the periodic driving constitutes an additional dimension that can be used to define the escape rate out of a basin of attraction in extended space (Jung and Hänggi, 1991b). The basin of attraction is then shown to be separated by an unstable periodic orbit in extended x - θ space. For a bistable potential, this rate is related to the smallest nonzero Floquet eigenvalue μ . Note that this very quantity rules the long-time relaxation of general statistical quantities such as time-dependent mean values or time-averaged correlations. This positive-valued, rate-determining eigenvalue, being at weak noise well separated from higher-order relaxational Floquet eigenvalues, has been investigated for a driven, symmetric double well by Jung (1989; see Figs. 4–6 therein). For the peculiarities that

occur in periodically driven dynamics in a periodic potential, where the rate-determining Floquet eigenvalue is related to the Floquet eigenvalue at one of the two boundaries of the first Brillouin zone, we refer the reader to the discussion given in Jung and Hänggi (1991b). A main result is that the rate (or Floquet eigenvalue), which increases proportional to A_0^2 , does not exhibit any kind of resonance-like behavior! Thus, as repeatedly demonstrated in this review, the stochastic resonance phenomenon is not due to a resonance for the rate of escape in the periodically driven system. In recent studies by Reichl and her collaborators (Alpatov and Reichl, 1994; Kim and Reichl, 1996) the higher-order Floquet eigenvalues have been investigated by formally mapping the periodically driven overdamped system onto an equivalent quantum dynamics at imaginary times. As a main result they find that in the regime where the quantum system exhibits a transition to chaos the spectrum of Floquet eigenvalues shows level repulsion.

Yet another quantity related to the rate-determining Floquet eigenvalue is the diffusion coefficient in a tilted periodic potential. Here, the interplay of subthreshold static drive and thermal fluctuations results in an enhancement of the ensuing stationary current, with a maximum for a certain value of the temperature. Such an effect, not observable in the overdamped limit (Hu, 1993; Gittermann, Khalfin, and Shapiro, 1994; Casado, Mejías, and Morillo, 1995), may be important in *weakly* damped systems (Marchesoni, 1997). If the static drive is replaced by a periodic tilt with suprathreshold amplitude, novel effects can be induced such as an enhancement of the diffusion rate that even exceeds (Hu, Daffnersthofer, and Haken, 1996) the rate of free diffusion! Likewise, the role of suprathreshold driving strengths applied to stochastic resonance systems has recently been studied for the phenomenon of a noise-induced failure mechanism for driven switch transitions (Apostolico *et al.*, 1997).

VIII. CONCLUSIONS AND OUTLOOK

In this review we have shown that adding noise to a system can sometimes improve its ability to transfer information reliably. This phenomenon—known as *stochastic resonance*—was originally proposed, almost seventeen years ago, to account for the periodicity in the Earth's ice ages, but has since been shown to occur in many systems. By now, understanding of the phenomenon of stochastic resonance has reached a mature level that we have attempted to review with this long paper. Numerous contributions to stochastic resonance have appeared in most physics journals and can be found scattered through many other scientific journals (e.g., see <http://www.umbrars.com/sr>), particularly in the fields of biology and physiology; it thus has reached the level of what we may term an “industry.”

Undoubtedly, the neurophysiological applications represent cornerstones in the field of stochastic resonance. Such applications have attracted continued inter-

est from scientists in biology, biomedical engineering, and medicine. A new interdisciplinary field beyond conventional stochastic resonance, with inputs from nonlinear dynamics, nonequilibrium statistical physics, biological and medical sciences has emerged.

The general feature of a system exhibiting stochastic resonance is its increased sensitivity to small perturbations when an appropriate dose of noise is added; see the Introduction and Sec. II. The idea that random noise can be beneficial to the formation of “order” sounds paradoxical, but must be taken seriously by now. Stochastic resonance simply stands for a new paradigm wherein noise represents a useful tool, rather than a nuisance. Given the basic three ingredients of stochastic resonance, which are (1) a form of threshold, (2) a source of “noise,” and (3) a generally weak input source, it is clear that stochastic resonance is generic enough to be observable in a large variety of systems.

For the benefit of the reader let us summarize here what we think has been achieved in the field of stochastic resonance, point out open questions, and finally share our views about future perspectives.

For simple physical systems that can be described by either one of the two generic models introduced in Sec. III (two-state model) and in Secs. IV.A and IV.B (continuous-state bistable model), the mechanism that underpins the stochastic resonance effect is by now well understood. The increased response of the system in the presence of noise is due to the *synchronization* of noise-induced hopping with the temporal profile of the weak perturbation. This response of the system is ruled by two competing aspects: starting out from the zero-noise limit, increasing noise allows—correlated with the small perturbation—excursions into the neighboring well. This causes an increased response. On the other hand, increasing the noise level counteracts the aforementioned correlation; thereby reducing the response. These two aspects are encoded mathematically by the susceptibility χ , which essentially is made up of two factors: the product of an Arrhenius factor, describing the activated hopping, and a factor proportional to the inverse noise intensity $1/D$, characterizing the degradation of the response. The result is a bell-shaped curve for the response amplitude vs noise intensity, hence the term stochastic resonance—an expression that for some may appear ill-defined. This physics in turn determines the most common quantifiers for stochastic resonance: the signal-to-noise ratio (SNR) of the output, and the spectral (power) amplification η ; see Secs. II.A and IV.B. Likewise, the statistical features of the driven residence-time distributions $N(T)$ (see Sec. IV.C), reflect the synchronization between random hopping and external modulation. The multi peaked signatures exhibited by the residence-time distribution at *odd* multiples of half the driving period are nothing but the fingerprints of this synchronization process that occurs in the competition between the active driving source and the passive dissipation. It should not go unnoticed that the understanding of this very concept of driven residence-time distri-

butions paved the way to interpreting a mass of physiological data from a new viewpoint; see Secs. IV.C and V.C.

Equipped with the basics of stochastic resonance theory, developed in some detail in Secs. III and IV, we discussed and interpreted in Sec. V several prominent applications and experiments taken from the fields of physics and neurophysiology (a more detailed list of experimental stochastic resonance demonstrations is given in Sec. II.C.3).

More recent developments in the field of stochastic resonance presently in the limelight of the activities of many research groups are discussed in Sec. VI. Both quantum stochastic resonance and spatiotemporal stochastic resonance have only just begun to be explored. Quantum tunneling assists the stochastic resonance effect in the semiclassical regime; in the deep cold, however, quantum coherence increasingly spoils the effect; see Sec. VI.A. The notion of stochastic resonance generalized to spatially extended pattern-forming systems has been the subject of Sec. VI.B: spatiotemporal patterns can be enhanced by adding the proper amount of noise. The notion of deterministic chaos, which intrinsically provides a source of disorder, has been studied by various groups, and has been reviewed in Sec. VI.C. The last section, Sec. VI.D, has been devoted to the study of the effects of finite correlation times (colored noise) of the background noise. For overdamped dynamics the role of colored noise generally results in a reduction of the efficiency of stochastic resonance. In contrast, finite inertia effects, induced by moderate friction, tend to boost the stochastic resonance response. The physics of stochastic resonance at extreme weak friction, however, still needs to be investigated in greater detail.

The field of stochastic resonance research has witnessed a remarkable flourishing during the last few years; needless to say it is no longer possible to present a detailed account on each single contribution. In Sec. VII we have discussed selected contributions that provide additional insight. Stochastic resonance and its connection with the dithering effect, globally coupled periodically modulated bistable elements, or the impact of additional multiplicative noise on stochastic resonance in the presence of additive noise (see Sec. VII.C.1) are all topics that relate closely to stochastic resonance. Recently, research on stochastic resonance in physical systems has diverged into neighboring fields such as the problem of noise-induced transport in Brownian ratchets; see Sec. VII.D.2. The modern topic of nonlinear stochastic resonance involves the impact of noise on the generation and mixing of higher harmonics; see Sec. VII.D.1. A peculiar effect, the suppression of higher harmonics at some specific noise strengths, is being reported there. This effective elimination of higher harmonics could be used to the effect of minimizing the distortion of information transfer in nonlinear systems.

In Sec. VII.C.3 on aperiodic stochastic resonance—i.e., the phenomenon being obtained in presence of a nonperiodic input signal—we touched on a still open problem. Can stochastic resonance be suitably charac-

terized by means of information theory concepts alone? In view of quantum stochastic resonance, which intrinsically avoids the notion of joint probability measures, a unifying answer seems anything but trivial.

It may be worthwhile to conclude with some speculations on what the future of stochastic resonance may look like. What is still lacking from a physics point of view is a detailed, microscopic approach to stochastic resonance that would account for the mutual interplay between the transfer of power among the system $x(t)$, the bath(s) [or sources of noise $\xi(t)$], and the external signal $A(t)$. Another promising area for fruitful further research is quantum stochastic resonance. For example, almost nothing is known about the quantum analog of stochastic resonance in threshold-crossing devices, stochastic resonance in arrays of coupled quantum systems, or—last but not least—the difficult problem of modeling quantum stochastic resonance in stationary nonequilibrium systems (i.e., the physics occurring in driven, dissipative quantum systems that are far from thermal equilibrium). The observation that classical concepts become increasingly invalid upon crossing the borderline between the classical and the quantum world, and beyond, is an indication that several surprises and novel stochastic resonance phenomena are waiting to be uncovered. The same holds true for spatiotemporal stochastic resonance, which yet has to be extended into three-dimensional structures.

Clearly, stochastic resonance constitutes an information-transmitting phenomenon that exploits the noise in a self-optimizing manner. Therefore, its promising role in complex systems such as the nervous system, or even the brain have not gone unnoticed in the communities of physiological, biological, and medical sciences; see the reviews by Moss, Pierson, and O’Gorman (1994) and Wiesenfeld and Moss (1995). For example, the question of whether extremely low-frequency electromagnetic fields actually affect biological function via the stochastic resonance phenomenon still remains open. What has been achieved so far is the successful demonstration of stochastic resonance with injected external noise in the peripheral nervous system of crayfish (Douglass *et al.*, 1993; Pei *et al.*, 1996), in crickets (Levin and Miller, 1996), in the human visual perception (Riani and Simonotto, 1995; Simonotto *et al.*, 1997), and in ion channels (Bezrukov and Vodyanoy, 1995), to name a few. Without doubt this latter area, too, is expected to prosper by providing numerous interesting new results and novel insights.

ACKNOWLEDGEMENTS

P.J. and P.H. would like to thank the Deutsche Forschungsgemeinschaft for financial support. F.M. and L.G. wish to thank the Istituto Nazionale di Fisica della Materia (INFM) and CNR for partial funding. F.M. would further like to thank B. I. Halperin for his kind hospitality at the Physical Laboratories of the Harvard University. Most of the analog simulations presented herein were carried out in collaboration with E. Menichella-

Saetta and S. Santucci (Perugia). We greatly appreciate insightful discussions with Th. Anastasio, R. Bartussek, A. Bulsara, D. K. Campbell, J. Collins, Th. Dittrich, M. Dykman, W. Ebeling, J. Grohs, R. F. Fox, H. Frauenfelder, G. Hu, M. Grifoni, A. Jackson, L. Kiss, A. Kittel, A. Longtin, R. Mantegna, G. Mayer-Kress, F. Moss, J. Parisi, E. Pollak, C. Presilla, H. Risken, R. Roy, L. Schimansky-Geier, B. Spagnolo, A. Suter, W. Sung, P. Talkner, A. Vulpiani, U. Weiss, K. Wiesenfeld, and C. Zerbe. Last but not least, we would like to thank our indispensable Mrs. Eva Seehuber for her continuous help while preparing this work.

P.H., P.J., and F.M. are deeply indebted to our friend and mentor Hannes Risken, whose wisdom and friendship we are missing deeply after his untimely death in January 1994.

APPENDIX: PERTURBATION THEORY

For weak external forcing, the time-inhomogeneous term in the Fokker-Planck equation can be treated as a small perturbation with the amplitude A_0 acting as small parameter. Perturbation theory (Presilla *et al.*, 1989; Hu *et al.*, 1990; Jung, 1993) provides explicit expressions for characteristic quantities of the driven systems, such as the Floquet eigenvalues, eigenfunctions, and mean values in terms of the eigenvalues and eigenfunctions of the unperturbed Fokker-Planck operator.

The starting point is the Floquet eigenvalue problem, obtained in Sec. IV.A by inserting the Floquet ansatz (4.9) into the Fokker-Planck equation (4.11) with $\varphi=0$

$$\left[\mathcal{L}_0 - A_0 \cos(\Omega t) \frac{\partial}{\partial x} - \frac{\partial}{\partial t} \right] p_\mu(x, t; 0) = -\mu p_\mu(x, t; 0), \quad (\text{A1})$$

with

$$\mathcal{L}_0 = -\frac{\partial}{\partial x} f(x) + D \frac{\partial^2}{\partial x^2}, \quad (\text{A2})$$

and

$$f(x) = x - x^3. \quad (\text{A3})$$

Expanding the Floquet eigenfunctions into a Fourier series

$$p_\mu(x, t; 0) = \sum_{n=-\infty}^{\infty} c_n(x) \exp[in\Omega t] \quad (\text{A4})$$

yields the hierarchy of coupled ordinary differential equations

$$(\mathcal{L}_0 - in\Omega + \mu)c_n(x) - \frac{A_0}{2}[c'_{n+1}(x) + c'_{n-1}(x)] = 0, \quad (\text{A5})$$

with $c'_n(x) = dc_n(x)/dx$.

For strictly real-valued Floquet eigenvalues μ , the Floquet eigenfunctions are real as well, i.e.,

$$c_n(x) = c_n^*(x). \quad (\text{A6})$$

In order for the perturbation theory to be applicable to more general situations such as tilted periodic potentials, we do not make use of this assumption. The zeroth order perturbation theory for $A_0 = 0$

$$(\mathcal{L}_0 - i n \Omega + \mu^{(0)}) c_n^{(0)}(x) = 0 \quad (\text{A7})$$

is solved by

$$\begin{aligned} c_n^{(0)}(x) &= \psi_l(x), \\ \mu_{ln}^{(0)} &= \lambda_l + i n \Omega, \end{aligned} \quad (\text{A8})$$

where the index l of $\{c_n(x)\}$ has been dropped throughout for convenience. Here, $\psi_l(x)$, $\psi_l^\dagger(x)$, and λ_l are the eigenfunctions and eigenvalues of the unperturbed (adjoint) Fokker-Planck operator, i.e.,

$$\begin{aligned} \mathcal{L}_0 \psi_l(x) &= -\lambda_l \psi_l(x), \\ \mathcal{L}_0^\dagger \psi_l^\dagger(x) &= -\lambda_l \psi_l^\dagger(x), \\ \langle l | m \rangle &\equiv \int_{-\infty}^{\infty} \psi_l(x) \psi_m^\dagger(x) dx = \delta_{lm}. \end{aligned} \quad (\text{A9})$$

To order $(A_0)^0$, all equations with label n have the same form. The eigenvalues differ by a multiple of $i\Omega$. Without loss of generality, we can start the perturbation expansion at $n=0$, i.e.,

$$\begin{aligned} \mu_{ln=0}^{(0)} &\equiv \mu_l^{(0)} = \lambda_l, \\ c_n^{(0)}(x) &= \delta_{n0} \psi_l(x). \end{aligned} \quad (\text{A10})$$

The first three equations of Eq. (A5) are written down explicitly as

$$(\mathcal{L}_0 + \mu) c_0(x) = \frac{A_0}{2} [c_1'(x) + c_{-1}'(x)], \quad (\text{A11})$$

$$(\mathcal{L}_0 - i\Omega + \mu) c_1(x) = \frac{A_0}{2} [c_2'(x) + c_0'(x)],$$

$$(\mathcal{L}_0 + i\Omega + \mu) c_{-1}(x) = \frac{A_0}{2} [c_0'(x) + c_{-2}'(x)]. \quad (\text{A12})$$

We now seek a solution of the latter system of differential equations in terms of the perturbation expansions

$$\begin{aligned} c_n(x) &= \delta_{n0} \psi_l(x) + A_0 c_n^{(1)}(x) + A_0^2 c_n^{(2)}(x) + \dots, \\ \mu_l &= \lambda_l + A_0 \mu_l^{(1)} + A_0^2 \mu_l^{(2)} + \dots. \end{aligned} \quad (\text{A13})$$

It follows immediately from Eq. (A11) that $\mu^{(1)}$ vanishes. Comparing terms of order A_0 and A_0^2 in Eqs. (A11) and (A12) yields

$$(\mathcal{L}_0 + \lambda_l) c_0^{(2)}(x) + \mu_l^{(2)} \psi_l(x) = \frac{1}{2} [c_1^{(1)'}(x) + c_{-1}^{(1)'}(x)], \quad (\text{A14})$$

$$\begin{aligned} (\mathcal{L}_0 - i\Omega + \lambda_l) c_1^{(1)}(x) &= \frac{1}{2} \psi_l'(x), \\ (\mathcal{L}_0 + i\Omega + \lambda_l) c_{-1}^{(1)}(x) &= \frac{1}{2} \psi_l'(x). \end{aligned} \quad (\text{A15})$$

Since $\lambda_l \pm i\Omega$ are not eigenvalues of the operator \mathcal{L}_0 , the operators $\mathcal{L}_0 + \lambda_l \pm i\Omega$ can be inverted and the functions $c_{\pm 1}^{(1)}(x)$ are obtained formally as

$$\begin{aligned} c_1^{(1)}(x) &= \frac{1}{2} (\mathcal{L}_0 - i\Omega + \lambda_l)^{-1} \psi_l'(x), \\ c_{-1}^{(1)}(x) &= \frac{1}{2} (\mathcal{L}_0 + i\Omega + \lambda_l)^{-1} \psi_l'(x). \end{aligned} \quad (\text{A16})$$

Inserting Eq. (A16) into Eq. (A14), multiplying by the eigenfunction ψ_l^\dagger and then integrating over x , we obtain

$$\begin{aligned} \mu_l^{(2)} &= \frac{1}{4} \int_{-\infty}^{\infty} \psi_l^\dagger(x) \frac{\partial}{\partial x} \mathcal{L}_1 \frac{\partial}{\partial x} \psi_l(x) dx \\ &= \frac{1}{4} \left\langle l \left| \frac{\partial}{\partial x} \mathcal{L}_1 \frac{\partial}{\partial x} \right| l \right\rangle, \end{aligned} \quad (\text{A17})$$

where

$$\mathcal{L}_1 = \frac{1}{\mathcal{L}_0 + i\Omega + \lambda_l} + \frac{1}{\mathcal{L}_0 - i\Omega + \lambda_l}. \quad (\text{A18})$$

For the Floquet eigenfunctions one obtains in leading-order perturbation theory

$$\begin{aligned} p_{\mu_l}(x, t) &= \psi_l(x) + \frac{A_0}{2} \{ \exp(-i\Omega t) [\mathcal{L}_0 + i\Omega + \lambda_l]^{-1} \\ &\quad + \exp(i\Omega t) [\mathcal{L}_0 - i\Omega + \lambda_l]^{-1} \} \psi_l'(x). \end{aligned} \quad (\text{A19})$$

In view of the identity

$$\begin{aligned} \frac{\partial}{\partial x} \psi_l(x) &= \sum_{q=0}^{\infty} \psi_q(x) \int_{-\infty}^{\infty} \psi_q^\dagger(x) \frac{\partial}{\partial x} \psi_l(x) dx \\ &= \sum_{q=0}^{\infty} \left\langle q \left| \frac{\partial}{\partial x} \right| l \right\rangle \psi_q(x), \end{aligned} \quad (\text{A20})$$

the Floquet eigenvalues and eigenfunctions can be expressed in terms of the eigenfunctions of the undriven Fokker-Planck operator, namely

$$\begin{aligned} \mu_l &= \lambda_l + \frac{A_0^2}{2} \sum_{q=0}^{\infty} \frac{\lambda_l - \lambda_q}{(\lambda_l - \lambda_q)^2 + \Omega^2} \left\langle l \left| \frac{\partial}{\partial x} \right| q \right\rangle \\ &\quad \times \left\langle q \left| \frac{\partial}{\partial x} \right| l \right\rangle \\ p_{\mu_l}(x, t) &= \psi_l(x) + A_0 \sum_{q=0}^{\infty} \frac{1}{\sqrt{(\lambda_l - \lambda_q)^2 + \Omega^2}} \\ &\quad \times \left\langle q \left| \frac{\partial}{\partial x} \right| l \right\rangle \cos(\Omega t + \alpha_{ql}) \psi_q(x), \end{aligned} \quad (\text{A21})$$

with

$$\tan(\alpha_{ql}) = \Omega/(\lambda_l - \lambda_q). \quad (\text{A22})$$

In particular, the asymptotic probability density $p_{as}(x, t)$, corresponding to the vanishing Floquet eigenvalue $\mu_0 = \lambda_0 = 0$, reads

$$\begin{aligned} p_{as}(x, t) &= p_0(x) + A_0 \sum_{q=1}^{\infty} \left\langle q \left| \frac{\partial}{\partial x} \right| 0 \right\rangle \\ &\quad \times \psi_q(x) \frac{1}{\sqrt{\lambda_q^2 + \Omega^2}} \cos(\Omega t + \alpha_{q0}) \\ &= p_0(x) + A_s(x) \sin(\Omega t) + A_c(x) \cos(\Omega t), \end{aligned} \quad (\text{A23})$$

with

$$\begin{aligned} A_c(x) &= A_0 \sum_{q=1}^{\infty} \frac{\lambda_q}{\lambda_q^2 + \Omega^2} \left\langle q \left| \frac{\partial}{\partial x} \right| 0 \right\rangle \psi_q(x), \\ A_s(x) &= A_0 \sum_{q=1}^{\infty} \frac{\Omega}{\lambda_q^2 + \Omega^2} \left\langle q \left| \frac{\partial}{\partial x} \right| 0 \right\rangle \psi_q(x). \end{aligned} \quad (\text{A24})$$

The spectral amplification [Eq. (4.21)] is obtained by inserting Eq. (A23) into the definition $\langle x(t) \rangle_{as} = \int x p_{as}(x, t) dx$, that is,

$$\begin{aligned} \eta &= \sum_{n,m=0}^{\infty} \frac{\lambda_n \lambda_m + \Omega^2}{(\lambda_n^2 + \Omega^2)(\lambda_m^2 + \Omega^2)} \left\langle n \left| \frac{\partial}{\partial x} \right| 0 \right\rangle \\ &\quad \times \left\langle m \left| \frac{\partial}{\partial x} \right| 0 \right\rangle \langle 0|x|n \rangle \langle 0|x|m \rangle. \end{aligned} \quad (\text{A25})$$

The expression (A25) is exact up to order $(A_0)^2$.

The expression for $\langle x(t) \rangle_{as}$ to leading order in A_0 coincides with the prediction of the thermal-equilibrium linear-response theory of Sec. IV.B. This last statement can be proved explicitly by inserting the completeness relation into the expression

$$\chi(t) = \int x e^{\mathcal{L}_0 t} \left(-\frac{\partial}{\partial x} \right) p_0(x) dx \quad (\text{A26})$$

for the response function $\chi(t)$ of the modulated system of Eq. (A.1), whence

$$\chi(t) = - \sum_{q=1}^{\infty} e^{-\lambda_q t} \left\langle q \left| \frac{\partial}{\partial x} \right| 0 \right\rangle \langle 0|x|q \rangle. \quad (\text{A27})$$

Note that Eq. (A26) follows immediately from the general definition (4.28) of $\chi(t)$ by substituting $\Gamma_{ext} = \delta'(y - z)$ for the perturbation kernel, and $P_0(x, t|y, 0) = e^{\mathcal{L}_0 t} \delta(x - y)$ for the unperturbed conditional probability density of the system under study. On further substituting the spectral representation (A27) of $\chi(t)$ into Eq. (4.26), we eventually reproduce the perturbation theory prediction for $\langle x(t) \rangle_{as}$:

$$\begin{aligned} \langle x(t) \rangle_{as} &= A_0 \sum_{q=1}^{\infty} \left\langle q \left| \frac{\partial}{\partial x} \right| 0 \right\rangle \langle 0|x|q \rangle \frac{1}{\sqrt{\lambda_q^2 + \Omega^2}} \\ &\quad \times \cos(\Omega t + \alpha_{q0}). \end{aligned} \quad (\text{A28})$$

On approximating $\langle 0|x|1 \rangle$ to 1 and $\langle 1|\partial/\partial x|0 \rangle$ to $-\lambda_1/D$, one eventually recovers Eq. (2.7a) for $\bar{x}(D)$ (Hu *et al.*, 1990).

REFERENCES

- Abramowitz, M., and I. A. Stegun, 1965, *Handbook of Mathematical Functions* (Dover, New York).
- Ajdari, A., D. Mukamel, L. Peliti, and J. Prost, 1994, *J. Phys. I* **4**, 1551.
- Ajdari, A., and J. Prost, 1992, *C.R. Acad. Sci.* **315**, 1635.
- Alpatov, P., and L. E. Reichl, 1994, *Phys. Rev. E* **49**, 2630.
- Amit, D. J., 1984, *Field Theory, The Renormalization Group and Critical Phenomena* (World Scientific, Singapore).
- Amit, D. J., 1989, *Modeling Brain Functions* (Cambridge University, Cambridge).
- Anishchenko, V. S., A. B. Neimann, and M. A. Safanova, 1993, *J. Stat. Phys.* **70**, 183.
- Anishchenko, V. S., M. A. Safonova, and L. O. Chua, 1994, *Int. J. Bifurcation Chaos Appl. Sci. Eng.* **4**, 441.
- Apostolico, F., L. Gammaitoni, F. Marchesoni, and S. Santucci, 1997, *Phys. Rev. E* **55**, 36.
- Arimondo, E., D. Dangoisse, E. Menchi, and F. Papoff, 1987, *J. Opt. Soc. Am. B* **4**, 892.
- Arimondo, E., and B. M. Dinelli, 1983, *Opt. Commun.* **44**, 277.
- Astumian, R. D., 1997, *Science* **276**, 917.
- Astumian, R. D., and M. Bier, 1994, *Phys. Rev. Lett.* **72**, 1766.
- Bartussek, R., P. Hänggi, and P. Jung, 1994, *Phys. Rev. E* **49**, 3930.
- Bartussek, R., P. Hänggi, and J. G. Kissner, 1994, *Europhys. Lett.* **28**, 459.
- Bartussek, R., P. Jung, and P. Hänggi, 1993, in *Noise in Physical Systems and 1/f Fluctuations*, edited by P. H. Handel, AIP Conf. Proc. 285, 661 (AIP, New York, 1993).
- Barzykin, A. V., and K. Seki, 1997, *Europhys. Lett.* **40**, 117.
- Benderskii, V. A., D. E. Makarov, and C. A. Wight, 1994, *Adv. Chem. Phys.* **85**, 1.
- Bennet, W. R., 1948, *Bell Syst. Tech. J.* **27**, 446.
- Benzi, R., G. Parisi, A. Sutera, and A. Vulpiani, 1982, *Tellus* **34**, 10.
- Benzi, R., A. Sutera, G. Parisi, and A. Vulpiani, 1983, *SIAM (Soc. Ind. Appl. Math.) J. Appl. Math.* **43**, 565.
- Benzi, R., A. Sutera, and A. Vulpiani, 1981, *J. Phys. A* **14**, L453.
- Benzi, R., A. Sutera, and A. Vulpiani, 1985, *J. Phys. A* **18**, 2239.
- Berdichevsky, V., and M. Gitterman, 1996, *J. Phys. A* **29**, L447.
- Berghaus, C., A. Hilgers, and J. Schnakenberg, 1996, *Z. Phys. B* **100**, 157.
- Bezrukov, S. M., and I. Vodyanoy, 1995, *Nature (London)* **378**, 362.
- Bezrukov, S. M., and I. Vodyanoy, 1997, *Nature (London)* **385**, 319.
- Blake, I. F., and W. C. Lindsey, 1973, *IEEE Trans. Inf. Theory* **IT-19**, 295.
- Bonifacio, R., and L. A. Lugiato, 1978, *Phys. Rev. A* **18**, 1192.
- Brey, J. J., and A. Prados, 1996, *Phys. Lett. A* **216**, 240.

- Bruce, A. D., 1980, *Adv. Phys.* **29**, 111.
- Bulsara, A., S. Chillemi, L. Kiss, P. V. E. McClintock, R. Mannella, F. Marchesoni, G. Nicolis, and K. Wiesenfeld, 1995, Eds., *International Workshop on Fluctuations in Physics and Biology: Stochastic Resonance, Signal Processing and Related Phenomena*, published in *Nuovo Cimento* **17D**, 653.
- Bulsara, A., T. C. Elston, C. R. Doering, S. B. Lowen, and K. Lindenberg, 1996, *Phys. Rev. E* **53**, 3958.
- Bulsara, A., and L. Gammaitoni, 1996, *Phys. Today* **49**, No. 3, 39.
- Bulsara, A., M. Inchiosa, and L. Gammaitoni, 1996, *Phys. Rev. Lett.* **77**, 2162.
- Bulsara, A., E. W. Jacob, T. Zhou, F. Moss, and L. Kiss, 1991, *J. Theor. Biol.* **152**, 531.
- Bulsara, A., S. B. Lowen, and C. D. Rees, 1994, *Phys. Rev. E* **49**, 4989.
- Bulsara, A., and G. Schmera, 1993, *Phys. Rev. E* **47**, 3734.
- Bulsara, A., and A. Zador, 1996, *Phys. Rev. E* **54**, R2185.
- Carroll, T. L., and L. M. Pecora, 1993a, *Phys. Rev. Lett.* **70**, 576.
- Carroll, T. L., and L. M. Pecora, 1993b, *Phys. Rev. E* **47**, 3941.
- Casado, J. M., J. J. Mejías, and M. Morillo, 1995, *Phys. Lett. A* **197**, 365.
- Castelpoggi, F., and H. S. Wio, 1997, *Europhys. Lett.* **38**, 91.
- Chapeau-Blondeau, F., 1997, *Phys. Rev. E* **55**, 2016.
- Chialvo, D. R., and A. V. Apkarian, *J. Stat. Phys.* **70**, 375.
- Choi, M., R. F. Fox, and P. Jung, 1998, *Phys. Rev. E* (in press).
- Chun, K., and N. O. Birge, 1993, *Phys. Rev. E* **48**, 11 4500.
- Claes, I., and C. Van den Broeck, 1991, *Phys. Rev. A* **44**, 4970.
- Clarke, J., A. N. Cleland, M. H. Devoret, D. Esteve, and J. M. Martinis, 1988, *Science* **239**, 992.
- Collins, J. J., C. C. Chow, A. C. Capela, and T. T. Imhoff, 1996, *Phys. Rev. E* **54**, R5575.
- Collins, J. J., C. C. Chow, and T. T. Imhoff, 1995a, *Phys. Rev. E* **52**, R3321.
- Collins, J. J., C. C. Chow, and T. T. Imhoff, 1995b, *Nature (London)* **376**, 236.
- Collins, J. J., T. T. Imhoff, and P. Grigg, 1996, *J. Neurophysiol.* **76**, 642.
- Coppinger, F., J. Genoe, D. K. Maude, U. Gennser, J. C. Portal, K. E. Singer, P. Rutter, T. Taskin, A. R. Peaker, and A. C. Wright, 1995, *Phys. Rev. Lett.* **75**, 3513.
- Crisanti, A., M. Falcioni, G. Paladin, and A. Vulpiani, 1994, *J. Phys. A* **27**, L597.
- Dakhnovskii, Yu., and R. D. Coalson, 1995, *J. Chem. Phys.* **103**, 2908.
- Dayan, I., M. Gitterman, and G. H. Weiss, 1992, *Phys. Rev. A* **46**, 757.
- Debnath, G., T. Zhou, and F. Moss, 1989, *Phys. Rev. A* **39**, 4323.
- Desai, R., and R. Zwanzig, 1978, *J. Stat. Phys.* **19**, 1.
- De Weese, M., and W. Bialek, 1995, *Nuovo Cimento* **17D**, 733.
- Dewel, G., P. Borckmanns, D. Walgraef, 1985, *Phys. Rev. A* **31**, 1983.
- Dittrich, T., B. Oelschlägel, and P. Hänggi, 1993, *Europhys. Lett.* **22**, 5.
- Doering, C. R., W. Horsthemke, and J. Riordan, 1994, *Phys. Rev. Lett.* **72**, 2984.
- Douglass, J. K., L. Wilkens, E. Pantazelou, and F. Moss, 1993, *Nature (London)* **365**, 337.
- Drozhdov, A. N., and M. Morillo, 1996, *Phys. Rev. E* **54**, 3304.
- Dykman, M. I., G. P. Golubev, I. Kh. Kaufman, D. G. Luchinsky, P. V. E. McClintock, and E. A. Zhukov, 1995, *Appl. Phys. Lett.* **67**, 308.
- Dykman, M. I., H. Haken, G. Hu, D. G. Luchinsky, R. Mannella, P. V. E. McClintock, C. Z. Ning, N. D. Stein, and N. G. Stocks, 1993, *Phys. Lett. A* **180**, 332.
- Dykman, M. I., D. G. Luchinsky, R. Mannella, P. V. E. McClintock, H. E. Short, N. D. Stein, and N. G. Stocks, 1994, *Phys. Rev. E* **49**, 1198.
- Dykman, M. I., D. G. Luchinsky, R. Mannella, P. V. E. McClintock, N. D. Stein, and N. G. Stocks, 1995, *Nuovo Cimento D* **17**, 661.
- Dykman, M. I., D. G. Luchinsky, P. V. E. McClintock, N. D. Stein, and N. G. Stocks, 1992, *Phys. Rev. A* **46**, R1713.
- Dykman, M. I., R. Mannella, P. V. E. McClintock, N. D. Stein, and N. G. Stocks, 1993a, *Phys. Rev. E* **47**, 1629.
- Dykman, M. I., R. Mannella, P. V. E. McClintock, N. D. Stein, and N. G. Stocks, 1993b, *Phys. Rev. E* **47**, 3996.
- Dykman, M. I., R. Mannella, P. V. E. McClintock, and N. G. Stocks, 1990a, *Phys. Rev. Lett.* **65**, 48.
- Dykman, M. I., R. Mannella, P. V. E. McClintock, and N. G. Stocks, 1990b, *Phys. Rev. Lett.* **65**, 2606.
- Dykman, M. I., R. Mannella, P. V. E. McClintock, and N. G. Stocks, 1992, *Phys. Rev. Lett.* **68**, 2985.
- Dykman, M. I., R. Mannella, P. V. E. McClintock, and N. G. Stocks, 1993, *Phys. Rev. Lett.* **70**, 874.
- Eckmann, J-P., and L. E. Thomas, 1982, *J. Phys. A* **15**, L261.
- Eigler, D. M., and E. K. Schweitzer, 1990, *Nature (London)* **344**, 523.
- Faetti, S., P. Grigolini, and F. Marchesoni, 1982, *Z. Phys. B* **47**, 353.
- Fauve, S., and F. Heslot, 1983, *Phys. Lett.* **97A**, 5.
- Feynman, R. P., R. B. Leighton, and M. Sands, 1966, *The Feynman Lectures on Physics* (Addison-Wesley, Reading, MA), Vol. I, Chap. 46.
- Fioretti, A., L. Guidoni, R. Mannella, and E. Arimondo, 1993, *J. Stat. Phys.* **70**, 403.
- Floquet, G., 1883, *Ann. de l'Ecole Norm. Suppl.* **12**, 47.
- Fox, R. F., 1978, *Phys. Rep.* **48**, 179.
- Fox, R. F., 1989, *Phys. Rev. A* **39**, 4148.
- Fox, R., and Y. Lu, 1993, *Phys. Rev. E* **48**, 3390.
- Fronzoni, L., F. Moss, and P. V. E. McClintock, 1989, eds., *Noise in Nonlinear Dynamical Systems*, Vol. 3 (Cambridge University, Cambridge), p. 222.
- Fuliński, A., 1995, *Phys. Rev. E* **52**, 4523.
- Gage, E. C., and L. Mandel, 1988, *Phys. Rev. A* **38**, 5166.
- Gammaitoni, L., 1995a, *Phys. Rev. E* **52**, 4691.
- Gammaitoni, L., 1995b, *Phys. Lett. A* **208**, 315.
- Gammaitoni, L., and F. Marchesoni, 1993, *Phys. Rev. Lett.* **70**, 873.
- Gammaitoni, L., F. Marchesoni, M. Martinelli, L. Pardi, and S. Santucci, 1991, *Phys. Lett. A* **158**, 449.
- Gammaitoni, L., F. Marchesoni, E. Menichella-Saetta, and S. Santucci, 1989, *Phys. Rev. Lett.* **62**, 349.
- Gammaitoni, L., F. Marchesoni, E. Menichella-Saetta, and S. Santucci, 1990, *Phys. Rev. Lett.* **65**, 2607.
- Gammaitoni, L., F. Marchesoni, E. Menichella-Saetta, and S. Santucci, 1993, *Phys. Rev. Lett.* **71**, 3625.
- Gammaitoni, L., F. Marchesoni, E. Menichella-Saetta, and S. Santucci, 1994, *Phys. Rev. E* **49**, 4878.
- Gammaitoni, L., F. Marchesoni, E. Menichella-Saetta, and S. Santucci, 1995, *Phys. Rev. E* **51**, R3799.

- Gammaitoni, L., F. Marchesoni, and S. Santucci, 1994, *Phys. Lett. A* **195**, 116.
- Gammaitoni, L., F. Marchesoni, and S. Santucci, 1995, *Phys. Rev. Lett.* **74**, 1052.
- Gammaitoni, L., M. Martinelli, L. Pardi, and S. Santucci, 1991, *Phys. Rev. Lett.* **67**, 1799.
- Gammaitoni, L., M. Martinelli, L. Pardi, and S. Santucci, 1992, *Mod. Phys. Lett. B* **6**, 197.
- Gammaitoni, L., M. Martinelli, L. Pardi, and S. Santucci, 1993, *J. Stat. Phys.* **70**, 425.
- Gammaitoni, L., E. Menichella-Saetta, S. Santucci, and F. Marchesoni, 1989, *Phys. Lett. A* **142**, 59.
- Gammaitoni, L., E. Menichella-Saetta, S. Santucci, F. Marchesoni, and C. Presilla, 1989, *Phys. Rev. A* **40**, 2114.
- Gerstein, G. L., and B. Mandelbrot, 1964, *Biophys. J.* **4**, 41.
- Gingl, Z., L. B. Kiss, and F. Moss, 1995, *Europhys. Lett.* **29**, 191.
- Gitterman, M., I. B. Khalfin, and B. Ya. Shapiro, 1994, *Phys. Lett. A* **184**, 339.
- Gluckman, B. J., T. I. Netoff, E. J. Neel, W. L. Ditto, M. Spano, and S. J. Schiff, 1996, *Phys. Rev. Lett.* **77**, 4098.
- Goel, N. S., and N. Richter-Dyn, 1974, *Stochastic Models in Biology* (Academic, New York).
- Golding, B., N. M. Zimmermann, and S. N. Coppersmith, 1992, *Phys. Rev. Lett.* **68**, 998.
- Gómez-Ordóñez, J., and M. Morillo, 1994, *Phys. Rev. E* **49**, 4919.
- Gong, D., G. Hu, X. Wen, C. Yang, G. Qin, R. Li, and D. Ding, 1992, *Phys. Rev. A* **46**, 3243; *Phys. Rev. E* **48**, 4862 (E).
- Gong, D., G. R. Qin, G. Hu, and X. D. Weng, 1991, *Phys. Lett. A* **159**, 147.
- Goychuk, I. A., E. G. Petrov, and V. May, 1996, *Chem. Phys. Lett.* **353**, 428.
- Grabert, H., and H. Wipf, 1990, in *Festkörperprobleme—Advances in Solid State Physics*, edited by U. Rössler (Vieweg, Braunschweig), Vol. 30, 1.
- Graham, R., and A. Schenzle, 1982, *Phys. Rev. A* **25**, 1731.
- Grebogi, C., E. Ott, and J. A. Yorke, 1987, *Physica D* **7**, 181.
- Grifoni, M., and P. Hänggi, 1996a, *Phys. Rev. Lett.* **76**, 1611.
- Grifoni, M., and P. Hänggi, 1996b, *Phys. Rev. E* **54**, 1390.
- Grifoni, M., L. Hartmann, S. Berchtold, and P. Hänggi, 1996, *Phys. Rev. E* **53**, 5890; **56**, 6213 (E).
- Grifoni, M., M. Sassetti, P. Hänggi, and U. Weiss, 1995, *Phys. Rev. E* **52**, 3596.
- Grifoni, M., M. Sassetti, J. Stockburger, and U. Weiss, 1993, *Phys. Rev. E* **48**, 3497.
- Grigolini, P., and F. Marchesoni, 1985, *Adv. Chem. Phys.* **62**, 29.
- Grigorenko, A. N., V. I. Konov, and P. I. Nikitin, 1990, *JETP Lett.* **52**, 592.
- Grigorenko, A. N., and P. I. Nikitin, 1995, *IEEE Trans. Magn.* **31**, 2491.
- Grigorenko, A. N., P. I. Nikitin, A. N. Slavin, and P. Y. Zhou, 1994, *J. Appl. Phys.* **76**, 6335.
- Grohs, J., S. Apanasevich, P. Jung, H. Issler, D. Burak, and C. Klingshirn, 1994, *Phys. Rev. A* **49**, 2199.
- Grohs, J., H. Issler, and C. Klingshirn, 1991, *Opt. Commun.* **86**, 183.
- Grohs, J., A. Schmidt, M. Kunz, C. Weber, A. Daunois, B. Schehr, A. Rupp, W. Dotter, F. Werner, and C. Klingshirn, 1989, *Proc. SPIE* **1127**, 39.
- Grossmann, F., T. Dittrich, P. Jung, and P. Hänggi, 1991, *Phys. Rev. Lett.* **67**, 516.
- Haken, H. 1970, *Laser Theory* (Springer, Berlin).
- Hänggi, P., 1978, *Helv. Phys. Acta* **51**, 202.
- Hänggi, P., 1993, in *Activated Barrier Crossing*, edited by G. R. Fleming and P. Hänggi (World Scientific, London), pp. 268–292.
- Hänggi, P., and R. Bartussek, 1996, in *Nonlinear Physics of Complex Systems: Current Status and Future Trends*, edited by J. Parisi, S. C. Müller, and W. Zimmermann, Lecture Notes in Physics 476 (Springer, Berlin, New York), p. 294.
- Hänggi, P., H. Grabert, G. L. Ingold, and U. Weiss, 1985, *Phys. Rev. Lett.* **55**, 761.
- Hänggi, P., and P. Jung, 1995, *Adv. Chem. Phys.* **89**, 239.
- Hänggi, P., P. Jung, and F. Marchesoni, 1989, *J. Stat. Phys.* **54**, 1367.
- Hänggi, P., P. Jung, C. Zerbe, and F. Moss, 1993, *J. Stat. Phys.* **70**, 25.
- Hänggi, P., F. Marchesoni, and P. Grigolini, 1984, *Z. Phys. B* **56**, 333.
- Hänggi, P., F. Marchesoni, and P. Sodano, 1988, *Phys. Rev. Lett.* **60**, 2563.
- Hänggi, P., P. Talkner, and M. Borkovec, 1990, *Rev. Mod. Phys.* **62**, 251.
- Hänggi, P., and H. Thomas, 1982, *Phys. Rep.* **88**, 207.
- Heneghan, C., C. C. Chow, J. J. Collins, T. T. Imhoff, S. B. Lowen, and M. C. Teich, 1996, *Phys. Rev. E* **54**, R2228.
- Hibbs, A. D., A. L. Singsaas, E. W. Jacobs, A. R. Bulsara, J. J. Bekkedahl, and F. Moss, 1995, *J. Appl. Phys.* **77**, 2582.
- Hohmann, W., J. Müller, and F. W. Schneider, 1996, *J. Chem. Phys.* **100**, 5388.
- Hu, G., 1993, *Phys. Lett. A* **174**, 247.
- Hu, G., A. Daffertshofer, and H. Haken, 1996, *Phys. Rev. Lett.* **76**, 4874.
- Hu, G., T. Ditzinger, C. Z. Ning, and H. Haken, 1993, *Phys. Rev. Lett.* **71**, 807.
- Hu, G., H. Haken, and C. Z. Ning, 1992, *Phys. Lett. A* **172**, 21.
- Hu, G., H. Haken, and C. Z. Ning, 1993, *Phys. Rev. E* **47**, 2321.
- Hu, G., H. Haken, and F. Xie, 1996, *Phys. Rev. Lett.* **77**, 1925.
- Hu, G., G. Nicolis, and C. Nicolis, 1990, *Phys. Rev. A* **42**, 2030.
- Hu, G., G. R. Qing, D. C. Gong, and X. D. Weng, 1991, *Phys. Rev. A* **44**, 6424.
- I, L., and J.-M. Liu, 1995, *Phys. Rev. Lett.* **74**, 3161.
- Iannelli, J. M., A. Yariv, T. R. Chen, and Y. H. Zhuang, 1994, *Appl. Phys. Lett.* **65**, 1983.
- Imbrie, J., A. C. Mix, and D. G. Martinson, 1993, *Nature (London)* **363**, 531.
- Inchiosa, M., and A. Bulsara, 1995a, *Phys. Rev. E* **52**, 327.
- Inchiosa, M., and A. Bulsara, 1995b, *Phys. Rev. E* **53**, 327.
- Inchiosa, M., and A. Bulsara, 1995c, *Phys. Lett. A* **200**, 283.
- Inchiosa, M., and A. Bulsara, 1996, *Phys. Rev. E* **51**, 2021.
- Ippen, E., J. Lindner, and W. Ditto, 1993, *J. Stat. Phys.* **70**, 148.
- Jost, B., and B. Sahleh, 1996, *Opt. Lett.* **21**, 287.
- Jülicher, F., A. Ajdari, and J. Prost, 1997, *Rev. Mod. Phys.* **69**, 1269.
- Jung, P., 1989, *Z. Phys. B* **76**, 521.
- Jung, P., 1993, *Phys. Rep.* **234**, 175.
- Jung, P., 1994, *Phys. Rev. E* **50**, 2513.
- Jung, P., 1995, *Phys. Lett. A* **207**, 93.
- Jung, P., 1997, in *Stochastic Dynamics*, edited by L. Schimansky-Geier and T. Pöschel, Lecture Notes in Physics No. 484 (Springer, Berlin), p. 23.
- Jung, P., and R. Bartussek, 1996, in *Fluctuations and Order: The New Synthesis*, edited by M. Millonas (Springer, New York, Berlin), pp. 35–52.

- Jung, P., U. Behn, E. Pantazelou, and F. Moss, 1992, *Phys. Rev. A* **46**, R1709.
- Jung, P., G. Gray, R. Roy, and P. Mandel, 1990, *Phys. Rev. Lett.* **65**, 1873.
- Jung, P., and P. Hänggi, 1989, *Europhys. Lett.* **8**, 505.
- Jung, P., and P. Hänggi, 1990, *Phys. Rev. A* **41**, 2977.
- Jung, P., and P. Hänggi, 1991a, *Phys. Rev. A* **44**, 8032.
- Jung, P., and P. Hänggi, 1991b, *Ber. Bunsenges. Phys. Chem.* **95**, 311.
- Jung, P., and P. Hänggi, 1993, *Z. Phys. B* **90**, 255.
- Jung, P., J. G. Kissner, and P. Hänggi, 1996, *Phys. Rev. Lett.* **76**, 3436.
- Jung, P., and G. Mayer-Kress, 1995, *Phys. Rev. Lett.* **74**, 2130.
- Jung, P., and P. Talkner, 1995, *Phys. Rev. E* **51**, 2640.
- Jung, P., and K. Wiesenfeld, 1997, *Nature (London)* **385**, 291.
- Kapitaniak, T., 1994, *Phys. Rev. E* **49**, 5855.
- Kim, S., and L. E. Reichl, 1996, *Phys. Rev. E* **53**, 3088.
- Kiss, L. B., Z. Gingl, Z. Marton, J. Kertesz, F. Moss, G. Schnera, and A. Bulsara, 1993, *J. Stat. Phys.* **70**, 451.
- Kittel, A., R. Richter, M. Hirsch, G. Flätgen, J. Peinke, and J. Parisi, 1993, *Z. Naturforsch. Teil A* **48**, 633.
- Kramers, H., 1940, *Physica (Utrecht)* **7**, 284.
- Krogh, A., and R. Palmer, 1991, *Introduction to the Theory of Neural Computation* (Addison Wesley, New York), 1991.
- Kubo, R., 1957, *J. Phys. Soc. Jpn.* **12**, 570.
- Kubo, R., 1966, *Rep. Prog. Phys.* **29**, 255.
- Kubo, R., M. Toda, and N. Hashitsume, 1985, *Statistical Physics II*, Springer Series in Solid State Sciences Vol. 31 (Springer, Berlin, New York).
- Lambsdorff, M., C. Dornfeld, and C. Klingshirn, 1986, *Z. Phys. B* **64**, 409.
- Leggett, A. J., S. Chakravarty, A. T. Dorsey, M. P. A. Fisher, A. Garg, and W. Zwerger, 1987, *Rev. Mod. Phys.* **59**, 1; **67**, 725(E).
- Leibler, S., 1994, *Nature (London)* **370**, 412.
- Levin, J. E., and J. P. Miller, 1996, *Nature (London)* **380**, 165.
- Lindner, J. F., B. K. Meadows, W. L. Ditto, M. E. Inchiosa, and A. Bulsara, 1995, *Phys. Rev. Lett.* **75**, 3.
- Löcher, M., G. A. Johnson, and E. R. Hunt, 1996, *Phys. Rev. Lett.* **77**, 4698.
- Löfstedt, R., and S. N. Coppersmith, 1994a, *Phys. Rev. Lett.* **72**, 1947.
- Löfstedt, R., and S. N. Coppersmith, 1994b, *Phys. Rev. E* **49**, 4821.
- Longtin, A., 1993, *J. Stat. Phys.* **70**, 309.
- Longtin, A., A. Bulsara, and F. Moss, 1991, *Phys. Rev. Lett.* **67**, 656.
- Longtin, A., A. R. Bulsara, D. Pierson, and F. Moss, 1994, *Biol. Cybern.* **70**, 569.
- Louis, A. A., and J. P. Sethna, 1995, *Phys. Rev. Lett.* **74**, 1363.
- Magnasco, M. O., 1993, *Phys. Rev. Lett.* **71**, 1477.
- Magnus, W., and S. Winkler, 1979, *Hill's Equation* (Dover, New York).
- Mahato, M. C., and S. R. Shenoy, 1994, *Phys. Rev. E* **50**, 2503.
- Makarov, D. E., and N. Makri, 1995, *Phys. Rev. E* **52**, 5863.
- Makri, N., 1997, *J. Chem. Phys.* **106**, 2286.
- Mandel, L., R. Roy, and S. Singh, 1981, in *Optical Bistability*, edited by C. M. Bowden, M. Cifan, and H. R. Robl (Plenum, New York), p. 127.
- Mannella, R., A. Fioretti, L. Fronzoni, B. Zambon, E. Arimondo, and S. Chillemi, 1995, *Phys. Lett. A* **197**, 25.
- Mansour, M., and G. Nicolis, 1975, *J. Stat. Phys.* **13**, 197.
- Mantegna, R. N., and B. Spagnolo, 1994, *Phys. Rev. E* **49**, R1792.
- Mantegna, R. N., and B. Spagnolo, 1995, *Nuovo Cimento D* **17**, 873.
- Mantegna, R. N., and B. Spagnolo, 1996, *Phys. Rev. Lett.* **76**, 563.
- Marchesoni, F., 1997, *Phys. Lett. A* **231**, 61.
- Marchesoni, F., L. Gammaitoni, and A. Bulsara, 1996, *Phys. Rev. Lett.* **76**, 2609.
- Marchesoni, F., and P. Grigolini, 1983, *Physica A* **121**, 269.
- Masoliver, J., A. Robinson, and G. H. Weiss, 1995, *Phys. Rev. E* **51**, 4021.
- Matteucci, G., 1989, *Climate Dynamics* **3**, 179.
- Matteucci, G., 1991, *Climate Dynamics* **6**, 67.
- McClintock, P. V. E., and F. Moss, 1989, in *Noise in Nonlinear Dynamical Systems*, Vol. 3, edited by F. Moss and P. V. E. McClintock, (Cambridge University, Cambridge), p. 243.
- McNamara, B., and K. Wiesenfeld, 1989, *Phys. Rev. A* **39**, 4854.
- McNamara, B., K. Wiesenfeld, and R. Roy, 1988, *Phys. Rev. Lett.* **60**, 2626.
- Melnikov, V. I., 1993, *Phys. Rev. E* **48**, 2481.
- Millman, J., 1983, *Microelectronics*, McGraw-Hill Series in Electrical Engineering, Electronics and Electronic Circuits (McGraw-Hill, New York).
- Morillo, M., J. Gómez-Ordóñez, and J. M. Casado, 1995, *Phys. Rev. E* **52**, 316.
- Moss, F., 1991, *Ber. Bunsenges. Phys. Chem.* **95**, 303.
- Moss, F., 1994, in *Contemporary Problems in Statistical Physics*, edited by G. H. Weiss (SIAM, Philadelphia), pp. 205–253.
- Moss, F., A. Bulsara, and M. F. Shlesinger, 1993, Eds., *The Proceedings of the NATO Advanced Research Workshop: Stochastic Resonance in Physics and Biology*, *J. Stat. Phys.* **70** (Plenum, New York), pp. 1–512.
- Moss, F., J. K. Douglass, L. Wilkins, D. Pierson, and E. Pantazelou, 1993, *Ann. (N.Y.) Acad. Sci.* **706**, 26.
- Moss, F., D. Pierson, and D. O'Gorman, 1994, *Int. J. Bifurcation Chaos Appl. Sci. Eng.* **4**, 1383.
- Moss, F., and K. Wiesenfeld, 1995a, *Sci. Am.* **273**, 50.
- Moss, F., and K. Wiesenfeld, 1995b, *Spektrum der Wissenschaft*, **273** (Oktober), 92.
- Murray, J. D., 1989, *Mathematical Biology* (Springer, Berlin).
- Néda, Z., 1995a, *Phys. Lett. A* **210**, 125.
- Néda, Z., 1995b, *Phys. Rev. E* **51**, 5315.
- Neiman, A., and L. Schimansky-Geier, 1994, *Phys. Rev. Lett.* **72**, 2988.
- Neiman, A., and L. Schimansky-Geier, 1995, *Phys. Lett. A* **197**, 379.
- Neiman, A., L. Schimansky-Geier, and F. Moss, 1997, *Phys. Rev. E* **56**, R9.
- Neimann, A., B. Shulgin, V. Anishchenko, W. Ebeling, L. Schimansky-Geier, and J. A. Freund, 1996, *Phys. Rev. Lett.* **76**, 4299.
- Neiman, A., and W. Sung, 1996, *Phys. Lett. A* **223**, 341.
- Nicolis, C., 1981, *Sol. Phys.* **74**, 473.
- Nicolis, C., 1982, *Tellus* **34**, 1.
- Nicolis, C., 1993, *J. Stat. Phys.* **70**, 3.
- Nicolis, C., and G. Nicolis, 1981, *Tellus* **33**, 225.
- Nicolis, C., G. Nicolis, and G. Hu, 1990, *Phys. Lett. A* **151**, 139.
- Nicolis, G., C. Nicolis, and D. McKernan, 1993, *J. Stat. Phys.* **70**, 125.
- Oelschlägel, B., T. Dittrich, and P. Hänggi, 1993, *Acta Phys. Pol. B* **24**, 845.

- Oppenheim, A. V., and R. W. Schaffer, 1975, *Digital Signal Processing* (Prentice Hall, New York).
- Papoulis, A., 1965, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, New York).
- Pei, X., J. Wilkens, and F. Moss, 1996, *J. Neurophysiol.* **76**, 3002.
- Pérez-Madrid, A., and J. M. Rubí, 1995, *Phys. Rev. E* **51**, 4159.
- Phillips, J. C., and K. Schulten, 1995, *Phys. Rev. E* **52**, 2473.
- Presilla, C., F. Marchesoni, and L. Gammaitoni, 1989, *Phys. Rev. A* **40**, 2105.
- Raikher, Y. L., and V. I. Stepanov, 1994, *J. Phys.: Condens. Matter* **6**, 4137.
- Rajaraman, R., 1982, *Solitons and Instantons* (North-Holland, Amsterdam).
- Rappel, W. J., and S. H. Strogatz, 1994, *Phys. Rev. E* **50**, 3249.
- Reibold, E., W. Just, J. Becker, and H. Benner, 1997, *Phys. Rev. Lett.* **78**, 3101.
- Reimann, P., R. Bartussek, W. Häussler, and P. Hänggi, 1996, *Phys. Lett. A* **215**, 26.
- Riani, M., and E. Simonotto, 1994, *Phys. Rev. Lett.* **72**, 3120.
- Riani, M., and E. Simonotto, 1995, *Nuovo Cimento D* **17**, 903.
- Rice, S. O., 1944, *Bell Syst. Tech. J.* **23**, 1; reprinted in N. Vax, 1954, *Noise and Stochastic Processes* (Dover, New York).
- Rice, S. O., 1948, *Bell Syst. Tech. J.* **27**, 109.
- Risken, H., 1984, *The Fokker-Planck Equation*, Springer Series in Synergetics Vol. 18 (Springer, Berlin, New York).
- Risken, H., and H. D. Vollmer, 1989, in *Noise in Nonlinear Dynamical Systems; Theory, Experiments, Simulation*, Vol. I, edited by F. Moss and P. V. E. McClintock (Cambridge University, Cambridge), p. 191.
- Rose, J. E., J. F. Brugge, D. J. Anderson, and J. E. Hind, 1967, *J. Neurophysiol.* **30**, 769.
- Rouse, R., S. Han, and J. E. Lukens, 1995, *Appl. Phys. Lett.* **66**, 108.
- Roy, R., P. A. Schulz, and A. Walther, 1987, *Opt. Lett.* **12**, 672.
- Sargent III, M., M. O. Scully, and W. E. Lamb Jr., 1974, *Laser Physics* (Addison-Wesley, Reading, MA).
- Schenzle, A., and H. R. Brand, 1979, *Phys. Rev. A* **20**, 1628.
- Schimansky-Geier, L., J. A. Freund, A. B. Neiman, and B. Shulgin, 1998, "Noise Induced Order: Stochastic Resonance," *Int. J. Bifurcation Chaos Appl. Sci. Eng.* (in press).
- Schimansky-Geier, L., and U. Siewert, 1997, in *Stochastic Dynamics*, edited by L. Schimansky-Geier, and T. Pöschel, *Lecture Notes in Physics* No. 484 (Springer, Berlin), p. 245.
- Schimansky-Geier, L., and Ch. Zülicke, 1990, *Z. Phys. B* **79**, 451.
- Schwartz, D. B., B. Sen, C. N. Archie, and J. E. Lukens, 1985, *Phys. Rev. Lett.* **55**, 1547.
- Shiino, M., 1987, *Phys. Rev. A* **36**, 2393.
- Shneidman, V. A., P. Jung, and P. Hänggi, 1994a, *Phys. Rev. Lett.* **72**, 2682.
- Shneidman, V. A., P. Jung, and P. Hänggi, 1994b, *Europhys. Lett.* **26**, 571.
- Shulgin, B., A. Neiman, and V. Anishchenko, 1995, *Phys. Rev. Lett.* **75**, 4157.
- Simon, A., and A. Libchaber, 1992, *Phys. Rev. Lett.* **68**, 3375.
- Simonotto, E., M. Riani, C. Seife, M. Roberts, J. Twitty, and F. Moss, 1997, *Phys. Rev. Lett.* **78**, 1186.
- Smoluchowski, M. v., 1912, *Phys. Z.* **8**, 1069; see p. 1078.
- Smoluchowski, M. v., 1914, in *Vorträge über die kinetische Theorie der Materie und der Elektrizität*, edited by M. Planck *et al.* (Teubner, Leipzig), pp. 89–121.
- Spano, M. L., M. Wun-Fogle, and W. L. Ditto, 1992, *Phys. Rev. A* **46**, R5253.
- Stocks, N. G., 1995, *Nuovo Cimento D* **17**, 925.
- Stratonovich, R. L., 1963, *Topics in the Theory of Random Noise*, Vol. I (Gordon and Breach, New York), p. 143ff.
- Teich, M. C., S. M. Khanna, and P. C. Guiney, 1993, *J. Stat. Phys.* **70**, 257.
- Thorwart, M., and P. Jung, 1997, *Phys. Rev. Lett.* **78**, 2503.
- Valls, O. T., and G. F. Mazenko, 1986, *Phys. Rev. B* **34**, 7941.
- Van den Broeck, C., J. M. R. Parrando, J. Armero, and A. Hernandez-Machado, 1994, *Phys. Rev. E* **49**, 2639.
- van Kampen, N. G., 1992, *Stochastic Processes in Physics and Chemistry*, 2nd ed. (North Holland, Amsterdam, New York).
- Vemuri, G., and R. Roy, 1989, *Phys. Rev. A* **39**, 4668.
- Vilar, J. M. G., and J. M. Rubi, 1997, *Phys. Rev. Lett.* **78**, 2886.
- Weiss, U., 1993, *Quantum Dissipative Systems*, Series in Modern Condensed Matter Physics (World Scientific, Singapore), Vol. 2.
- Wiesenfeld, K., and F. Moss, 1995, *Nature (London)* **373**, 33.
- Wiesenfeld, K., D. Pierson, E. Pantazelou, C. Dames, and F. Moss, 1994, *Phys. Rev. Lett.* **72**, 2125.
- Winograd, I. J., T. B. Coplen, J. M. Landwehr, A. C. Riggs, K. R. Ludwig, B. J. Szabo, P. T. Kolesar, and K. M. Revesz, 1992, *Science* **258**, 255.
- Wio, H. S., 1996, *Phys. Rev. E* **54**, R3075.
- Yang, W., M. Ding, and G. Hu, 1995, *Phys. Rev. Lett.* **74**, 3955.
- Zambon, B., F. De Tomasi, D. Hennequin, and E. Arimondo, 1989, *Phys. Rev. A* **40**, 3782.
- Zhou, T., and F. Moss, 1990, *Phys. Rev. A* **41**, 4255.
- Zhou, T., F. Moss, and P. Jung, 1990, *Phys. Rev. A* **42**, 3161.

nature insight

Complex systems



Cover illustration

A computer simulation of the ground displacement due to the 1992 Landers earthquake — an example of one of the many systems that show complex dynamical behaviour. (Image: D. Massonnet/CNES/SPL.)

The science of complexity, as befits its name, lacks a simple definition. It has been used to refer to the study of systems that operate at the 'edge of chaos' (itself a loosely defined concept); to infer structure in the complex properties of systems that are intermediate between perfect order and perfect disorder; or even as a simple restatement of the cliché that the behaviour of some systems as a whole can be more than the sum of their parts.

Notwithstanding these difficulties over formal definition, the study of complex systems has seen tremendous growth. Numerous research programmes, institutes and scientific journals have been established under this banner. And the new concepts emerging from these studies are now influencing disciplines as disparate as astronomy and biology, physics and finance. The richness of the field and the diversity of its application lends itself naturally to the Insight format, although our choice of themes to review is necessarily somewhat eclectic.

We begin by considering systems in which the microscopic properties and processes can be immensely complex and seemingly noisy, yet on larger scales they exhibit certain classes of simple behaviour that seem insensitive to the mechanistic details. On page 242, Sethna *et al.* show that the seemingly random, impulsive events by which many physical systems evolve exhibit universal — and, to some extent, predictable — behaviour. Shinbrot and Muzzio on page 251 offer a different perspective on noise, describing how order and patterns can emerge from intrinsically noisy systems.

We then shift our focus to systems where both the properties of the individual components and the nature of their interactions are reasonably well understood, yet the collective behaviour of the ensemble can still defy simple explanation. On page 259, Debenedetti and Stillinger show how recent theoretical progress on describing the dynamics of systems of many identical interacting particles — in the form of a multidimensional 'energy landscape' — is shedding light on the age-old phenomena of supercooling and glass formation. And for extensive networks of simple interacting systems, Strogatz shows on page 268 how network topology can be as important as the interactions between elements.

But complex systems do not always lend themselves to such easy (if qualitative) categorization. For example, the many complex rhythms encountered in living organisms arise not just from intrinsic stochastic or nonlinear dynamical processes, but also from their interaction with an external fluctuating environment. Yet, according to Glass on page 277, decoding the essential features of these rhythms might ultimately be of benefit to medicine, even in the absence of a simple mathematical interpretation.

As should be clear from these articles, the science of complexity is in its infancy, and some research directions that today seem fruitful might eventually prove to be academic cul-de-sacs. Nevertheless, it seems reasonable to suppose that the general principles emerging from these studies will help us to better understand the complex world around us.

Karl Ziemelis Physical Sciences Editor

Liz Allen Publisher

Crackling noise

James P. Sethna*, Karin A. Dahmen† & Christopher R. Myers‡

*Laboratory of Atomic and Solid State Physics, Clark Hall, Cornell University, Ithaca, New York 14853-2501, USA (sethna@lassp.cornell.edu)

†Department of Physics, 1110 West Green Street, University of Illinois at Urbana-Champaign, Illinois 61801-3080, USA

(dahmen@physics.uiuc.edu)

‡Cornell Theory Center, Frank H. T. Rhodes Hall, Cornell University, Ithaca, New York 14853-3801, USA (myers@tc.cornell.edu)

Crackling noise arises when a system responds to changing external conditions through discrete, impulsive events spanning a broad range of sizes. A wide variety of physical systems exhibiting crackling noise have been studied, from earthquakes on faults to paper crumpling. Because these systems exhibit regular behaviour over a huge range of sizes, their behaviour is likely to be independent of microscopic and macroscopic details, and progress can be made by the use of simple models. The fact that these models and real systems can share the same behaviour on many scales is called universality. We illustrate these ideas by using results for our model of crackling noise in magnets, explaining the use of the renormalization group and scaling collapses, and we highlight some continuing challenges in this still-evolving field.

In the past decade or so, science has broadened its purview to include a new range of phenomena. Using tools developed to understand second-order phase transitions^{1–5} in the 1960s and 70s, stochastic models of turbulence⁶ in the 1970s, and disordered systems^{7–9} in the 1980s, scientists now claim that they should be able to explain how and why things crackle.

Many systems crackle; when pushed slowly, they respond with discrete events of a variety of sizes. The Earth responds¹⁰ with violent and intermittent earthquakes as two tectonic plates rub past one another (Fig. 1). A piece of paper¹¹ (or a candy wrapper at the cinema^{12,13}) emits intermittent, sharp noises as it is slowly crumpled or rumpled. (Try it, but preferably not with this page.) A magnetic material in a changing external field magnetizes in a series of jumps^{14,15}. These individual events span many orders of magnitude in size — indeed, the distribution of sizes forms a power law with no characteristic size scale. In the past few years, scientists have been making rapid progress in developing models and theories for understanding this sort of scale-invariant behaviour in driven, nonlinear, dynamical systems.

Interest in these sorts of phenomena goes back several decades. The work of Gutenberg and Richter¹⁰ in the 1940s and 1950s established the well-known frequency-magnitude relationship for earthquakes that bears their names (Fig. 1). A variety of many-degree-of-freedom dynamical models^{16–28}, with and without disorder, have been introduced in the years since to investigate the nature of slip complexity in earthquakes. More recent impetus for work in this field came from the study of the depinning transition in sliding charge-density wave (CDW) conductors in the 1980s and early 1990s^{29–35}. Interpretation of the CDW depinning transition as a dynamic critical phenomenon sprung from Fisher's early work^{29,30}, and several theoretical and numerical studies followed. This activity culminated in the renormalization-group solution by Narayan and Fisher³² and the numerical studies by Middleton³³ and Myers³⁴, which combined to provide a clear picture of depinning in CDWs and open the doors to the study of other disordered, non-equilibrium systems.

Bak, Tang and Wiesenfeld inspired much of the succeeding work on crackling noise^{36,37}. They introduced the connection between dynamical critical phenomena and crackling noise, and they emphasized how systems may end

up naturally at the critical point through a process of self-organized criticality. (Their original model was that of avalanches in growing sandpiles — sand has long been used as an example of crackling noise^{38,39}, but we now know that real sandpiles do not crackle at the longest scales^{40,41}.)

Researchers have studied many systems that crackle. Simple models have been developed to study bubbles rearranging in foams as they are sheared⁴², biological extinctions⁴³ (where the models are controversial^{44,45} — they ignore catastrophic external events like asteroids), fluids invading porous materials and other problems involving invading fronts^{46–51} (where the model we describe was invented^{46,47}), the dynamics of superconductors^{52–54} and superfluids^{55,56}, sound emitted during martensitic phase transitions⁵⁷, fluctuations in the stock market^{58,59}, solar flares⁶⁰, cascading failures in power grids^{61,62}, failures in systems designed for optimal performance^{63–65}, group decision-making⁶⁶, and fracture in disordered materials^{67–72}. These models are driven systems with many degrees of freedom, which respond to the driving in a series of discrete avalanches spanning a broad range of scales — what in this paper we term crackling noise.

There has been healthy scepticism by some established professionals in these fields to the sometimes-grandiose claims by newcomers claiming support for an overarching paradigm. But often confusion arises because of the unusual kind of predictions the new methods provide. If such models apply at all to a physical system, they should be able to predict most behaviour on long scales of length and time, independent of many microscopic details of the real world. But this predictive capacity comes at a price: the models typically do not make clear predictions of how the real-world microscopic parameters affect the behaviour at long length scales.

Here we provide an overview of the renormalization group^{1–5} used by many researchers to understand crackling noise. Briefly, the renormalization group discusses how the effective evolution laws of a system change as measurements are made on longer and longer length scales. (It works by generating a coarse-graining mapping in system space, the abstract space of possible evolution laws.) The broad range of event sizes are attributed to a self-similarity, where the evolution laws look the same at different length scales. This self-similarity leads to a method for scaling experimental data. In the simplest case this yields power laws and fractal

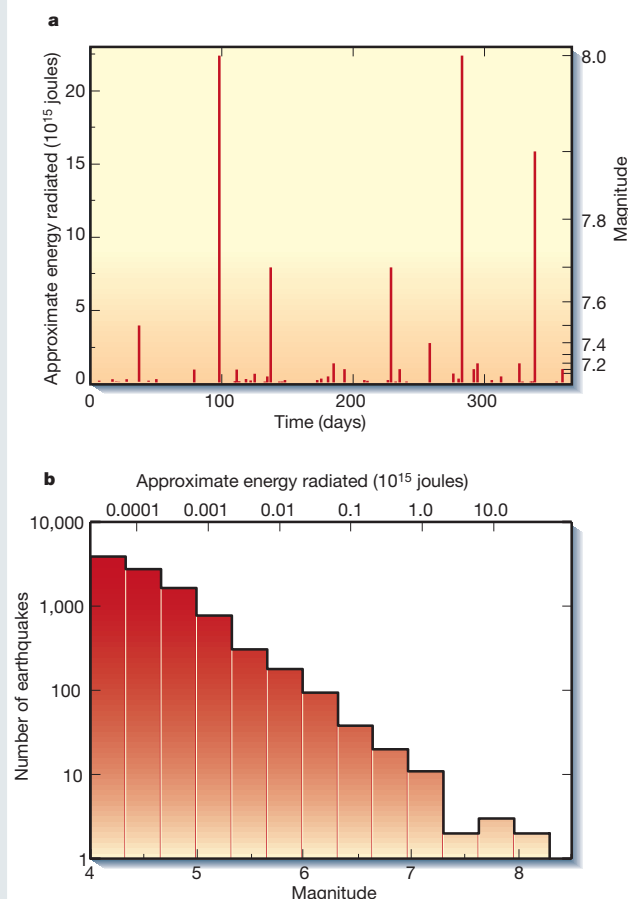


Figure 1 The Earth cracks. **a**, Time history of radiated energy from earthquakes throughout all of 1995^{108–110}. The Earth responds to the slow strains imposed by continental drift through a series of earthquakes (impulsive events well separated in space and time). This time series, when sped up, sounds remarkably like the crackling noise of paper, magnets and Rice Krispies (listen to it in ref. 110). **b**, Histogram of number of earthquakes in 1995 as function of their magnitude (or, alternatively, their energy release). Earthquakes come in a wide range of sizes, from unnoticeable trembles to catastrophic events. The smaller earthquakes are much more common: the number of events of a given size forms a power law¹⁰ called the Gutenberg–Richter law. (Earthquake magnitude scales with the logarithm of the strength of the earthquake. On a log–log plot of number versus radiated energy, the power law is a straight line, as we observe in the plotted histogram.) One would hope that such a simple law should have an elegant explanation.

structures, but more generally it leads to universal scaling functions — where we argue the real predictive power lies. We will only touch upon the complex analytical methods used in this field, but we believe we can explain faithfully and fully both what our tools are useful for, and how to apply them in practice. The renormalization group is perhaps the most impressive use of abstraction in science.

Why should crackling noise be comprehensible?

Not all systems crackle. Some respond to external forces with many similar-sized, small events (for example, popcorn popping as it is heated). Others give way in one single event (for example, chalk snapping as it is stressed). In broad terms, crackling noise is in between these limits: when the connections between parts of the system are stronger than in popcorn but weaker than in the grains making up chalk, the yielding events can span many size scales. Crackling forms the transition between snapping and popping.

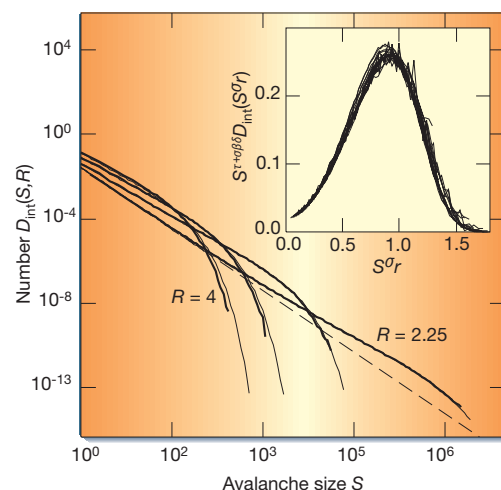


Figure 2 Magnets crackle^{73–76}. Magnets respond to a slowly varying external field by changing their magnetization in a series of bursts, or avalanches. These bursts, called Barkhausen noise, are very similar (albeit on different time and size scales) to those shown in Fig. 1 for earthquakes. The avalanches in our model have a power-law distribution only at a special value of the disorder, $R_c = 2.16$. Shown is a histogram giving the number of avalanches $D_{\text{int}}(S, R)$ of a given size S at various disorders R ranging from 4 to 2.25; the thin lines are theoretical predictions from our model. (D_{int} gives all the avalanches during our simulation, integrated over the external field $-\infty < H(t) < +\infty$). The straight dashed line shows the power-law distribution at the critical point. Notice that fitting power laws to the data would work only very near to R_c : even with a range of a million in avalanche sizes, the slope has not converged to the asymptotic value. On the other hand, scaling-function predictions (theoretical curves) work well far from the critical point. The inset shows a scaling collapse of the avalanche size distribution (scaled probability versus scaled size), which is used to provide the theoretical curves as described in the text.

Figure 1b presents a simple relationship between earthquake number and magnitude. We expect that there ought to be a simple, underlying reason why earthquakes occur on all different sizes. The properties of very small earthquakes probably depend in detail on the kind of dirt (fault gouge) in the crack. The very largest earthquakes will depend on the geography of the continental plates. But the smooth power-law behaviour indicates that something simpler is happening in between, independent of either the microscopic or the macroscopic details.

There is an analogy here with the behaviour of a fluid. A fluid is very complicated on the microscopic scale, where molecules are bumping into one another: the trajectories of the molecules are chaotic, and depend both on exactly what direction they are moving and what they are made of. However, a simple law describes most fluids on long time and size scales. This law, the Navier–Stokes equation, depends on the constituent molecules only through a few parameters (the density and viscosity). Physics works because simple laws emerge on large scales. In fluids, these microscopic fluctuations and complexities disappear on large scales: for crackling noise, they become scale-invariant and self-similar.

How do we derive the laws for crackling noise? There are two approaches. First, we can calculate analytically the behaviour on long time and size scales by formally coarse-graining over the microscopic fluctuations. This leads us to renormalization-group methods^{1–5}, which we discuss in the next section. The analytic approach can be challenging, but it can give useful results and (more important) is the only explanation for why events on all scales should occur. Second, we can make use of universality. If the microscopic details do not matter for the behaviour at long length scales, why not make up a simple model with the same behaviour (in the same universality class) and solve it?

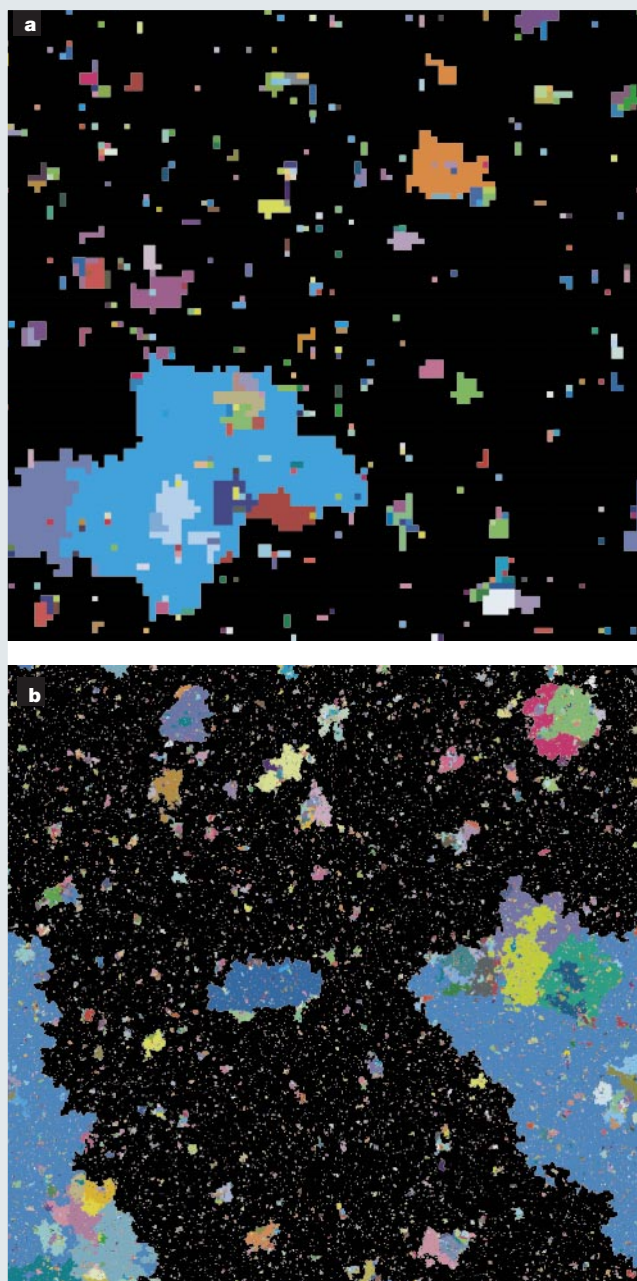


Figure 3 Self-similarity. Cross-sections of the avalanches during the magnetization of our model^{15,73–76}. Here each avalanche is drawn in a separate colour. **a**, A 100^3 simulation; **b**, a $1,000^3$ simulation (a billion domains⁷⁴); both are run at the critical point $R_c = 2.16$ J where avalanches just barely continue. The black background represents a large avalanche that spans the system: the critical point occurs when avalanches would first span an infinite system.

The model we focus on here is a caricature of a magnetic material^{15,46,47,73–77}. A piece of iron will ‘crackle’ as it enters a strong magnetic field, giving what is called Barkhausen noise. We model the iron as a cubic grid of magnetic domains S_i , whose north pole is either pointing upwards ($S_i = +1$) or downwards ($S_i = -1$). The external field pushes on our domain with a force $H(t)$, which will increase with time. Iron can be magnetized because neighbouring domains prefer to point in the same direction: if the six neighbours of our cubic domain are S_j , then in our model we let their force on our domain be $\sum_j JS_j$ (where we set the coupling $J=1$). Finally, we model dirt, randomness in the domain shapes, and other kinds of disorder by introducing a random field h_i , different for each domain and chosen

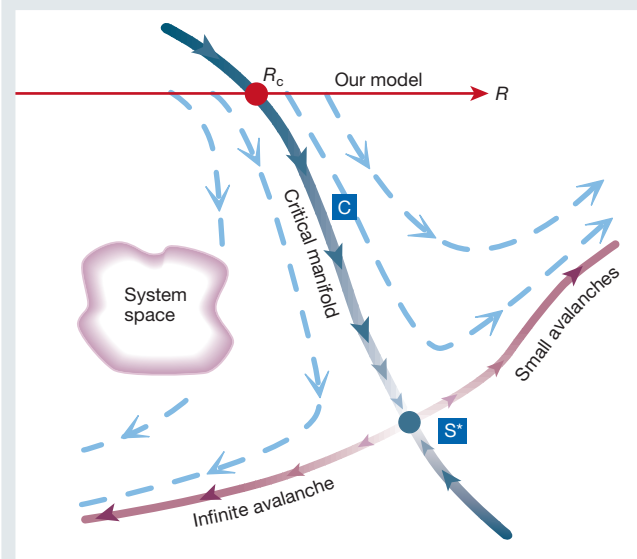


Figure 4 Renormalization-group flows. The renormalization group is a theory of how coarse-graining to longer length scales introduces a mapping from the space of physical systems to itself. Consider the space of all possible models of magnetic hysteresis. Each model can be coarse-grained, removing some fraction of the microscopic degrees of freedom and introducing more complicated rules so that the remaining ones still flip at the same external fields. This defines a mapping from our space into itself. A fixed point S^* in this space will be self-similar: because it maps to itself upon coarse-graining, it must have the same behaviour on different length scales. Points that flow into S^* under coarse-graining share this self-similar behaviour on sufficiently long length scales: they all share the same universality class.

at random from a normal distribution with standard deviation R , which we call the disorder. The net force on our domain is thus

$$\text{Force on domain } i = H(t) + \sum_j JS_j + h_i \quad (1)$$

The domains in our model all start with their north pole pointing down (-1), and flip up as soon as the net force on them becomes positive. This can occur either because $H(t)$ increases sufficiently (spawning a new avalanche), or because one of their neighbours flipped up, kicking them over (propagating an existing avalanche). (Thermal fluctuations are ignored: a good approximation in many experiments because the domains are large.) If the disorder R is large, so the h_i are typically big compared to J , then most domains flip independently: all the avalanches are small, and we get popping noise. If the disorder is small compared to J , then typically most of the domains will be triggered by one of their neighbours: one large avalanche will snap up most of our system. In between, we get crackling noise. When the disorder R is just large enough so that each domain flip on average triggers one of its neighbours (at the critical disorder R_c), then we find avalanches on all scales (Figs 2, 3).

What do these avalanches represent? In nonlinear systems with many degrees of freedom, there are often large numbers of metastable states. Local regions in the system can have multiple stable configurations, and many combinations of these local configurations are possible. (A state is metastable when it cannot lower its energy by small rearrangements. It is distinguished from the globally stable state, which is the absolute lowest energy possible for the system.) Avalanches are the rearrangements that occur as our system shifts from one metastable state to another. Our specific interest is in systems with a broad distribution of avalanche sizes, where shifting between metastable states can rearrange anything between a few domains and millions of domains.

There are many choices we made in our model that do not matter at long time and size scales. Because of universality, we can argue^{78,79}

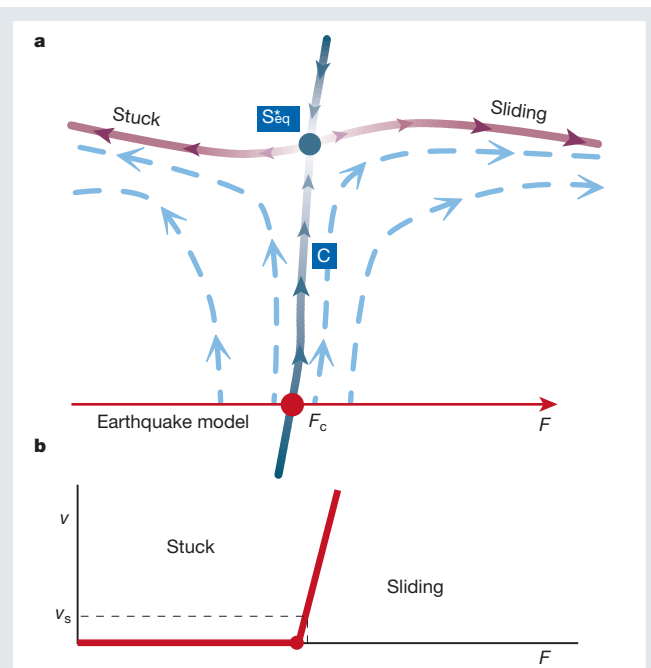


Figure 5 Flows in the space of earthquake models. A model for earthquakes will have a force F applied across the front. In models ignoring inertia and velocity-dependent friction²⁸, there is a critical force F_c that just allows the fault to slip forward. **a**, Coarse-graining defines a flow on the space of earthquake models. The fixed point S^* will have a different local flow field from other renormalization-group fixed points, yielding its own universality class of critical exponents and scaling functions. The critical manifold C , consisting of models that flow into S^* , separates the stuck faults from those that slide forward with an average velocity $v(F)$. **b**, The velocity varies with the external force as a power law $v(F) \sim F^\beta$. The motion of the continental plates, however, does not fix the force F across the fault: rather, it sets the average relative velocity to a small value v_s (centimetres per year). This automatically sets the force across the fault very close to its critical force F_c . This is one example of self-organized criticality^{36,37}.

that the behaviour would be the same if we chose a different grid of domains, or if we changed the distribution of random fields, or if we introduced more realistic random anisotropies and random coupling constants. Were this not the case, we could hardly expect our simple model to explain real experiments.

The renormalization group and scaling

To study crackling noise, we use renormalization-group^{1–5,78,80} tools developed in the study of second-order phase transitions. The word renormalization has roots in the study of quantum electrodynamics, where the effective charge changes in size (norm) as a function of length scale. The word group refers to the family of coarse-graining operations that are basic to the method: the group product is composition (coarsening repeatedly). The name is unfortunate, however, as the basic coarse-graining operation does not have an inverse, so that the renormalization group does not have the mathematical structure of a group.

The renormalization group studies the way the space of all physical systems maps into itself under coarse-graining (Fig. 4). The coarse-graining operation shrinks the system and removes degrees of freedom on short length scales. Under coarse-graining, we often find a fixed point S^* : many different models flow into the fixed point and hence share long-wavelength properties. Figure 3 provides a schematic view of coarse-graining: the $1,000^3$ cross-section looks (statistically) like the 100^3 section if you blur your eyes by a factor of ten. Much of the mathematical complexity of this field involves finding analytical tools for computing the flow diagram in Fig. 4. Using methods developed to study thermodynamical phase

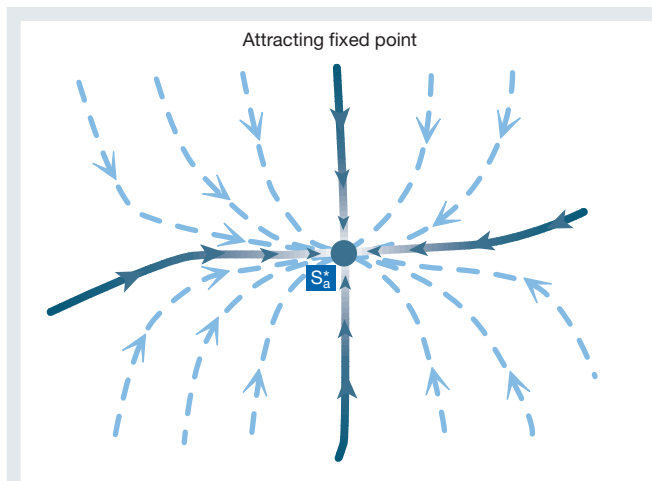


Figure 6 Attracting fixed point. Often there will be fixed points that attract in all directions. These fixed points describe phases rather than phase transitions. Most phases are rather simple, with fluctuations that die away on long length scales⁸¹. When fluctuations remain important, they will exhibit self-similarity and power laws called generic scale invariance^{83,84}.

transitions² and the depinning of charge-density waves³², we can calculate for our model the flows for systems in dimensions close to six (the so-called ϵ -expansion^{78–80}, where $\epsilon = 6 - d$, with d being the dimension of the system). Interpolating between dimensions may seem a surprising thing to do. In our system it gives good predictions even in three dimensions (that is, $\epsilon = 3$), but it is difficult and is not discussed here. Nor will we discuss real-space renormalization-group methods¹ or series-expansion methods. We focus on the relatively simple task of using the renormalization group to justify and explain the universality, self-similarity and scaling observed in nature.

Consider the 'system space' for disordered magnets. There is a separate dimension in system space for each possible parameter in a theoretical model (disorder, coupling, next-neighbour coupling, dipolar fields, and so on) or in an experiment (for example, temperature, annealing time and chemical composition). Coarse-graining, however one implements it, gives a mapping from system space into itself: shrinking the system and ignoring the shortest length scales yields a new physical system with identical long-distance physics, but with different (renormalized) values of the parameters. We have abstracted the problem of understanding crackling noise in magnets into understanding a dynamical system acting on a space of dynamical systems.

Figure 4 represents a two-dimensional cross-section of this infinite-dimensional system space. We have chosen the cross-section to include our model (equation (1)): as we vary the disorder R , our model sweeps out a straight line (red) in system space. The cross-section also includes a fixed point S^* , which maps into itself under coarse-graining. The system S^* looks the same on all scales of length and time, because it coarse-grains into itself. We can picture the cross-section of Fig. 4 either as a plane in system space (in which case the arrows and flows depict projections, as in general the real flows will point somewhat out of the plane), or as the curved manifold swept out by our one-parameter model as we coarse-grain (in which case the flows above our model and below the maroon curved line should be ignored).

The flow near S^* has one unstable direction, leading outwards along the maroon curve (the unstable manifold). In system space, there is a surface of points C which flow into S^* under coarse-graining. Because S^* has only one unstable direction, C divides system space into two phases. To the left of C , the systems will have one large, system-spanning avalanche (a snapping noise). To the

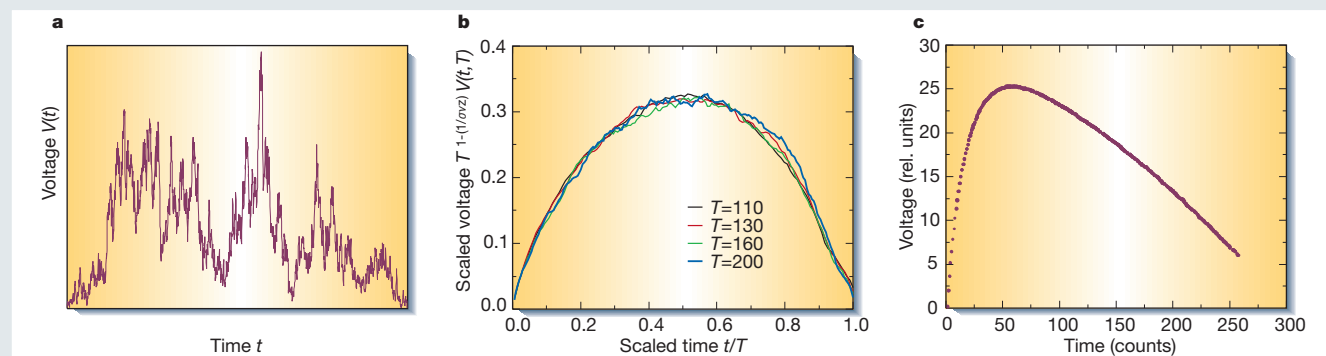


Figure 7 Scaling of avalanche shapes. **a**, Voltage pulse (number of domains flipped per unit time) during a single large avalanche (arbitrary units). Notice how the avalanche almost stops several times: if the forcing were slightly smaller, this large avalanche would have broken up into two or three smaller ones. The fact that the forcing is just large enough to keep the avalanche growing is the cause of the self-similarity: on average a partial avalanche of size S will trigger one other of size S . **b**, Average avalanche shapes⁹⁴ for avalanches of different duration for our model (A. Mehta and K.A.D., unpublished results). In addition to predicting power laws,

our theories should describe all behaviour on long scales of length and time (at least in a statistical sense). In particular, by fixing parameters one can predict what are called scaling functions. If we average the voltage as a function of time over all avalanches of a fixed duration, we obtain an average shape. In our simulation, this shape is the same for different durations. **c**, Experimental data of Spasojević *et al.*⁹⁵ showing all large avalanches averaged after scaling to fixed duration and area. The experimental average shape is very asymmetric and is not described correctly by our model.

right of **C**, all avalanches are finite and under coarse-graining they all become small (popping noise). As it crosses **C** at the value R_c , our model goes through a phase transition.

Our model at R_c is not self-similar on the shortest length scales (where the square lattice of domains still is important), but because it flows into S^* as we coarse-grain we deduce that it is self-similar on long length scales. Some phase transitions, such as ice melting into water, are abrupt and do not exhibit self-similarity. Continuous phase transitions like ours almost always have self-similar fluctuations on long length scales. In addition, note that our model at R_c will have the same self-similar structure as S^* does. Indeed, any experimental or theoretical model lying on the critical surface **C** will share the same long-wavelength critical behaviour. This is the fundamental explanation for universality.

The flows in system space can vary from one class of problems to another: the system space for some earthquake models (Fig. 5a) will have a different flow, and its fixed point will have different scaling behaviour (yielding a different universality class). In some cases, a fixed point will attract all the systems in its vicinity (no unstable directions; Fig. 6). Usually at such attracting fixed points the fluctuations become unimportant at long length scales: the Navier–Stokes equation for fluids described earlier can be viewed as a stable fixed point^{81,82}. The coarse-graining process, averaging over many degrees of freedom, naturally smoothens out fluctuations, if they are not amplified near a critical point by the unstable direction. Fluctuations can remain important when a system has random noise in a conserved property, so that fluctuations can die away only by diffusion: in these cases, the whole phase will have self-similar fluctuations, leading to generic scale invariance^{83,84}.

Sometimes, even when the system space has an unstable direction as in Fig. 4, the observed behaviour always has avalanches of all scales. This can occur simply because the physical system averages over a range of model parameters (that is, averaging over a range of R including R_c in Fig. 4). For example, this can occur by the sweeping of a parameter⁸⁵ slowly in time, or varying it gradually in space — either deliberately or through large-scale inhomogeneities.

Self-organized criticality^{36,37} can also occur, where the system is controlled so that it sits naturally on the critical surface. Self-organization to the critical point can occur through many mechanisms. In some models of earthquake faults (Fig. 5b), the external force naturally stays near the rupture point because the plates move at a fixed, but very small²⁰, velocity with respect to one another (Fig. 5b). (This probably does not occur during large earthquakes, where

inertial effects lead to temporary strain relief^{28,86}.) Sandpile models self-organize (if sand is added to the system at an infinitesimal rate) when open boundary conditions⁸⁷ are used (which allows sand to leave until the sandpile slope falls to the critical value). Long-range interactions^{88–90} between domains can act as a negative feedback in some models, yielding a net external field that remains at the critical point. For each of these cases, once the critical point is understood, adding the mechanism for self-organization is relatively easy.

The case shown in Fig. 4 of ‘plain old criticality’ is what is seen in some^{15,73–76} but not all^{88–92} models of magnetic materials, in foams⁴², and in some models of earthquakes²⁸.

Beyond power laws

The renormalization group is the theoretical basis for understanding why universality and self-similarity occur. Once we accept that different systems should sometimes share long-distance properties, though, we can quite easily derive some powerful predictions.

To take a tangible example, consider the relation between the duration of an avalanche and its size. In paper crumpling, this is not interesting: all the avalanches seem to be without internal temporal structure¹¹. But in magnets, large events take longer to finish, and have an interesting internal statistical self-similarity (Fig. 7a). If we look at all avalanches of a certain duration T in an experiment, they will have a distribution of sizes S around some average $\langle S \rangle_{\text{experiment}}(T)$. If we look at a theoretical model, it will have a corresponding average size $\langle S \rangle_{\text{theory}}(T)$. If our model describes the experiment, these functions must be essentially the same at large S and large T . We must allow for the fact that the experimental units of time and size will be different from the ones in the model: the best we can hope for is that $\langle S \rangle_{\text{experiment}}(T) = A \langle S \rangle_{\text{theory}}(T/B)$, for some rescaling factors A and B .

Now, instead of comparing our model to experiment, we can compare it to itself on a slightly larger timescale⁹³. If the timescale is expanded by a small factor $B = 1/(1 - \delta)$, then the rescaling of the size will also be small, say $1 + a\delta$. Now

$$\langle S \rangle(T) = (1 + a\delta) \langle S \rangle((1 - \delta)T) \quad (2)$$

Making δ very small yields the simple relation $a \langle S \rangle = T d \langle S \rangle / dT$, which can be solved to give the power-law relation $\langle S \rangle(T) = S_0 T^a$. The exponent a is called a critical exponent, and is a universal prediction of a given theory. (This means that if the theory correctly describes an experiment, the critical exponents will agree.) In our

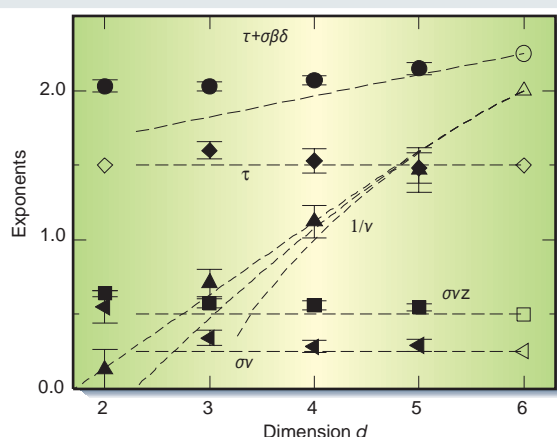


Figure 8 Critical exponents in various dimensions. We test our ε -expansion predictions^{78,80} by measuring⁷⁶ the various critical exponents numerically in up to five spatial dimensions. The various exponents are described in the text. All exponents are calculated only to linear order in ε , except for the correlation length exponent ν , where we use results from other models^{78,80}. The agreement even in three dimensions is remarkably good, considering that we are expanding in ε where $\varepsilon = 3$. Note that perturbing in dimension for our system is not only complicated, but also controversial¹¹¹ (see also section VI.C of ref. 78 and section V of ref. 76).

work, we write the exponent a relating time to size in terms of three other critical exponents, $a = 1/\sigma\nu z$.

There are several basic critical exponents, which arise in various combinations depending on the physical property being studied. The details of the naming and relationships between these exponents are not a focus of this article. Briefly, the cutoff in the avalanche size distribution in Fig. 2 gets larger as one approaches the critical disorder as $(R - R_c)^{-1/\sigma}$ (Fig. 2). The typical length L of the largest avalanche is proportional to $(R - R_c)^{-\nu}$. At R_c , the probability of having an avalanche of size S is $S^{-(\tau + \sigma\beta\delta)}$ (Fig. 2); and just at the critical field it is $S^{-\tau}$. (Note that the small change in scale δ should not be confused with the critical exponent δ .) The fractal dimension of the avalanches is $1/\sigma\nu$, meaning the spatial extent L of an avalanche is proportional to the size $S^{\sigma\nu}$. The duration T of an avalanche of spatial extent L is L^z .

To specialists in critical phenomena, these exponents are central; whole conversations will seem to rotate around various combinations of Greek letters. Critical exponents are one of the relatively easy parameters to calculate from the various analytic approaches, and so have attracted the most attention. They are derived from the eigenvalues of the linearized flows about the fixed point S^* in Fig. 4. Figure 8 shows our numerical estimates⁷⁶ for several critical exponents in our model in various spatial dimensions, together with our 6- ε expansions^{78,80} for them. Of course the key challenge is not to get analytical work to agree with numerics: it is to get theory to agree with experiment. Figure 9 shows that our model does well in describing a wide variety of experiments, but that two alternative models (with different flows around their fixed points) also fit.

Critical exponents are not everything; many other scaling predictions, explaining wide varieties of behaviour, are relatively straightforward to extract from numerical simulations. Universality extends even to those properties of long length scales for which written formulas do not yet exist. Perhaps the most important of these other predictions are the universal scaling functions. For example, consider the time history of the avalanches, $V(t)$, denoting the number of domains flipping per unit time. (We call it V because it is usually measured as a voltage in a pickup coil.) Each avalanche has large fluctuations, but we can average over many avalanches to get a typical shape. Figure 7b shows the average over all avalanches of fixed duration T , which we shall call $\langle V \rangle(T, t)$. Universality again suggests that this average should be the same for experiment and a successful

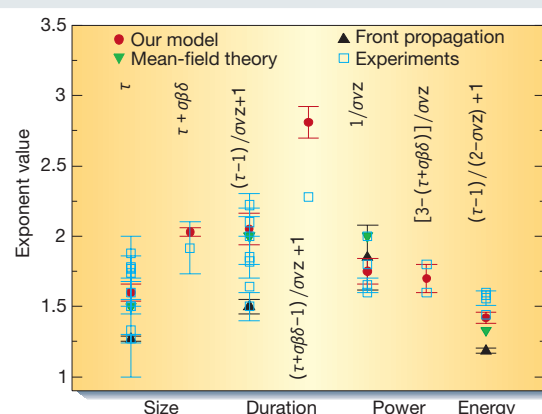


Figure 9 Comparing experiments with theory: critical exponents. Different experiments on crackling noise in magnets measure different combinations of the universal critical exponents. Here we compare experimental measurements^{88,95,112–122} (see also Table I of ref. 73) to the theoretical predictions for three models: our model^{15,73–76}, the front-propagation model^{46–51,121} and mean-field theory. (In mean-field theory our coupling J in equation (1) couples all pairs of spins: such long-range interactions occur because of the boundaries in models with magnetic dipolar forces⁹⁰. Mean-field theory is equivalent to a successful model with a single degree of freedom^{123,124}.) Shown are power laws that give the probability of obtaining an avalanche of a given size, duration or energy at the critical point; also shown is the critical exponent giving the power as a function of frequency⁹⁴ (due to the internal structure of the avalanches; Fig. 7a). In each pair of columns, the first column includes only avalanches at external fields H in equation (1) where the largest avalanches occur, and the second column (when it exists) includes all avalanches. The various combinations of the basic critical exponents can be derived from exponent equality calculations similar to the one discussed in the text^{73,80,94}. Many of the experiments were done years before the theories were developed, and many did not report error bars. All three theories do well (especially considering the possible systematic errors in fitting power laws to the experimental measurements; see Fig. 2). Recent work indicates a clumping of experimental values around the mean-field and front-propagation predictions¹²¹.

theory, apart from an overall shift in time and voltage scales: $\langle V \rangle_{\text{experiment}}(T, t) = A \langle V \rangle_{\text{theory}}(T/B, t/B)$. Comparing our model to itself with a shifted timescale becomes straightforward if we change variables: let $\nu(T, t/T) = \langle V \rangle(T, t)$, so $\nu(T, t/T) = A\nu(T/B, t/T)$. Here t/T is a particularly simple example of a scaling variable. Now, if we rescale time by a small factor $B = 1/(1 - \delta)$, we have $\nu(T, t/T) = (1 + b\delta)\nu(t/T, (1 - \delta)T)$. Again, making δ small we find $b\nu = T\partial\nu/\partial T$, with solution $\nu = \nu_0 T^b$. However, the integration constant ν_0 will now depend on t/T , $\nu_0 = V(t/T)$, so we arrive at the scaling form

$$\langle V \rangle(t, T) = T^b \mathbf{V}(t/T) \quad (3)$$

where the entire scaling function \mathbf{V} is a universal prediction of the theory.

Figure 7b,c shows the universal scaling functions \mathbf{V} for our model⁹⁴ and an experiment⁹⁵. For our model, we have drawn what are called scaling collapses, a simple but powerful way both to check that we are in the scaling regime, and to measure the universal scaling function. Using the form of the scaling equation (3), we simply plot $T^{-b} \langle V \rangle(t, T)$ versus t/T , for a series of long times T . All the plots fall onto the same curve, which means that our avalanches are large enough to be self-similar. (If in the scaling collapse the corresponding plots do not all look alike, then any power laws measured are probably accidental.) The scaling collapse also provides us with a numerical evaluation of the scaling function \mathbf{V} . Note that we use $1/\sigma\nu z - 1$ for the critical exponent b . This is an example of an

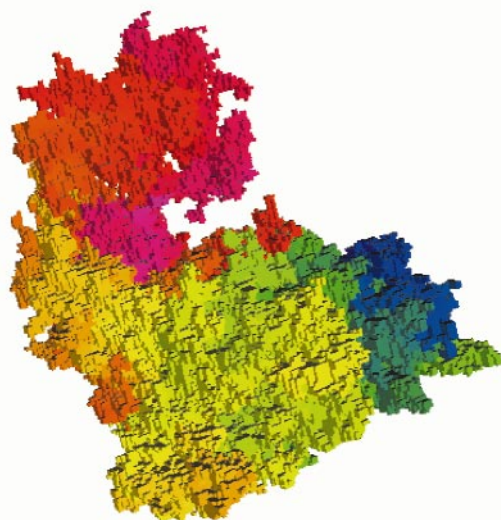


Figure 10 Fractal spatial structure of an avalanche⁷⁴. Fractal structures, as well as power laws, are characteristic of systems at their critical point. This moderate-sized avalanche involved the flipping of 282,785 domains in our simulation. The colours represent time: the first domains to flip are coloured blue, and the last pink. So far, there have not been many experiments showing the spatial structure of avalanches¹²⁵. When experiments become available, there are a wealth of predictions of the scaling theories that we could test. Other systems^{28,55,126} display a qualitatively different kind of avalanche spatial structure, where the avalanche is made up of many small disconnected pieces, which trigger one another through the waves emitted as they flip.

exponent equality and is easily derived from the fact that $\langle S \rangle(T) = \int \langle V \rangle(t, T) dt = \int T^b V(t/T) dt \sim T^{b+1}$, and the scaling relation $\langle S \rangle(T) \sim T^{1/\nu}$.

Notice that our model and the experiment have different shapes for V . The other two models from Fig. 9 also give much more symmetrical forms for V than the experiment does⁹⁴. How do we react to this? Our models are falsified if any of the predictions are shown to be wrong asymptotically on long scales of length and time. If duplication of this measurement by other groups continues to show this asymmetry, then our theory is obviously incomplete. Even if later experiments in other systems agree with our predictions, it would seem that this particular system is described by an undiscovered universality class. Incorporating insights from careful experiments to refine the theoretical models has always been crucial in the broad field of critical phenomena. The message we emphasize here is that scaling functions can provide a sharper tool for discriminating between different universality classes than critical exponents.

In broad terms, most common properties that involve large scales of length and time have scaling forms. Using self-similarity, we can write functions of N variables in terms of scaling functions of $N-1$ variables: $F(x, y, z) = z^{-\alpha} F(x/z^\beta, y/z^\gamma)$. In the inset to Fig. 2, we show the scaling collapse for the avalanche size distribution: $D(S, R) = S^{-(\tau + \alpha\beta\delta)} D((R - R_c)/S^{-\sigma})$. (This example illustrates that scaling works not only at R_c but also near R_c ; the maroon unstable manifold in Fig. 4 governs the behaviour for systems near the critical manifold C .)

Many other kinds of properties beyond critical exponents and scaling functions can be predicted from these theories. Figure 10 shows the spatial structure of a large avalanche in our model: notice that not only is it fractal (rugged on all scales), but also that it is longer than it is wide⁹⁶, and that it is topologically interesting⁹⁷. (It has tunnels, and sometimes during the avalanche it forms a tunnel and later winds itself through it, forming a knot. It is interesting that the topology of the interfaces in the three-dimensional Ising model have

applications in quantum gravity⁹⁷.) In other systems, the statistics of all of these properties have been shown to be universal on long scales of length and time.

Continuing challenges

Despite recent developments, our understanding of crackling noise is far from complete. There are only a few systems^{28,32,48,49,53,54,78–80} where the renormalization-group framework has substantially explained even the behaviour of the numerical models. There are several other approaches^{63,64,87,98–100} that have been developed to study crackling noise, many of which share our view of developing effective descriptions on long scales of length and time. But the successes remain dwarfed by the bewildering variety of systems that crackle. Achieving a global perspective on the universality classes for crackling noise remains an open challenge.

An even more important challenge is to make quantitative comparison between the theoretical models and experimental systems. We believe that going beyond power laws will be crucial in this endeavour. The past focus on critical exponents has sometimes been frustrating: it is too easy to find power laws over limited scaling ranges¹⁰¹, and too easy to find models which roughly fit them. It also seems unfulfilling, summarizing a complex morphology into a single critical exponent. We believe that measuring a power law is almost never definitive by itself: a power law in conjunction with evidence that the morphology is scale invariant (for example, a scaling collapse) is crucial. By aggressively pursuing quantitative comparisons of other, richer measures of morphology such as the universal scaling functions, we will be better able both to discriminate among theories and to ensure that a measured power law corresponds to a genuine scaling behaviour.

Another challenge is to start thinking about the key ways that these complex spatiotemporal systems differ from the phase transitions we understand from equilibrium critical behaviour. (The renormalization-group tools developed by our predecessors are seductively illuminating, and it is always easy to focus where the light is good.) For example, in several of these systems there are collective, dynamical ‘memory’ effects^{15,102–105} that may even have practical applications¹⁰⁶. The quest for a scaling theory of crackling phenomena needs to be viewed as part of the larger process of understanding the dynamics of these nonlinear, non-equilibrium systems.

A final challenge is to make the study of crackling noise profitable. Less noise from candy wrappers^{11–13} in cinemas and theatres is not the most pressing of global concerns. Making money from fluctuations in stock prices is already big business^{58,59}. Predicting earthquakes over the short term probably will not be feasible using these approaches¹⁰⁷, but longer-term prediction of impending large earthquakes may be both possible⁸⁶ and useful (for example, for guiding local building codes). Understanding that the large-scale behaviour relies on only a few emergent material parameters (disorder and external field for our model of magnetism) will lead to the study of how these parameters depend on the microphysics. We might dream, for example, of learning eventually how to shift an active earthquake fault into a harmless, continuously sliding regime by adding lubricants to the fault gouge. In the meantime, crackling noise is basic research at its elegant, fundamental best. □

- Kadanoff, L. P. Scaling laws for Ising models near T_c . *Physics* **2**, 263–272 (1966).
- Wilson, K. G. Problems in physics with many scales of length. *Sci. Am.* **241**, 140–157 (1979).
- Pfeuty, P. & Toulouse, G. *Introduction to the Renormalization Group and to Critical Phenomena* (Wiley, London, 1977).
- Yeomans, J. M. *Statistical Mechanics of Phase Transitions* (Oxford Univ. Press, Oxford, 1992).
- Fisher, M. E. Renormalization group theory: its basis and formulation in statistical physics. *Rev. Mod. Phys.* **70**, 653–681 (1998).
- Martin, P. C., Siggia, E. D. & Rose, H. A. Statistical dynamics of classical systems. *Phys. Rev. A*, **423–437** (1973).
- De Dominicis, C. Dynamics as a substitute for replicas in systems with quenched random impurities. *Phys. Rev. B* **18**, 4913–4919 (1978).
- Sompolinsky, H. & Zippelius, A. Relaxational dynamics of the Edwards–Anderson model and the mean-field theory of spin-glasses. *Phys. Rev. B* **25**, 6860–6875 (1982).
- Zippelius, A. Critical-dynamics of spin-glasses. *Phys. Rev. B* **29**, 2717–2723 (1984).
- Gutenberg, B. & Richter, C. F. *Seismicity of the Earth and Associated Phenomena* (Princeton Univ.

- Press, Princeton, 1954).
11. Houle, P. A. & Sethna, J. P. Acoustic emission from crumpling paper. *Phys. Rev. E* **54**, 278–283 (1996).
12. Kramer, E. M. & Lobkovsky, A. E. Universal power law in the noise from a crumpled elastic sheet. *Phys. Rev. E* **53**, 1465–1469 (1996).
13. Glanz, J. No hope of silencing the phantom crinklers of the opera. *New York Times* 1 June 2000, A14 (2000).
14. Sethna, J. P. Hysteresis and avalanches <<http://www.lassp.cornell.edu/sethna/hysteresis/hysteresis.html>> (1996).
15. Sethna, J. P. *et al.* Hysteresis and hierarchies: dynamics of disorder-driven first-order phase transformations. *Phys. Rev. Lett.* **70**, 3347–3351 (1993).
16. Burridge, R. & Knopoff, L. Model and theoretical seismicity. *Bull. Seismol. Soc. Am.* **57**, 3411–3471 (1967).
17. Rice, J. R. & Ruina, A. L. Stability of steady frictional slipping. *J. Appl. Mech.* **50**, 343 (1983).
18. Carlson, J. M. & Langer, J. S. Mechanical model of an earthquake fault. *Phys. Rev. A* **40**, 6470–6484 (1989).
19. Bak, P. & Tang, C. Earthquakes as a self-organized critical phenomenon. *J. Geophys. Res.* **94**, 15635–15637 (1989).
20. Chen, K., Bak, P. & Obukhov, S. P. Self-organized criticality in a crack-propagation model of earthquakes. *Phys. Rev. A* **43**, 625–630 (1991).
21. Olami, Z., Feder, H. J. S. & Christensen, K. Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. *Phys. Rev. Lett.* **68**, 1244–1247 (1992).
22. Miltenberger, P., Sornette, D. & Vanette, C. Fault self-organization and optimal random paths selected by critical spatiotemporal dynamics of earthquakes. *Phys. Rev. Lett.* **71**, 3604–3607 (1993).
23. Crowie, P. A., Vanette, C. & Sornette, D. Statistical physics model for the spatiotemporal evolution of faults. *J. Geophys. Res. Solid Earth* **98**, 21809–21821 (1993).
24. Carlson, J. M., Langer, J. S. & Shaw, B. E. Dynamics of earthquake faults. *Rev. Mod. Phys.* **66**, 657–670 (1994).
25. Myers, C. R., Shaw, B. E. & Langer, J. S. Slip complexity in a crustal-plane model of an earthquake fault. *Phys. Rev. Lett.* **77**, 972–975 (1996).
26. Shaw, B. E. & Rice, J. R. Existence of continuum complexity in the elastodynamics of repeated fault ruptures. *J. Geophys. Res.* **105**, 23791–23810 (2000).
27. Ben-Zion, Y. & Rice, J. R. Slip patterns and earthquake populations along different classes of faults in elastic solids. *J. Geophys. Res.* **100**, 12959–12983 (1995).
28. Fisher, D. S., Dahmen, K., Ramanathan, S. & Ben-Zion, Y. Statistics of earthquakes in simple models of heterogeneous faults. *Phys. Rev. Lett.* **78**, 4885–4888 (1997).
29. Fisher, D. S. Threshold behavior of charge-density waves pinned by impurities. *Phys. Rev. Lett.* **50**, 1486–1489 (1983).
30. Fisher, D. S. Sliding charge-density waves as a dynamic critical phenomenon. *Phys. Rev. B* **31**, 1396–1427 (1985).
31. Littlewood, P. B. Sliding charge-density waves: a numerical study. *Phys. Rev. B* **33**, 6694–6708 (1986).
32. Narayan, O. & Fisher, D. S. Critical behavior of sliding charge-density waves in 4- ϵ dimensions. *Phys. Rev. B* **46**, 11520–11549 (1992).
33. Middleton, A. A. & Fisher, D. S. Critical behavior of charge-density waves below threshold: numerical and scaling analysis. *Phys. Rev. B* **47**, 3530–3552 (1993).
34. Myers, C. R. & Sethna, J. P. Collective dynamics in a model of sliding charge-density waves. I. Critical behavior. *Phys. Rev. B* **47**, 11171–11192 (1993).
35. Thorne, R. E. Charge-density-wave conductors. *Phys. Today* **49**, 42–47 (1996).
36. Bak, P., Tang, C. & Wiesenfeld K. Self-organized criticality: an explanation for 1/f noise. *Phys. Rev. Lett.* **59**, 381–384 (1987).
37. Bak, P., Tang, C. & Wiesenfeld K. Self-organized criticality. *Phys. Rev. A* **38**, 364–374 (1988).
38. deGennes, P. G. *Superconductivity of Metals and Alloys* p. 83 (Benjamin, New York, 1966).
39. Feynman, R. P., Leighton, R. B. & Sands, M. *The Feynman Lectures on Physics* Vol. II Sect. 37–3 (Addison Wesley, Reading, MA, 1963–1965).
40. Jaeger, H. M., Liu, C. & Nagel, S. R. Relaxation at the angle of repose. *Phys. Rev. Lett.* **62**, 40–43 (1989).
41. Nagel, S. R. Instabilities in a sandpile. *Rev. Mod. Phys.* **64**, 321–325 (1992).
42. Tewari, S. *et al.* Statistics of shear-induced rearrangements in a two-dimensional model foam. *Phys. Rev. E* **60**, 4385–4396 (1999).
43. Solé, R. V. & Manrubia, S. C. Extinction and self-organized criticality in a model of large-scale evolution. *Phys. Rev. E* **54**, R42–R45 (1996).
44. Newman, M. E. J. Self-organized criticality, evolution, and the fossil extinction record. *Proc. R. Soc. Lond. B* **263**, 1605–1610 (1996).
45. Newman, M. E. J. & Palmer, R. G. Models of extinction: a review. Preprint [adap-org/9908002](http://xxx.lanl.gov) at <<http://xxx.lanl.gov>> (1999).
46. Cieplak, M. & Robbins, M. O. Dynamical transition in quasistatic fluid invasion in porous media. *Phys. Rev. Lett.* **60**, 2042–2045 (1988).
47. Koiller, B. & Robbins, M. O. Morphology transitions in three-dimensional domain growth with Gaussian random fields. *Phys. Rev. B* **62**, 5771–5778 (2000).
48. Nattermann, T., Stepanow, S., Tang, L. H. & Leschhorn N. Dynamics of interface depinning in a disordered medium. *J. Phys. II (Paris)* **2**, 1483–1488 (1992).
49. Narayan, O. & Fisher, D. S. Threshold critical dynamics of driven interfaces in random media. *Phys. Rev. B* **48**, 7030–7042 (1993).
50. Leschhorn, H., Nattermann, T., Stepanow, S. & Tang, L.-H. Driven interface depinning in a disordered medium. *Ann. Phys. (Leipzig)* **6**, 1–34 (1997).
51. Roters, L., Hucht, A., Lubeck, S., Nowak, U. & Usadel, K. D. Depinning transition and thermal fluctuations in the random-field Ising model. *Phys. Rev. E* **60**, 5202–5207 (1999).
52. Field, S., Witt, J., Nori, F. & Ling, X. Superconducting vortex avalanches. *Phys. Rev. Lett.* **74**, 1206–1209 (1995).
53. Ertas, D. & Kardar, M. Anisotropic scaling in depinning of a flux line. *Phys. Rev. Lett.* **73**, 1703–1706 (1994).
54. Ertas, D. & Kardar, M. Anisotropic scaling in threshold critical dynamics of driven directed lines. *Phys. Rev. B* **53**, 3520–3542 (1996).
55. Lilly, M. P., Wootters, A. H. & Hallock, R. B. Spatially extended avalanches in a hysteretic capillary condensation system: superfluid He-4 in nuclepore. *Phys. Rev. Lett.* **77**, 4222–4225 (1996).
56. Guyer, R. A. & McCall, K. R. Capillary condensation, invasion percolation, hysteresis, and discrete memory. *Phys. Rev. B* **54**, 18–21 (1996).
57. Ortin, J. *et al.* Experiments and models of avalanches in martensites. *J. Phys. IV (Paris)* **5**, 209–214 (1995).
58. Bouchaud, J. P. Power-laws in economy and finance: some ideas from physics. (Proc. Santa Fe Conf. Beyond Efficiency.) *J. Quant. Finance* (in the press); also available as preprint cond-mat/0008103 at <<http://xxx.lanl.gov>>.
59. Bak, P., Paczuski, M. & Shubik, M. Price variations in a stock market with many agents. *Physica A* **246**, 430–453 (1997).
60. Lu, E. T., Hamilton, R. J., McTiernan, J. M. & Bromond, K. R. Solar flares and avalanches in driven dissipative systems. *Astrophys. J.* **412**, 841–852 (1993).
61. Carreras, B. A., Newman, D. E., Dobson, I. & Poole, A. B. Initial evidence for self-organized criticality in electrical power system blackouts. In *Proc. 33rd Hawaii Int. Conf. Syst. Sci.* (ed. Sprague, R. H. Jr) (IEEE Comp. Soc., Los Alamitos, CA, 2000).
62. Sachtjen, M. L., Carreras, B. A. & Lynch, V. E. Disturbances in a power transmission system. *Phys. Rev. E* **61**, 4877–4882 (2000).
63. Carlson, J. M. & Doyle, J. Highly optimized tolerance: a mechanism for power laws in designed systems. *Phys. Rev. E* **60**, 1412–1427 (1999).
64. Carlson, J. M. & Doyle, J. Highly optimized tolerance: robustness and design in complex systems. *Phys. Rev. Lett.* **84**, 2529–2532 (2000).
65. Newman, M. The power of design. *Nature* **405**, 412–413 (2000).
66. Galam, S. Rational group decision making: a random field Ising model at $T=0$. *Physica A* **238**, 66–80 (1997).
67. Petri, A., Paparo, G., Vespignani, A., Alippi, A. & Costantini, M. Experimental evidence for critical dynamics in microfracturing processes. *Phys. Rev. Lett.* **73**, 3423–3426 (1994).
68. Garcimartin, A., Guarino, A., Bellon, L. & Ciliberto, S. Statistical properties of fracture precursors. *Phys. Rev. Lett.* **79**, 3202–3205 (1997).
69. Curtin, W. A. & Scher, H. Analytic model for scaling of breakdown. *Phys. Rev. Lett.* **67**, 2457–2460 (1991).
70. Herrman, H. J. & Roux, S. (eds) *Statistical Models for the Fracture of Disordered Media* (North Holland, Amsterdam, 1990).
71. Chakrabarti, B. K. & Benguigui, L. G. *Statistical Physics of Fracture and Breakdown in Disordered Systems* (Clarendon, Oxford, 1997).
72. Zapperi, S., Ray, P., Stanley, H. E. & Vespignani, A. First-order transition in the breakdown of disordered media. *Phys. Rev. Lett.* **78**, 1408–1411 (1997).
73. Perković, O., Dahmen, K. A. & Sethna, J. P. Avalanches, Barkhausen noise, and plain old criticality. *Phys. Rev. Lett.* **75**, 4528–4531 (1995).
74. Kuntz, M. C., Perković, O., Dahmen, K. A., Roberts, B. W. & Sethna, J. P. Hysteresis, avalanches, and noise: numerical methods. *Comput. Sci. Eng.* **1**, 73–81 (1999).
75. Kuntz, M. C. & Sethna, J. P. Hysteresis, avalanches, and noise: numerical methods <<http://www.lassp.cornell.edu/sethna/hysteresis/code/>> (1998).
76. Perković, O., Dahmen, K. A. & Sethna, J. P. Disorder-induced critical phenomena in hysteresis: numerical scaling in three and higher dimensions. *Phys. Rev. B* **59**, 6106–6119 (1999).
77. Berger, A., Inomata, A., Jiang, J. S., Pearson, J. E. & Bader, S. D. Experimental observation of disorder-driven hysteresis-loop criticality. *Phys. Rev. Lett.* **85**, 4176–4179 (2000).
78. Dahmen, K. A. & Sethna, J. P. Hysteresis, avalanches, and disorder induced critical scaling: a renormalization group approach. *Phys. Rev. B* **53**, 14872–14905 (1996).
79. da Silva, R. & Kardar, M. Critical hysteresis for N-component magnets. *Phys. Rev. E* **59**, 1355–1367 (1999).
80. Dahmen, K. A. & Sethna, J. P. Hysteresis loop critical exponents in 6- ϵ dimensions. *Phys. Rev. Lett.* **71**, 3222–3225 (1993).
81. Visscher, P. B. Renormalization-group derivation of Navier-Stokes equation. *J. Stat. Phys.* **38**, 989–1013 (1985).
82. Kadanoff, L. P., McNamara, G. R. & Zanetti, G. From automata to fluid flow: comparisons of simulation and theory. *Phys. Rev. A* **40**, 4527–4541 (1989).
83. Hwa, T. & Kardar, M. Dissipative transport in open systems: an investigation of self-organized criticality. *Phys. Rev. Lett.* **62**, 1813–1816 (1989).
84. Grinstein, G., Lee, D.-H. & Sachdev, S. Conservation laws, anisotropy, and “self-organized criticality” in noisy non-equilibrium systems. *Phys. Rev. Lett.* **64**, 1927–1930 (1990).
85. Sornette, D. Sweeping of an instability—an alternative to self-organized criticality to get power laws without parameter tuning. *J. Phys. I (Paris)* **4**, 209–221 (1994).
86. Sykes, L. R., Shaw, B. E. & Scholz, C. H. Rethinking earthquake prediction. *Pure Appl. Geophys.* **155**, 207 (1999).
87. Carlson, J. M., Chayes, J. T., Grannan, E. R. & Swindle, G. H. Self-organized criticality and singular diffusion. *Phys. Rev. Lett.* **65**, 2547–2550 (1990).
88. Urbach, J. S., Madison, R. C. & Markert, J. T. Interface depinning, self-organized criticality, and the Barkhausen effect. *Phys. Rev. Lett.* **75**, 276–279 (1995).
89. Narayan, O. Self-similar Barkhausen noise in magnetic domain wall motion. *Phys. Rev. Lett.* **77**, 3855–3857 (1996).
90. Zapperi, P., Cizeau, P., Durin, G. & Stanley, H. E. Dynamics of a ferromagnetic domain wall: avalanches, depinning transition, and the Barkhausen effect. *Phys. Rev. B* **58**, 6353–6366 (1998).
91. Pazmandi F., Zarand G. & Zimanyi G. T. Self-organized criticality in the hysteresis of the Sherrington-Kirkpatrick model. *Phys. Rev. Lett.* **83**, 1034–1037 (1999).
92. Pazmandi F., Zarand G. & Zimanyi G. T. Self-organized criticality in the hysteresis of the Sherrington-Kirkpatrick model. *Physica B* **275**, 207–211 (2000).
93. Perković, O., Dahmen, K. A. & Sethna, J. P. Disorder-induced critical phenomena in hysteresis: a numerical scaling analysis. Preprint cond-mat/9609072, appendix A, at <<http://xxx.lanl.gov>> (1996).
94. Kuntz, M. C. & Sethna, J. P. Noise in disordered systems: the power spectrum and dynamic exponents in avalanche models. *Phys. Rev. B* **62**, 11699–11708 (2000).
95. Spasojević, D., Bukvić, S., Milošević, S. & Stanley, H. E. Barkhausen noise: elementary signals, power laws, and scaling relations. *Phys. Rev. E* **54**, 2531–2546 (1996).
96. Family, F., Vicsek, T. & Meakin, P. Are random fractal clusters isotropic? *Phys. Rev. Lett.* **55**, 641–644 (1985).
97. Dotsenko, V. S. *et al.* Critical and topological properties of cluster boundaries in the 3D Ising model. *Phys. Rev. Lett.* **71**, 811–814 (1993).
98. Kadanoff, L. P., Nagel, S. R., Wu, L. & Zhou, S.-M. Scaling and universality in avalanches. *Phys. Rev. A* **39**, 6524–6537 (1989).
99. Dhar, D. The Abelian sandpile and related models. *Physica A* **263**, 4–25 (1999).
100. Paczuski, M., Maslov, S. & Bak, P. Avalanche dynamics in evolution, growth, and depinning models. *Phys. Rev. E* **414**–443 (1996).
101. Malcai, O., Lidar, D. A., Biham, O. & Avnir, D. Scaling range and cutoffs in empirical fractals. *Phys. Rev. E* **56**, 2817–2828 (1997).
102. Fleming, R. M. & Schneemeyer, L. F. Observation of a pulse-duration memory effect in $K_{0.30}MoO_3$. *Phys. Rev. Lett.* **33**, 2930–2932 (1986).

103. Coppersmith, S. N. & Littlewood, P. B. Pulse-duration memory effect and deformable charge-density waves. *Phys. Rev. B* **36**, 311–317 (1987).
104. Middleton, A. A. Asymptotic uniqueness of the sliding state for charge-density waves. *Phys. Rev. Lett.* **68**, 670–673 (1992).
105. Amengual, A. *et al.* Systematic study of the martensitic transformation in a Cu–Zn–Al alloy—reversibility versus irreversibility via acoustic emission. *Thermochim. Acta* **116**, 195–308 (1987).
106. Perković, O. & Sethna, J. P. Improved magnetic information storage using return-point memory. *J. Appl. Phys.* **81**, 1590–1597 (1997).
107. Pepke, S. L., Carlson, J. M. & Shaw, B. E. Prediction of large events on a dynamical model of a fault. *J. Geophys. Res.* **99**, 6769 (1994).
108. Council of the National Seismic System. Composite Earthquake Catalog Archive <<http://www.cnss.org>> (2000).
109. US Geological Survey National Earthquake Information Center. Earthquake information for the world <<http://www.neic.cr.usgs.gov>> (2001).
110. Sethna, J. P., Kuntz, M. C., & Houle, P. A. Crackling noise <<http://simscience.org/crackling>>. (1999).
111. Brézin E. & De Dominicis C. Dynamics versus replicas in the random field Ising model. *C.R. Acad. Sci. II* **327**, 383–390 (1999).
112. Cote, P. J. & Meisel, L. V. Self-organized criticality and the Barkhausen effect. *Phys. Rev. Lett.* **67**, 1334–1337 (1991).
113. Meisel, L. V. & Cote, P. J. Power laws, flicker noise, and the Barkhausen effect. *Phys. Rev. B* **46**, 10822–10828 (1992).
114. Stierstadt, K. & Boeckh, W. Die Temperaturabhängigkeit des Magnetischen Barkhauseneffekts. 3. Die Sprunggrößenverteilung längs der Magnetisierungskurve. *Z. Phys.* **186**, 154 (1965).
115. Bertotti, G., Durin, G. & Magni, A. Scaling aspects of domain wall dynamics and Barkhausen effect in ferromagnetic materials. *J. Appl. Phys.* **75**, 5490–5492 (1994).
116. Bertotti, G., Fiorillo, F. & Montorsi, A. The role of grain size in the magnetization process of soft magnetic materials. *J. Appl. Phys.* **67**, 5574–5576 (1990).
117. Lieneweg, U. Barkhausen noise of 3% Si-Fe strips after plastic deformation. *IEEE Trans. Magn.* **10**, 118–120 (1974).
118. Lieneweg, U. & Grosse-Nobis, W. Distribution of size and duration of Barkhausen pulses and energy spectrum of Barkhausen noise investigated on 81% nickel-iron after heat treatment. *Int. J. Magn.* **3**, 11–16 (1972).
119. Bittel, H. Noise of ferromagnetic materials. *IEEE Trans. Magn.* **5**, 359–365 (1969).
120. Montalenti, G. Barkhausen noise in ferromagnetic materials. *Z. Angew. Phys.* **28**, 295–300 (1970).
121. Durin, G. & Zapperi, S. Scaling exponents for Barkhausen avalanches in polycrystalline and amorphous ferromagnets. *Phys. Rev. Lett.* **84**, 4705–4708 (2000).
122. Petta, J. R. & Weissmann, M. B. Barkhausen pulse structure in an amorphous ferromagnet: characterization by high-order spectra. *Phys. Rev. E* **57**, 6363–6369 (1998).
123. Alessandro, B., Beatrice, C., Bertotti, G., & Montorsi, A. Domain-wall dynamics and Barkhausen effect in metallic ferromagnetic materials. 1. Theory. *J. Appl. Phys.* **68**, 2901–2907 (1990).
124. Alessandro, B., Beatrice, C., Bertotti, G. & Montorsi, A. Domain-wall dynamics and Barkhausen effect in metallic ferromagnetic materials. 2. Experiment. *J. Appl. Phys.* **68**, 2908–2915 (1990).
125. Walsh, B., Austvold, S. & Proksch, R. Magnetic force microscopy of avalanche dynamics in magnetic media. *J. Appl. Phys.* **84**, 5709–5714 (1998).
126. Kryscak, L. C. & Maynard, J. D. Evidence for the role of propagating stress waves during fracture. *Phys. Rev. Lett.* **81**, 4428–4431 (1998).

Acknowledgements

The perspective on this field described in this paper grew out of a collaboration with M. Kuntz. We thank A. Mehta for supplying the data for Fig. 7b, and D. Dolgert, M. Newman, J.-P. Bouchaud, L. C. Kryscak, D. Fisher and J. Thorpe for helpful comments and references. This work was supported by NSF grants, the Cornell Theory Center and IBM.

Noise to order

Troy Shinbrot & Fernando J. Muzzio

Department of Chemical & Biochemical Engineering, Rutgers University, Piscataway, New Jersey 08854, USA (e-mail: shinbrot@sol.rutgers.edu)

Patterns in natural systems abound, from the stripes on a zebra to ripples in a riverbed. In many of these systems, the appearance of an ordered state is not unexpected as the outcome of an underlying ordered process. Thus crystal growth, honeycomb manufacture and floret evolution generate regular and predictable patterns. Intrinsically noisy and disordered processes such as thermal fluctuations or mechanically randomized scattering generate surprisingly similar patterns. Here we discuss some of the underlying mechanisms believed to be at the heart of these similarities.

If one rubs a piece of flannel across a blackboard, chalk from the board becomes distributed uniformly across both flannel and board, erasing a pattern drawn earlier. The converse does not occur: rubbing flannel on a blackboard does not cause chalk to leave a more ordered state than found initially. This is a crude but straightforward consequence of the second law of thermodynamics, which roughly stated dictates that disorder tends to increase. On the other hand — as is well known to any parent who has organized a child's party — if one rubs the same flannel on a plastic balloon, charge transfer can take place between cloth and balloon, against the apparent electrostatic gradient¹, leaving the balloon highly charged and predisposed to be stuck to a nearby wall. How is it that chalk, when rubbed, tends to become more uniformly distributed, whereas plastic, when rubbed, causes surface charges to become more strongly separated?

This example, although unsophisticated, is representative of a growing number of problems in modern physics in which noisy processes (for example, rubbing) can result in increased ordering². Noise-induced ordering can take many forms, ranging from mere separation of different materials into a heterogeneous final state (as in the example of charges on a balloon) to the formation of a rich variety of regular patterns, including stripes, squares, hexagons and spirals. These phenomena are seen in catalysis³, cosmology^{4–6}, marine biology⁷, reactive mixing^{8–10}, colloidal chemistry¹¹ and geophysics^{12–14}. In some of these problems, most notably in colloids research^{11,15}, mechanisms such as steric stabilization have long been understood, whereas in others the causes and implications of spontaneous pattern formation continue to generate developments that are theoretically meaningful and of practical importance. In this article we review some of the high points in historical progress in the field and focus on research in two of the fastest growing areas: granular physics and phase separation.

The use of noise for productive purposes has a long and colourful history. For over a century it has been reported by sailors that disordered raindrops falling on the ocean will calm rough seas^{16–18}. By comparison, noise in the form of heat has for many decades been used in the electronics industry to instigate charge release from surfaces and so amplify signals (for example, in vacuum tubes), deposit ions (for example, in thin-film fabrication), or generate displays (for example, in cathode ray tubes). Noise in the form of tiny disturbances provokes instabilities that lead to regular oscillations in systems ranging from musical instruments and humming power lines to road corrugation and catastrophic structural collapse¹⁹. Noise is also important in

increasing signal-processing effectiveness in applications as diverse as neuroprocessing and palaeoclimatology²⁰, and the addition of noise to pattern-forming systems can actually enhance the pattern-formation mechanism in several problems^{10,21,22}. In other applications, aspects of noise can improve the regularity of long-term biological rhythms²³.

For our purposes, it is useful to differentiate between problems in which noise enhances or initiates an existing process, and qualitatively different problems in which order would not appear at all without the presence of noise. This distinction is not always clear-cut, as in the case of convection rolls that are arguably both initiated and sustained by noise in the form of heat. Nevertheless, the issues we seek to explore are how noise itself produces ordered states, and what can be learnt about patterns in natural systems from the mechanisms that permit this to occur.

Noise, order and the second law of thermodynamics

Noise at the macroscopic scale, or equivalently diffusion at the microscopic scale, has long been investigated in chemical and thermodynamic systems. Historically, study of the emergence of periodic patterns from diffusive processes has had a lively development, beginning arguably with the travails of Boris Belousov, who in the 1950s reported to a disbelieving scientific community²⁴ that chemical reactions can evolve periodically in time towards a heterogeneous final state (see Box 1). The reason for conventional disbelief in this, now well accepted, result lies at the heart of the present discussion. That is, diffusion is characterized by a monotonic tendency towards homogeneity as dictated by the equation:

$$\frac{\partial C_A}{\partial t} = D_A \nabla^2 C_A \quad (1)$$

where C_A is the concentration of a substance of interest, and D_A is a diffusion coefficient. It is self-evident that this equation is first order in time and prescribes a monotonic change in C_A . It therefore seems paradoxical that diffusion combined with any sequence of terminal reactions can produce evolving patterns. Heuristic analysis on the microscale produces a similar result: diffusion is nothing more than random motion of particles caused by brownian agitation, and it is not at all obvious that anything other than homogeneity can result from such motion. Indeed, the second law of thermodynamics hinges on this very principle. We stress at the outset that the second law is robust in the face of all of the examples that we will present, as can be confirmed through a careful accounting of energy and entropy²⁵. Nevertheless, the overt behaviours found in these problems can be strikingly counterintuitive.

As an example, consider a macroscale version of the Maxwell's Demon problem. Maxwell's Demon is the

paradigmatic thought experiment used to challenge the second law²⁶, and consists of the hypothetical process of extracting heat from a gas by opening an imaginary door separating two chambers only when high-speed (that is, hot) molecules approach the door from, say, the right. In this way, heat could conceivably be transferred from one chamber to its neighbour without the expenditure of energy. A meticulous analysis of the actual energy required to implement such a machine reveals that the second law is sound because the minimum expenditure of energy needed to detect and respond to fast particles is

Box 1

Liesegang bands

Geology is rich with patterned states, ranging from large sand-dune patterns⁷⁶ and cellular stone formations² to striated rocks⁷⁷, gems^{15,78} and sedimentary structures^{79,80}. Many of these patterns are so complex that scientists continue to debate whether they were produced by organized, biological processes or disorganized, inorganic means^{12,81–83}. One of the earliest situations where geological patterns governed by inorganic diffusion have been analysed is in 'Liesegang' bands, proposed to be responsible for the striped appearance of iris agates⁷⁸. These bands, shown in the figure below, occur in gels or other media where convection is suppressed and only diffusive mixing is possible³⁵. In its simplest analysis, the mechanism is as follows. A solute (added from above in the experiment illustrated) diffuses through a static medium that may contain a second reactant. If the solute or its reaction product is present in sufficient concentration, it can supersaturate the medium and subsequently precipitate out of solution in a local region. The precipitate depletes the region of solute, thus delaying its accumulation downstream. The solute diffuses through the obstacle presented by the precipitate to a distance downstream where it once again accumulates to reach supersaturation levels. At this point the cycle repeats itself. In the wake of these cycles of diffusion, supersaturation and precipitation, bands of precipitate and low-concentration solute are left behind.

This analysis can be embellished by including the effects of unavoidable complications such as differential solubilities and crystallization speeds as well as weak convective cycles^{79,80}. Notwithstanding these, sometimes significant, refinements, a precipitation–diffusion cycle seems to be the dominant mechanism at work in the formation of Liesegang patterns. We note in particular that, as presaged by Belousov (see text), although the kinetics are governed by a first-order diffusion equation, spatial pattern formation occurs as a result of coupling between this first-order equation and a more complicated, supersaturation–precipitation cycle.



Box 1 Figure Liesegang patterns in a cylinder containing agar gel with dissolved K_2CrO_4 to which $CuSO_4$ has been added from above. See <http://qsad.bu.edu/ogaf/lies/> for details of this system suitable for instructional use. (Photo: P. Garik, M. B. Gillespie and K. Brecher, Mathematics Education Center, Boston University.)

nonzero and in fact exactly equals that predicted by quantitative forms of the law²⁵.

But in the case of granular physics, where much recent research on noise and pattern formation has focused, a surprising result is obtained. This result, first discussed²⁷ in 1996 and since analysed in detail²⁸, is duplicated in Fig. 1, where we display the results of an experiment in which a number of steel beads were initially distributed uniformly across a vibrated acrylic container. The container is separated into two identical chambers by a foamboard barrier containing an open window near its bottom.

Despite being distributed uniformly at the start, after a short time (2–3 minutes) most of the beads migrate spontaneously to one chamber or the other. This separates the system into a first chamber containing a sparse and gas-like state of high-speed particles, and a second containing a dense and close-packed state of nearly motionless beads. This separation of 'hot' and 'cold' states occurs in the presence of noisy agitation in the form of a macroscopic analogue of brownian motion: experiments²⁹ and simulations^{30,31} have shown that the particles' horizontal velocities are effectively randomized when vibrated energetically in this way.

Unlike the classic Maxwell's Demon experiment, the window remains open, the two sides are identical, and particles are free to travel between the chambers at all times. Also unlike the classic case — and central to the ordering mechanism in this and all other problems — this system has an intrinsic and pronounced source of dissipation. The dissipation here results from inelastic collisions between the beads, and the mechanism by which this dissipation leads to spontaneous heating and freezing of adjacent chambers is straightforward: each time a bead suffers a collision, it loses kinetic energy to internal heat owing to inelasticity of the bead material. Because energy is lost with every collision, energy delivered to the bottom of a bed of beads must be lost exponentially with height (as measured in numbers of beads) through the bed, and a small surplus of beads in one of the two chambers can produce a large excess in dissipation in that chamber. This means that once, by chance, a small surplus of beads appears on either side of the barrier, that side will lose energy more rapidly than the opposing side, and will therefore be able to eject fewer beads through the window than it receives. For suitable vibration conditions²⁸, this instability accelerates rapidly, causing one chamber to lose so much energy that its beads effectively become frozen, while the other chamber loses increasingly more particles, and each particle remaining experiences fewer collisions and so remains highly energetic.

This is the state shown in Fig. 1, where a 'hot gas' has apparently separated spontaneously from a 'frozen solid'. This separation occurs only in this macroscopic view: a full analysis both of the actual heating of the grains and of the heating of the environment as a result of energy input through an external shaker would readily reveal that entropy overall in this nonequilibrium system increases (markedly) with time. Nevertheless, the apparent effect is that a disordered and homogenous initial state has been transformed into an ordered and heterogeneous state as a result of noisy agitation.

Patterns from microscale diffusion

Against expectation, ordering of macroscopic particles into gas-like and frozen states can arise spontaneously. The ingredients required to produce this outcome are a source of noisy energy (here the shaking of the container) and a source of dissipation (the inelasticity of the particles). The same ingredients on the microscopic scale produce regular ordering in a variety of natural situations as well. Historically, one of the earliest mentions of this kind of ordering — and one of the most elegant systems in which it can be produced — is in the production of layered geological specimens, where 'Liesegang bands' of precipitating reactants can form spontaneously, resulting in stripes and other patterns in otherwise static gels (see Box 1).

The possibility that patterns could form in natural systems due only to reaction combined with molecular noise (that is, diffusion) was proposed in 1952 by Alan Turing. The emergence of these

Figure 1 Separation of hot and cold beads in a granular 'Maxwell's Demon' experiment.

a, Snapshot of spontaneous separation between gas-like and solid-like steel beads (diameter 2 mm) from an experiment in which an acrylic container is vibrated sinusoidally at 10 Hz and a maximum acceleration of 1.3g. Initially the beads are distributed uniformly

across the container, but owing to collisional inelasticity they migrate to whichever side randomly acquires a slight excess of beads. **b**, Schematic of container which is divided into two chambers by a barrier containing a window 6 mm high, beginning 6 mm from the bottom of the container.



patterns from an initially homogeneous state seems to run counter to the simple analysis of equation (1). The paradoxical appearance of patterns in experiments occurs when two reactants diffuse at different rates, thereby generating differing reaction outcomes as their gradients change. This is a subtle effect with profound consequences, and permits the transmission of neuronal waves in biological systems such as the heart^{32,33}, as well as the emergence of complex ecological³⁴, chemical^{8–10} and bacterial² patterns. In all of these varied problems, equation (1) is modified first by the inclusion of a reacting term and second by a coupling to a second 'reaction-diffusion' equation describing another species. In neurons, the two equations relate to discharge and repolarization across membranes; in ecology they relate to differential flow and uptake of nutrients; and in catalytic and other reactions they relate to different diffusivities of constituent reactants and by-products.

Example

One of the simplest scenarios in which this effect is seen is the following set of reactions, simulated^{9,35} in 1993 for reactants A and B:



Reaction such as (2a) are seen in biological problems such as when a nutrient, A, is added to a population of cells, B, yielding more cells, or in crystallization problems where an antisolvent, A, is added to a saturated solution to increase the abundance of a crystalline product, B.

Reaction (2b) represents a decay such as occurs during ageing or oxidation of biological or chemical systems.

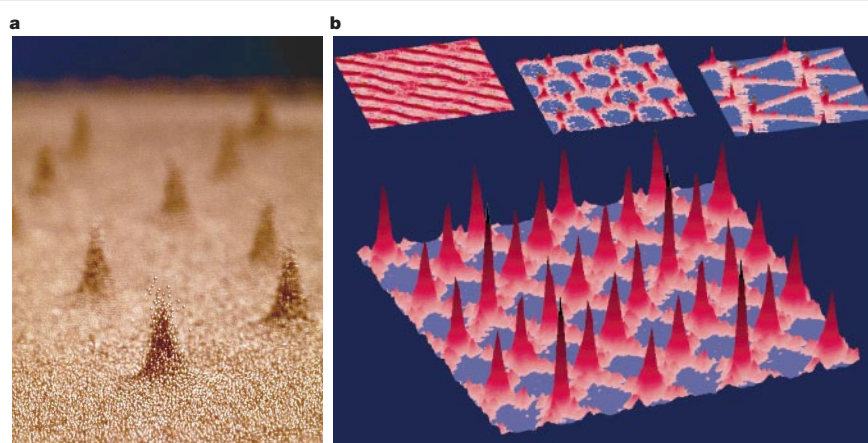
When reactant A is fed continuously and uniformly into the system, these reactions, suitably normalized, can be described by the coupled equations:

$$\frac{\partial C_A}{\partial t} = D_A \nabla^2 C_A - C_A C_B^2 + S_A (1 - C_A) \quad (3a)$$

$$\frac{\partial C_B}{\partial t} = D_B \nabla^2 C_B + C_A C_B^2 - (S_A + S_B) C_B \quad (3b)$$

where $S_{A,B}$ are constants regulating feed and reaction rates and $D_{A,B}$ are diffusion coefficients. The first term on the right side of each equation describes diffusion of each reactant, the second describes the reaction (2a), and the third describes feed of reactant A and consumption of A and B through reactions (2a) and (2b). When $D_A > D_B$, this system can give rise to complex patterns including stripes, spots, labyrinths and spirals. In this system, the heuristic mechanism for pattern formation is (similar to that of Liesegang bands discussed in Box 1) that reactant A spreads more rapidly than reactant B, and when the feed rate defined by S_A is sufficiently small, A becomes depleted in local regions through reaction (2a). When this occurs, the resulting excess of reactant B leads to a new product C in that region, as indicated in reaction (2b). This in turn produces a local depletion of B, enabling the feed to produce a local surplus of

Figure 2 Granular patterns from experiment and model. **a**, Array of 'oscillons' from a granular vibration experiment. (Photo: P. Umbanhowar and H. L. Swinney, Univ. Texas at Austin.) **b**, Patterns from simulations of 32,767 idealized particles that are randomized periodically and collide inelastically (see text). Plotted are densities of particles as a function of position, in square state (main plot: $f=0.5$, $V=2,500$, $T=1.25 \times 10^{-4}$), in stripes (left inset: $f=0.5$, $V=1,500$, $T=1.5 \times 10^{-4}$), in hexagons (centre inset: $f=0.5$, $V=2,000$, $T=2 \times 10^{-4}$) and in triangles (right inset: $f=0.5$, $V=3,500$, $T=2 \times 10^{-4}$).



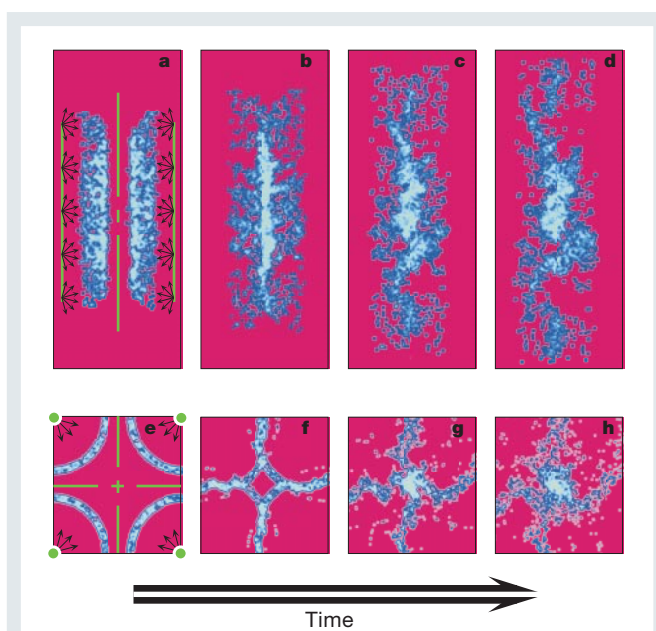


Figure 3 Stability of striped and square lattice patterns. Response of particles distributed near parallel stripes (a–d) or square gridpoints (e–h) to an instantaneous noisy impulse in the model simulation described in the text. The dashed green lines identify reflection symmetries of each configuration, and the solid green lines and small black arrows indicate locations and directions of initial particles. a–b, The striped initial state generates a new stripe at symmetry line. c–d, The new stripe is unstable to transverse fluctuations, causing it to decompose into clusters. e–h, The instability of the striped state causes clusters to form at the intersections of symmetry lines, thus stabilizing the square and other patterned states.

unreacted A. This surplus diffuses towards regions that have become depleted of A, and the cycle repeats itself.

Although this scenario might appear to be implausibly complicated at first blush, with some reflection the underlying mechanism can be seen to be so simple as to actually be quite probable. This is perhaps most clear in problems involving excitable media. In these media, a biological or chemical fuel is consumed in a local region, and the region cannot support further activity until the fuel has been replenished. Thus as an activation wave travels through neuronal tissue, cells left in its wake must wait for a refractory period to elapse before being able to fire again³². Sequential wave fronts, separated characteristically by the product of the wave speed and the refractory period, can thus be supported in such media. A similar explanation is obtained in combustion^{36,37} or catalysis³ experiments, when fuel on an extended surface is consumed more rapidly than the available oxygen can sustain. In this case, the reaction becomes extinguished by its own by-products, which must diffuse away before new reactions can proceed.

In reaction-diffusion problems, the difference between diffusivities D_A and D_B is central to the emergence of patterns. In a system with identical diffusivities, the spatial term can be eliminated from equations (3a) and (3b), leaving a single first-order equation that is incapable of describing spatial patterns. But the case $D_A \equiv D_B$ is non-generic and so there are numerous practical problems that are subject to pattern formation. The case $D_A < D_B$, on the other hand, is much more common, and in this situation, reactant A diffuses slower than heterogeneities can develop, thus stabilizing the system and leading to eventual homogeneity.

The reason for the required asymmetry of D_A and D_B is that B reacts nonlinearly, whereas A does not. This nonlinearity is a manifestation of the reproduction of B in reaction (2a), which augments local concentrations of B in much the same way that elastic dissipation promotes the concentrated frozen state in the granular Maxwell's Demon experiment. The same is not true of reactant A, which responds linearly and can increase its concentration only in

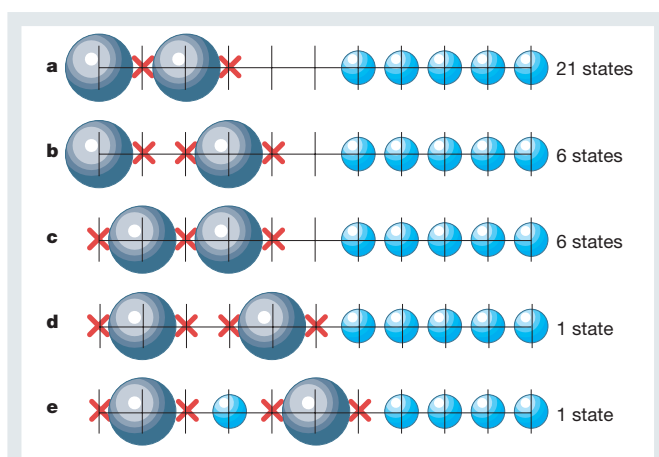


Figure 4 Entropic ordering example in simple one-dimensional lattice. Several of the 270 possible configurations consisting of 2 large (grey) and 5 small (blue) particles constrained to lie on a one-dimensional lattice of gridpoints. Excluded volumes surrounding the large particles are indicated by red x's, and the numbers of distinct states of small particles for each configuration of large particles depicted are indicated on the right. For randomly placed particles, states in which like particles lie together are favoured, and states in which large particles lie near an edge are greatly favoured.

response to the external supply. It is competition between diffusive gradients (that is, spreading caused by noise) and localization (resulting from nonlinear reactions) that produces patterns in reaction-diffusion systems. In the next section we describe larger-scale analogues of this same competition.

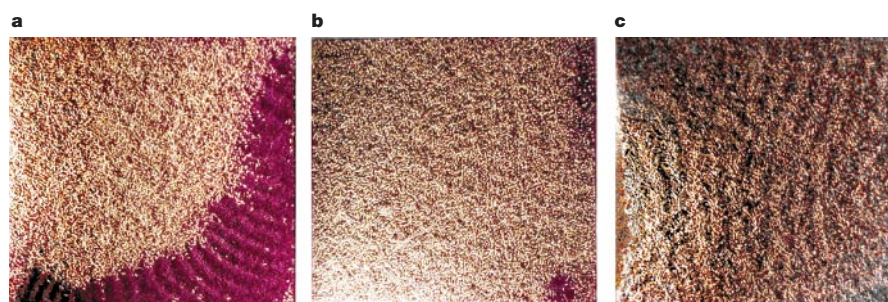
Patterns from macroscale noise

Patterns are seen on all scales, from the microscopic to the astrophysical. Thus complementing chemical and microbial² patterns, the regular spiral structure of many galaxies is believed to be due to dissipative collisions imposed on a background of initially random star locations^{5,6,38}. At larger scales still, scientific debate has intensified during the past decade over the possible presence of periodic patterns of clusters and superclusters of galaxies extending over as much as hundreds of millions of parsecs^{4,39,40}. These patterns are strongly reminiscent of results from simulations⁴¹ of inelastic hard spheres used to model granular flows, which we turn to next.

On laboratory scales, a variety of striking patterns have been reported^{42,43} in vibrated granular experiments similar to the granular Maxwell's Demon system of Fig. 1. Among the most intriguing are 'oscillons', which consist of a state of peaks and dips that alternate on successive vibration cycles⁴⁴. An array of oscillon peaks from a granular vibration experiment⁴⁴ is shown in Fig. 2a. As with chemical and astrophysical problems, many of the patterns seen in vibrated grains can be duplicated by models^{30,45–47} that combine the effects of noise or diffusion with dissipative or nonlinear interactions. Results of one such model are shown in Fig. 2b. Note that the square pattern in Fig. 2b differs from the oscillon array in Fig. 2a in that the peaks in this simulation appear in periodic lattices, whereas oscillons have been generated in both periodic and aperiodic arrangements^{44,46}.

Vibrated granular beds represent a particularly illuminating archetype for understanding the mechanism of pattern formation from noise. We have seen (Fig. 1) that the significance of dissipation for the generation of localized states is easily understood in granular beds. But localization (referred to as 'clustering' in the granular literature) is only part of the process: a clustered state does not necessarily display any regular pattern, such as the squares, stripes, hexagons or triangles shown in Fig. 2b. The other characteristic of importance is the regularity of the pattern itself. This regularity requires an additional ingredient, namely an underlying set of symmetries. Thus each of the patterns shown in Fig. 2b represents a particular symmetry. In

Figure 5 Possible transition from segregation to mixing in a vibrated granular bed. Possible granular transition between mixing and segregation of 1.8-mm (gold) and 0.8-mm (maroon) glass beads. Vibration is approximately sinusoidal with peak acceleration of 3.9g and frequency of 20 Hz. Under fixed conditions, the concentrations (as a weight fraction of the total granular bed) of small beads in the three snapshots are: **a**, 30%; **b**, 12%; **c**, 11%. Striped patterns, characteristic of vibrated grains, are evident in the figure but have no known role in the entropic segregation mechanism.



many problems, the source of these symmetries can be abstruse; in the granular case it is comparatively transparent.

Example

The model of Fig. 2b was engineered to capture the essential causes of pattern formation due to noise combined with dissipation. The model is a simplification of these two effects applied to the particular problem of vibrated beds of grains. High-speed photographs⁴⁸ of granular patterns reveal two distinct stages of motion. First, the supporting plate accelerates upwards to strike the granular bed. Patterns are at their sharpest at the moment of impact, and scatter apart thereafter. Second, the supporting plate accelerates downwards faster than gravity (accelerations of 3–8g are typical). The granular bed then separates from the supporting plate, and particles collide with one another to reform a patterned state, typically with the vertices of the new pattern at the centres of the previous pattern⁴⁹. Thus the simplified model for the system subjects a large number of identical discrete particles to two sequential processes: first, once per driving period, T , particles are scattered apart; and second, they suffer inelastic collisions thereafter³⁰.

The scattering process is simulated by instantaneously randomizing all particle velocities once per cycle. Randomization can be due to many different effects in problems as diverse as chemical waves and interstellar interactions; the randomization in the model of Fig. 2b is associated with multiple collisions between convex particles which effectively disorder particle vectors^{50,51}, in agreement with experimental and direct numerical data for the vibrated bed^{29,42}. Explicitly, particle velocities in Fig. 2b are reassigned periodically according to

$$\mathbf{U}_i \rightarrow f\mathbf{U}_i + V\eta_i \quad (4)$$

where \mathbf{U}_i is the velocity of the i -th particle, η_i is a noise vector that is random in direction for each particle and is fixed in magnitude, and the arrow symbolizes evolution of the velocity after impact. The terms f and V are constants that respectively determine the particles' memories of their prior velocities and the strength of the impact. In the square pattern shown in Fig. 2b, we choose $f = 0.5$, $V = 2,500$, and we randomize velocities of each particle every 1.25×10^{-4} natural units (parameters for the other patterns shown are defined in the figure caption).

The dissipative process is intended to mimic the results of inelastic particle collisions, shown in Fig. 1 to generate spontaneous concentrations of particles^{41,49,52}. This too can be modelled in many ways; in Fig. 2b we show the results of a mean-field approach in which collisions are non-binary and follow the rule

$$\mathbf{U}_i \rightarrow \frac{(1 + n_i)\mathbf{U}_i + (1 - \varepsilon)n_i\mathbf{U}_j}{1 + n_i} \quad (5)$$

where \mathbf{U}_i and \mathbf{U}_j are respectively the velocities of the i -th particle and the mean velocity of its n_i neighbours within a specified radius, and ε

is the effective coefficient of restitution for collisions. When not under the influence of periodic randomization (equation (4)) or collisions with neighbours (equation (5)), particles travel ballistically in the plane. The model is strictly two-dimensional and may be thought of as a planar projection of the true three-dimensional problem. In Fig. 2b, we use $\varepsilon = 0.25$, and integrate trajectories of 32,767 initially randomly placed particles with a time step of 5×10^{-6} natural units. The figure shows the density of particles on a planar computational domain with periodic boundaries.

In addition to the square pattern, this simple model produces (insets to Fig. 2b) stripes, hexagons, triangles and other states as parameters such as the randomization period T and the amplitude V are varied. It may seem surprising that such a simplified model, containing little more than frequent randomization combined with dissipation, and entirely neglecting gravity, friction and other realistic effects, can produce an apparent wealth of spontaneously organized and highly structured patterns. As we discuss in the next section, an alternative view is that these patterns actually form *because* the system is randomized and dissipative. These features can by themselves dictate the formation of patterns independent of specifics of the problem under study. This is an extremely powerful conclusion that goes some way to explaining how so many different physical, chemical and biological systems develop such similar patterned states².

Role of symmetries

In the granular case, the mechanism for the formation and selection³⁰ of patterns, rather than disorganized clusters⁴¹, in the presence of noise is especially straightforward to analyse. Consider the case depicted in the top panels of Fig. 3, where a large number (4,000) of stationary particles, initially distributed near two parallel stripes, are instantaneously given a noise impulse of constant amplitude V and random direction. The random distribution of directions assures that particles will expand outwards from their initial configuration, and the half of particles that travel towards the centre line separating the initial stripes are plotted in Fig. 3a–d. These particles cannot avoid colliding, and when they do so they spawn a new stripe midway between the original ones (Fig. 3b). If the particle velocities are again randomized at close to this time (approximately D/V where D is the separation between the initial stripes), the new stripe will expand outwards as before. In a plane tiled initially with stripes, the striped state will reinforce itself, generating new stripes after each randomizing event, interlaced with the stripes before the event. This is what is seen both in experiments⁴⁸ and in simulations^{30,42}.

We emphasize that the only function of the noisy impulse in the pattern-formation mechanism is to cause high-concentration regions to expand outwards. From this perspective, any process that produces such an outwards expansion is equivalent — whether due to an instantaneous impulse, or to steady diffusion, or to some other (for example, higher-order) mechanism. It is for this reason that patterns caused by noise so closely resemble patterns found in deterministic systems and that patterns generated by different processes are so similar.

Which particular patterned state will predominate at given parameter values, on the other hand, is a question of stability, and here the specifics of a given problem enter into consideration. In the model system of Fig. 3, for example, the striped state is unstable to transverse fluctuations that spoil the striped pattern if the randomizing event is delayed much beyond the time of Fig. 3b. This is shown in Fig. 3c, d, where we display snapshots of the continued evolution of the two-stripe state after successive, equal increments of time. The root cause of the instability in this particular system is that conservation of momentum guarantees that mean particle velocities parallel to the stripes will persist after collisions have run their course. Consequently, clusters form along newly spawned stripes; these clusters are evident both in Fig. 3d and in the upper left inset to Fig. 2b.

So when particles are randomized at periods close to $\tau = D/V$, transverse fluctuations do not have time to grow and the striped state is stable. But if randomizing events come much further apart than τ , this state loses stability to a clustered state. One such state is the square pattern in Fig. 2e–h, which shows a time series after noise has been added instantaneously to velocities of 4,000 particles clustered near the corners of a square. In this case, particles again explode away from their initial locations (Fig. 3e), collide inelastically along the symmetry lines separating the initial clusters (Fig. 3f), and then form new clusters at locations centred between the former clusters (Fig. 3g, h). If the plane is tiled with clusters located at the gridpoints of a square lattice, this state can reinforce itself, and so the very effect that destabilizes the striped pattern can stabilize the square pattern.

In this model system, the form of the particular instability that leads from one patterned state to another is heuristically transparent. In other problems, the instability will differ and may be more difficult to determine. In all pattern-formation problems, whether associated with noise or not, reflection, translation and rotation symmetries can be found to underlie the overt pattern. Thus the striped state is associated with symmetries of reflection along lines (dashed line in Fig. 3a) separating the initial stripes and of translation along the stripe direction. The square state is associated with reflection about two perpendicular lines (dashed lines in Fig. 3e). The square state appears exactly when the translational symmetry of the striped state is broken. Without these symmetries (as in problems where no interlaced dual pattern exists because, for example, periodic driving is absent^{41,53}) clustering can occur as a result of dissipation, but periodic pattern formation will not.

When do disordered states occur?

Spontaneously organized patterns can be formed by combining little more than noise (or diffusion) with dissipation (or nonlinear reactions) in a symmetric system. Why is it, then, that in some problems — such as the rubbing of a blackboard with flannel — a disordered, homogeneous state results, whereas in similar circumstances — the rubbing of a balloon with the same flannel — ordered, heterogeneous states appear⁵⁴?

To address this question, we paradoxically again turn to entropy. We have mentioned that the second law guarantees that the continual infusion of energy needed to maintain nonequilibrium states such as those shown in Figs 1 and 2a will always be sufficient to increase the entropy of the Universe. This coarse calculation is reassuring, but more can be learned by evaluating the entropy on a microcanonical basis — that is, on probabilistic grounds, one can conclude that the physically realized states that nature chooses tend to correspond to the ones that maximize the number of possible particle rearrangements. By counting all possible states, one finds that so-called ‘entropic ordering’ can favour either heterogeneous or homogeneous states with a transition at calculable values of order parameters, such as constituent concentrations.

As an example, suppose that we calculate the number of possible rearrangements of large and small particles constrained so that particle centres must lie on a one-dimensional grid⁵⁵. In Fig. 4 we sketch a much-simplified case where statistics of the entire microcanonical

Box 2

Liquid-crystal patterns

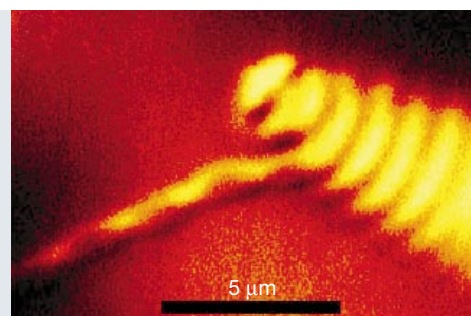
Colloidal ordering has been known for many years to be governed by entropy⁵⁴. This is important in several technological problems, including the generation of liquid-crystal displays and the fabrication of polymer composites. Intriguing examples in biological systems have also been found. As an example, the figure below displays an intermediate stage in the kinetics of the isotropic–smectic transition of a polymer/virus mixture. Initially, droplets of the metastable nematic phase are formed out of the isotropic background. Next, a single layer of the smectic phase is nucleated on the surface of the nematic drop. Growth proceeds by coalescence of similar drops until large droplets wrapped in smectic layers are formed. When a droplet is fully wound with smectic layers, a smectic strand peels away from the droplet, as shown in the figure. The strand has a uniform diameter of one-rod length. The rods are oriented perpendicular to the axis of the strand and rotate in this plane as seen by the modulations in intensity. The strands grow hundreds of microns long, unwind, and settle to the bottom of the container. Over a period of several months they transform into large, single-layer smectic sheets. These then stack on top of one other and eventually a single-phase, multilayer smectic monolith is formed.

Box 2 Figure

Entropic ordering of viral particles.

A tactoid of spontaneously aggregated rod-like viral particles. This micrograph was taken using a $\times 60$ water immersion lens

and differential-interference contrast optics and was subsequently false coloured. (Photo: Z. Dogic and S. Fraden, Brandeis University.)



ensemble can be evaluated using simple arithmetic. If we choose a representative example in which the large particles have radii just over 1.5 grid spacings, then these particles can exclude other particles from occupying adjacent gridpoints. This is indicated in Fig. 4, where the large grey particles are surrounded by excluded gridpoints indicated by red x's. The small blue particles can be arranged in $R = g!/s!(g-s)!$ distinct configurations (that is, excluding rearrangements of identical particles), where g is the number available gridpoints and s is the number of small particles. Notice that the volume excluded around the large particles increases either when the large particles are separated or when the large particles stray away from the edges of the grid. Correspondingly, in either of these situations the number of possible rearrangements of the small particles plunges from $R = 21$ to $R = 1$. This 21-fold change in the number of possible rearrangements implies that, just based on counting, we can conclude that it is highly probable that a randomly placed state will have large particles near one another, and that it is even more probable that the large particles will be near the edges^{56,57}.

If we calculate all possible rearrangements of both large and small particles, which is possible in this tiny illustrative problem, we find 154 distinct ‘separated’ states where all of one or the other species is contiguous, and only 116 ‘mixed’ states having at least one different particle intervening between like particles. Of the separated states, all but 24 have one particle adjacent to an edge. Thus there is a higher likelihood that we will find randomly distributed particles in a separated state than in a mixed state, and by far the highest likelihood

is that the large particles will cluster beside an edge. Notice that once again, irrespective of the specifics of any given problem, *any* system that is subject to random (that is, noisy) rearrangements will tend preferentially towards these separated states.

States in which particles separate into heterogeneous arrangements can be significantly favoured in randomized systems. In larger, more realistic problems, the probability contrast becomes much more extreme and separated states can be enormously more probable than any mixed state. This enhanced probability of obtaining separation due to 'volume-exclusion' effects either in the bulk^{58,59} or near boundaries^{56,57} persists in continuous^{56–64} as well as in simplified lattice-based examples, and has been reported widely in a variety of practical colloidal systems (see Box 2 for a topical example). But notice that this mechanism depends crucially on the concentrations of the two species of particles. By changing concentrations, one can tune the relative likelihood of obtaining a separated or a mixed state. This fact has been exploited in practical problems. For example, simulations^{58,60–64} and experiments¹¹ have confirmed that polymers of various shapes and sizes separate into distinct phases precisely when statistical calculations indicate that separation is favoured entropically over mixing. This result is of technological importance, for it implies that patterned displays, nanostructures and polymer blends can be engineered scientifically from first principles^{11,56}.

Using the volume-exclusion model, we can serendipitously directly compare microscale, entropically governed phase separation, observed in colloidal and polymeric systems, and macroscale noise-dominated segregation^{59,65}, long known to be ubiquitous (and troublesome) in granular systems⁶⁶. It is notable in this regard that there is little difference between research methods, results, or directions on phase separations of polymers and colloids and on segregation of grains — similar entropic considerations, scaling laws and even hard-sphere simulations have been used in the two fields.

As a test case, therefore, we make alternative use of our previous observation that thin vibrated beds of grains exhibit effective randomization of their velocities by examining how segregation between different size grains changes as relative concentrations of the grains are varied. Results of such an experiment are presented in Fig. 5. Here we have vibrated a 5-mm-deep blend of large (gold) and small (maroon) glass beads under identical conditions, but at three different sets of concentrations. At high concentration of small beads (Fig. 5a), the small and large beads separate cleanly, each species favouring positions near walls. As the relative concentration of small beads is reduced, the entropic advantage of mixing grows with respect to that of separation, and only a small region (bottom right of Fig. 5b) of segregated small beads is seen. Finally, the entropy of the mixed state exceeds that of the homogeneous state, which prevails when the weight fraction of small beads in the bed drops below about 11% (Fig. 5c). We cannot rule out the possibility that the small beads are in fact segregated, perhaps forming an island hidden by the larger grains, but if such islands exist, they are not visible and they do not change the bed dynamics in any perceptible way. Coincidentally, the fraction of smaller beads at which the transition between mixing and separation appears is close to the transition concentration predicted using hard-sphere simulations applied to polymers⁶³.

This system is far from the equilibrium case considered using traditional statistical mechanics. Moreover, the microscale physics of polymer blends, viral crystal aggregates and other systems, kept under controlled temperature and chemical balance, are quite different from those of large, energetically vibrated glass beads. Nevertheless the granular system is highly randomized and *because of this fact*, the probabilistic considerations that favour macroscale patterned structures over homogeneous mixed states apply irrespective of these differences in detail.

Outlook

It has been said that any sufficiently advanced technology is indistinguishable from magic. The inseparable facts that continue to attract

researchers to the study of patterns generated from noise are that the mechanism for their formation is extremely basic, yet even once the mechanism has been thoroughly dissected and understood, its outcome still seems like magic. The paradox that the spontaneous formation of these patterns seems simultaneously to be completely implausible and hugely probable is difficult to reconcile with a rational and systematic analysis of the natural world. This situation has given rise to several unresolved, and possibly unresolvable, questions.

For example, if random processes generate similar patterns to those generated by biological systems, is there any way to interpret patterns to determine which of these causes have produced putative fossil remains? In one sense this is a religious question: are the patterns that constitute life itself possible through only random influences, or was a deterministic, intelligent intervention involved? At this level, the question may be unresolvable, but on another front, several researchers^{49,67–69} have developed concrete relations between pattern selection and underlying symmetries. What are the practical lessons that this research can teach us about patterns and the mechanisms that formed them? More philosophically, in statistical physics, profound debates have developed historically over the direction of time — which is often characterized based on the tendency of ordered states become disordered over time. These debates have so far been partially resolved by establishing that the tendency toward disorder is due to its vastly greater probability of occurrence²⁶. How shall we interpret these debates when ordered states are actually more probable?

Finally, as research becomes directed increasingly towards pressing technological issues in problems such as nanofabrication, smart materials, semiconductor design and thin-film coating, to name a few, the engineering of spontaneously assembling patterns has begun to come into its own, and the question of potential limits to the design of entropically controlled patterns arises. For example, just using the model described in Fig. 2 that was designed deliberately to generate patterns using noise and dissipation alone, we have seen that squares, stripes, hexagons and triangles can be generated. The first three of these have been observed experimentally, but so far the triangles have not. Yet they respect perfectly well defined symmetries — does this mean that there are additional, currently unknown constraints preventing the physical manifestation of certain mathematically allowed patterns? To the triangles example, one can add an enormous variety of other patterns obtained from symmetry analysis, including quasipatterns^{70,71}, superlattices^{72,73} and three-dimensional patterns^{40,74,75}. Understanding the natural limits on pattern selection remains an on-going challenge. □

- Shinbrot, T. A novel model of contact charging. *J. Electrostat.* **17**, 113–123 (1985).
- Ball, P. *The Self-made Tapestry: Pattern Formation in Nature* (Oxford Univ. Press, Oxford, 1999).
- Graham, M. D. et al. Effects of boundaries on pattern formation: catalytic oxidation of CO on platinum. *Science* **264**, 80–82 (1994).
- Vogel, G. Goodness gracious, great walls afar. *Science* **274**, 343–343 (1996).
- Toomre, A. & Toomre, J. Galactic bridges and tails. *Astrophys. J.* **178**, 623–666 (1972).
- Sellwood, J. A. Spiral structure as a recurrent instability. *Astrophys. Space Sci.* **272**, 31–43 (2000).
- Kondo, S. & Asai, R. A reaction-diffusion wave on the skin of the marine angelfish *Pomacanthus*. *Nature* **376**, 765–768 (1995).
- Haim, D. et al. Breathing spots in a reaction-diffusion system. *Phys. Rev. Lett.* **77**, 190–193 (1996).
- Pearson, J. E. Complex patterns in a simple system. *Science* **261**, 189–192 (1993).
- Kadar, S., Wang, J. & Showalter, K. Noise supported traveling waves in sub-excitable media. *Nature* **391**, 770–772 (1998).
- Adams, M., Dogic, Z., Keller, S. L. & Fraden, S. Entropically driven microphase transitions in mixtures of colloidal rods and spheres. *Nature* **393**, 349–352 (1998).
- Grotzinger, J. P. & Rothman, D. H. An abiotic model for stromatolite morphogenesis. *Nature* **383**, 423–425 (1996).
- Csahók, I., Z., Misbah, I., Rioual, F. & Valance, A. Dynamics of aeolian sand ripples. *Eur. Phys. J. E* **3**, 71–86 (2000).
- Makse, H. A. Grain segregation mechanisms in aeolian sand ripples. *Eur. Phys. J. E* **1**, 127–135 (2000).
- Lekkerkerker, H. N. W. & Stroobants, A. Ordering entropy. *Nature* **393**, 305–306 (1998).
- Reynolds, O. *On the Action of Rain to Calm the Sea. Papers on Mechanical and Physical Subjects* 86–88 (Cambridge Univ. Press, London, 1900).
- Tsimplis, M. N. The effect of rain in calming the sea. *J. Phys. Oceanogr.* **22**, 404–412 (1993).
- Herzel, H. Stabilization of chaotic orbits by random noise. *A. Angew. Math. Mech.* **68**, 11–12 (1988).
- Walker, J. *The Flying Circus of Physics* 14–8, 99–100 (Wiley, New York, 1975).
- Weisenfeld, K. & Moss, F. Stochastic resonance and the benefits of noise: from ice ages to crayfish and SQUIDS. *Nature* **373**, 33–36 (1995).
- Barnsley, M. *Fractals Everywhere* (Academic, Boston, 1988).

22. Jung, P. & Mayer-Kress, G. Spatiotemporal stochastic resonance in excitable media. *Phys. Rev. Lett.* **74**, 2130–2133 (1995).
23. Shinbrot, T. & Scarbrough, K. Using variability to regulate long term biological rhythms. *J. Theor. Biol.* **196**, 455–471 (1999).
24. Strogatz, S. H. *Nonlinear Dynamics and Chaos* 254–255 (Addison Wesley, Reading, MA, 1994).
25. Caves, C. M., Unruh, W. G. & Zurek, W. H. Comment on 'Quantitative limits on the ability of a Maxwell demon to extract work from heat'. *Phys. Rev. Lett.* **65**, 1387 (1990).
26. Brush, S. G. *Statistical Physics and the Atomic Theory of Matter* 86–90 (Princeton Univ. Press, Princeton, 1983).
27. Schlichting, H. J. & Nordmeier, V. Strukturen im Sand—Kollektives Verhalten und Selbstorganisation bei Granulaten Math. *Naturwissenschaften* **49**, 323–332 (1996).
28. Eggers, J. Sand as Maxwell's demon. *Phys. Rev. Lett.* **83**, 5322–5325 (1999).
29. Sano, O. Random motion of a marker particle on square cells formed on vertically vibrated granular layer. *J. Phys. Soc. Jpn* **68**, 1769–1777 (1999).
30. Shinbrot, T. Competition between randomizing impacts and inelastic collisions in granular pattern formation. *Nature* **389**, 574–576 (1997).
31. Bizon, C., Shattuck, M. D., Swift, J. B. & Swinney, H. L. Transport coefficients for granular media from molecular dynamics simulations. *Phys. Rev. E* **60**, 4340–4351 (1999).
32. Winfree, A. Heart muscle as a reaction-diffusion medium: the roles of electric potential, diffusion, activation front curvature, and anisotropy. *Int. J. Bif. Chaos* **7**, 487–526 (1997).
33. Winfree, A. Electrical turbulence in 3-dimensional heart muscle. *Science* **266**, 1003–1006 (1994).
34. Klausmeier, C. A. Regular and irregular patterns in semiarid vegetation. *Science* **284**, 1826–1828 (1999).
35. Lee, K. J., McCormick, W. D., Ouyang, Q. & Swinney, H. L. Pattern formation by interacting chemical fronts. *Science* **261**, 192–194 (1993).
36. Gorman, M., el-Hamdi, M. & Pearson, B. Ratcheting motion of concentric rings in cellular flames. *Phys. Rev. Lett.* **76**, 228–231 (1996).
37. Aldushin, A. P. & Matkowsky, B. J. *Diffusion Driven Combustion Waves in Porous Media* (Gordon & Breach, Newark, NJ, 1999).
38. Sellwood, J. A. & Carlberg, R. G. Spiral instabilities provoked by accretion and star formation. *Astrophys. J.* **282**, 61–74 (1984).
39. Wu, K. K. S., Lahav, O. & Rees, M. J. The large-scale smoothness of the universe. *Nature* **397**, 225–230 (1999).
40. Einasto, J. *et al.* A 120-mpc periodicity in the three-dimensional distribution of galaxy superclusters. *Nature* **385**, 139–141 (1997).
41. Goldhirsch, I. & Zanetti, G. Clustering instability in dissipative gases. *Phys. Rev. Lett.* **70**, 1619–1622 (1993).
42. Bizon, C., Shattuck, M. D., Swift, J. B., McCormick, W. D. & Swinney, H. L. Patterns in 3D vertically oscillated granular layers: simulation and experiment. *Phys. Rev. Lett.* **80**, 57–60 (1998).
43. Shinbrot, T., Lomelo L. & Muzzio, F. J. Harmonic patterns in fine granular vibrated beds. *Gran. Matt.* **2**, 65–69 (2000).
44. Umbanhowar, P. B., Melo, F. & Swinney, H. L. Localized excitations in a vertically vibrated granular layer. *Nature* **382**, 793–796 (1996).
45. Eggers, J. & Riecke, H. A continuum description of vibrated sand. *Phys. Rev. E* **59**, 4476–4483 (1999).
46. Tsimring, L. S. & Aranson, I. S. Localized and cellular patterns in a vibrated granular layer. *Phys. Rev. Lett.* **79**, 213–216 (1997).
47. Venkataramani, S. C. & Ott, E. Spatio-temporal bifurcation phenomena with temporal period doubling: patterns in vibrated sand. *Phys. Rev. Lett.* **80**, 3495–3498 (1998).
48. Metcalf, T. H., Knight, J. B. & Jaeger, H. M. Standing wave patterns in shallow beds of vibrated granular material. *Physica A* **236**, 202–210 (1997).
49. Crawford, J. D., Gollub, J. P. & Lane, D. Hidden symmetries of parametrically forced waves. *Nonlinearity* **6**, 119–164 (1993).
50. Sinai, Ya. G. Dynamical systems with elastic reflections: ergodic properties of dispersing billiards. *Russ. Math. Surv.* **25**, 137–147 (1970).
51. Ding, M.-Z., Grebogi, C., Ott, E. & Yorke, J. A. Transition to chaotic scattering. *Phys. Rev. A* **42**, 7025–7040 (1990).
52. Luding, S., Clément, E., Blumen, A., Rajchenbach, J. & Duran, J. Studies of columns of beads under external vibrations. *Phys. Rev. E* **49**, 1634–1646 (1994).
53. Aranson, I. S. *et al.* Electrostatically driven granular media: phase transitions and coarsening. *Phys. Rev. Lett.* **84**, 3306–3309 (2000).
54. Grier, D. G. A surprisingly attractive couple. *Nature* **393**, 621–622 (1998).
55. Gujrati, P. D. Entropy-driven phase separation and configurational correlations on a lattice: some rigorous results. *Phys. Rev. E* (in the press).
56. Dinsmore, A. D., Yodh, A. G. & Pine, D. J. Entropic control of particle motion using passive surface microstructures. *Nature* **383**, 239–242 (1996).
57. Crocker, J. C. & Grier, D. G. When like charges attract: the effects of geometrical confinement on long-range colloidal interactions. *Phys. Rev. Lett.* **77**, 1897–1900 (1996).
58. Crocker, J. C., Matteo, J. A., Dinsmore, A. D. & Yodh, A. G. Entropic attraction and repulsion in binary colloids probed with a line optical tweezer. *Phys. Rev. Lett.* **82**, 4352–4355 (1999).
59. Duran, J. & Jullien, R. Attractive forces in a granular cocktail. *Phys. Rev. Lett.* **80**, 3547–3550 (1998).
60. van Roij, R., Mulder, B. & Dijkstra, M. Phase behavior of binary mixtures of thick and thin hard rods. *Physica A* **261**, 374–390 (1998).
61. Dijkstra, M. & Frenkel, D. Evidence for entropy-driven demixing in hard-core fluids. *Phys. Rev. Lett.* **72**, 298–300 (1994).
62. Dijkstra, M. & van Roij, R. Entropy-driven demixing in binary hard-core mixtures: from hard spherocylinders towards hard spheres. *Phys. Rev. E* **56**, 5594–5602 (1997).
63. Dijkstra, M., van Roij, R. & Evans, R. Phase diagram of highly asymmetric binary hard-sphere mixtures. *Phys. Rev. E* **59**, 5744–5771 (1999).
64. König, A. & Ashcroft, N. W. Structure and effective interactions three-component hard sphere liquids. *Phys. Rev. E* (in the press).
65. Hong, D., Quinn, P. V. & Luding, S. Reverse Brazilian nut problem: competition between percolation and condensation. *Phys. Rev. Lett.* (in the press); preprint cond-mat/0010459 also available at <http://xxx.lanl.gov>.
66. Shinbrot, T. & Muzzio, F. J. Nonequilibrium patterns in granular mixing and segregation. *Phys. Today* 25–30 (March 2000).
67. Stewart, I. & Golubitsky, M. *Fearful Symmetry: Is God a Geometer?* (Penguin, London, 1992).
68. Silber, M. & Skeldon, A. C. Parametrically excited surface waves: two-frequency forcing, normal form symmetries, and pattern selection. *Phys. Rev. E* **59**, 5446–5456 (1999).
69. Bowden, N., Terfort, A., Carbeck, J. & Whitesides, G. M. Self-assembly of mesoscale objects into ordered two-dimensional arrays. *Science* **276**, 233–236 (1997).
70. Silber, M., Topaz, C. M. & Skeldon, A. C. Two-frequency forced Faraday waves: weakly damped modes and pattern selection. *Physica D* **143**, 205–225 (2000).
71. Edwards, W. S. & Fauve, S. Patterns and quasi-patterns in the Faraday experiment. *J. Fluid Mech.* **278**, 123–148 (1994).
72. Kudrolli, A., Pier, B. & Gollub, J. P. Superlattice patterns in surface waves. *Physica D* **22**, 99–111 (1998).
73. Silber, M. & Proctor, M. R. E. Nonlinear competition between small and large hexagonal patterns. *Phys. Rev. Lett.* **81**, 2450–2453 (1998).
74. van Blaaderen, A., Ruel, R. & Wiltzius, P. Template-directed colloidal crystallization. *Nature* **385**, 321–323 (1997).
75. Prause, B., Glazier, J. A., Gravina, S. & Montemagno, C. Magnetic resonance imaging of a three dimensional foam. *J. Phys.* **7**, L511–L516 (1995).
76. Bagnold, R. A. *The Physics of Blown Sand and Desert Dunes* (Methuen, London, 1941).
77. Forrest, S. B. & Haff, P. K. Mechanics of wind ripple stratigraphy. *Science* **255**, 1240–1243 (1992).
78. Heaney, P. J. & Davis, A. M. Observation and origin of self-organized textures in agates. *Science* **269**, 1562–1565 (1995).
79. Hosoi, A. E. & Dupont, T. F. Layer formation in monodisperse suspensions and colloids. *J. Fluid Mech.* **328**, 297–311 (1996).
80. Mueth, D. M., Crocker, J. C., Esipov, S. E. & Grier, D. G. Origin of stratification in creaming emulsions. *Phys. Rev. Lett.* **77**, 578–581 (1996).
81. Mullin, T. Coarsening of self-organized clusters in binary mixtures of particles. *Phys. Rev. Lett.* **84**, 4741–4744 (2000).
82. Oliver, S., Kuperman, A., Coombs, N., Lough, A. & Ozin, G. A. Lamellar aluminophosphates with surface patterns that mimic diatom and radiolarian microstructures. *Nature* **378**, 47–50 (1995).
83. Dogic, Z. & Fraden, S. Smectic phase in a colloidal suspension of semiflexible virus particles. *Phys. Rev. Lett.* **78**, 2417–2420 (1997).
84. Kestenbaum, D. Gentle force of entropy bridges disciplines. *Science* **279**, 1849–1849 (1998).

Acknowledgements

We acknowledge helpful remarks and vital contributions from M. Dijkstra, Z. Dojic, S. Fraden, P. Garik, R. van Roij, H. Swinney and P. Umbanhowar.

Supercooled liquids and the glass transition

Pablo G. Debenedetti* & Frank H. Stillinger†‡

*Department of Chemical Engineering and ‡Princeton Materials Institute, Princeton University, Princeton, New Jersey 08544, USA (e-mail: pdebene@princeton.edu)

†Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA

Glasses are disordered materials that lack the periodicity of crystals but behave mechanically like solids. The most common way of making a glass is by cooling a viscous liquid fast enough to avoid crystallization. Although this route to the vitreous state — supercooling — has been known for millennia, the molecular processes by which liquids acquire amorphous rigidity upon cooling are not fully understood. Here we discuss current theoretical knowledge of the manner in which intermolecular forces give rise to complex behaviour in supercooled liquids and glasses. An intriguing aspect of this behaviour is the apparent connection between dynamics and thermodynamics. The multidimensional potential energy surface as a function of particle coordinates (the energy landscape) offers a convenient viewpoint for the analysis and interpretation of supercooling and glass-formation phenomena. That much of this analysis is at present largely qualitative reflects the fact that precise computations of how viscous liquids sample their landscape have become possible only recently.

The glassy state is ubiquitous in nature and technology¹. It is crucial in the processing of foods², the commercial stabilization of labile biochemicals³, and the preservation of insect life under extremes of cold or dehydration³.

Window glass, composed mostly of sand, lime and soda, is the best-known example of an engineered amorphous solid⁴. Optical fibres are made of very pure amorphous silica, occasionally carefully doped. Most engineering plastics are amorphous solids, as are some metallic glasses and alloys of interest because of their soft magnetism and corrosion resistance⁵. The silicon used in many photovoltaic cells is amorphous, and it is possible that most water in the Universe may be glassy⁶. Most of these examples entail supercooling of a liquid to take advantage of viscous retardation of nucleation and crystallization. Understanding quantitatively the extraordinary viscous slow-down that accompanies supercooling and glass formation is a major scientific challenge⁷.

We begin by reviewing the phenomenology of vitrification and supercooling. A useful approach for relating this complex phenomenology to molecular-scale events is to focus attention on the liquid's energy landscape, that is, the multidimensional surface generated by the system's potential energy as a function of molecular coordinates. Accordingly, basic landscape concepts and a discussion of the important theoretical and computational progress currently being made in this area are presented next. This is followed by a discussion of alternative viewpoints, in which narrowly avoided singularities are assumed to occur well above the glass-transition temperature. We then close with a summary of the important open questions.

It is impossible to do justice to the entire field of supercooled liquids and amorphous solids in an article of this length. We have therefore limited the scope to the dynamics and thermodynamics of viscous liquids above and close to the glass-transition temperature T_g — in other words, to the glass transition viewed 'from the liquid'. The view 'from the solid', including such topics as relaxation both relatively near

(for example, during annealing or ageing) and far below T_g , is not discussed. The reader is referred to an excellent recent review⁸ for a thorough coverage of these and other topics.

Phenomenology of supercooling and glass formation

Figure 1 illustrates the temperature dependence of a liquid's volume (or enthalpy) at constant pressure^{4,9}. Upon cooling below the freezing point T_m , molecular motion slows down. If the liquid is cooled sufficiently fast, crystallization can be avoided^{10,11}. Eventually molecules will rearrange so slowly that they cannot adequately sample configurations in the available time allowed by the cooling rate. The liquid's structure therefore appears 'frozen' on the laboratory timescale (for example, minutes). This falling out of equilibrium occurs across a narrow transformation range where the characteristic molecular relaxation time becomes of the order of 100 seconds, and the rate of change of volume or enthalpy with respect to temperature decreases abruptly (but continuously) to a value comparable to that of a crystalline solid. The resulting material is a glass. The intersection of the liquid and vitreous portions of the volume versus temperature curve provides one definition of T_g , which usually occurs around $2T_m/3$. The behaviour depicted in Fig. 1 is not a true phase transition, as it does not involve discontinuous changes in any physical property.

The slower a liquid is cooled, the longer the time available for configurational sampling at each temperature, and hence the colder it can become before falling out of liquid-state equilibrium. Consequently, T_g increases with cooling rate^{12,13}. The properties of a glass, therefore, depend on the process by which it is formed. In practice, the dependence of T_g on the cooling rate is weak (T_g changes by 3–5 °C when the cooling rate changes by an order of magnitude¹⁴), and the transformation range is narrow, so that T_g is an important material characteristic.

Slowing down

Another definition of T_g is the temperature at which the shear viscosity reaches 10^{13} poise. Close to T_g the viscosity η

is extraordinarily sensitive to temperature. For silica this dependence is reasonably well described by the Arrhenius functionality, $\eta = A \exp(E/k_B T)$, where A and E are temperature-independent and k_B is Boltzmann's constant. Other liquids exhibit an even more pronounced viscous slow-down close to the glass transition, which is reasonably well represented, over 2–4 orders of magnitude in viscosity⁸, by the Vogel–Tammann–Fulcher (VTF) equation^{15–17}

$$\eta = A \exp[B/(T - T_0)] \quad (1)$$

where A and B are temperature-independent constants. Understanding the origin of this extraordinary slow-down of relaxation processes is one of the main challenges in the physics of glasses.

Figure 2 shows a T_g -scaled Arrhenius representation of liquid viscosities^{18–21}. Angell has proposed a useful classification of liquids

along a 'strong' to 'fragile' scale. The viscosity and relaxation times (for example, dielectric relaxation) of the former behave in nearly Arrhenius fashion, whereas fragile liquids show marked deviations from Arrhenius behaviour. Silica (SiO_2) is the prototypical strong liquid, whereas *o*-terphenyl (OTP) is the canonical fragile glass-former. Strong liquids, such as the network oxides SiO_2 and germanium dioxide (GeO_2), have tetrahedrally coordinated structures, whereas the molecules of fragile liquids exert largely non-directional, dispersive forces on each other. Alternative scaling descriptions that attempt to extract universal aspects of viscous slow-down have been proposed^{22–24}. Their relative merits are still being assessed^{8,25}.

Viscous liquids close to T_g exhibit non-exponential relaxation. The temporal behaviour of the response function $F(t)$ (for example, the polarization in response to an applied electric field, the strain (deformation) resulting from an applied stress, or the stress in

Box 1

Entropy crises

Boltzmann's entropy formula establishes the connection between the microscopic world of atoms and molecules and the bulk properties of matter:

$$S(N, V, E) = k_B \ln \Omega$$

In this equation, S is the entropy, k_B is Boltzmann's constant and Ω is the number of quantum states accessible to N particles with fixed energy E in a volume V . Because Ω cannot be less than one, the entropy cannot be negative. When a crystal is cooled sufficiently slowly, it approaches a unique state of lowest energy, and hence its entropy approaches zero as $T \rightarrow 0$. If the entropy of a supercooled liquid were to become smaller than that of the stable crystal at the Kauzmann temperature, its entropy would eventually become negative upon further cooling. This impossible scenario constitutes an entropy crisis^{46–48}.

The Kauzmann temperature T_K is given by⁹

$$\Delta s_m = \int_{T_K}^{T_m} \frac{\Delta C_p}{T} dT$$

where Δs_m is the melting entropy (the difference between liquid and crystal entropies at the melting temperature), T_m is the melting temperature at the given pressure, and ΔC_p is the temperature-dependent difference between the heat capacity of the liquid and the crystal at the given pressure. The rate of change of entropy with temperature at constant pressure is given by

$$\left(\frac{\partial S}{\partial T}\right)_P = \frac{C_p}{T}$$

The entropy crisis arises because the heat capacity of a liquid is greater than that of the stable crystal. The entropy of fusion is therefore consumed upon supercooling, and vanishes at T_K . The entropy crisis entails no conflict with the second law of thermodynamics, as the difference in chemical potential $\Delta\mu$ between the supercooled liquid and the stable crystal at T_K is a positive quantity. Because the chemical potential is the Gibbs free energy per unit mass, this means that the system can reduce its Gibbs free energy by freezing, in accord with experience. The chemical potential difference at T_K is given by⁹

$$\Delta\mu(T_K) = \int_{T_K}^{T_m} \Delta C_p \left(\frac{T_m}{T} - 1\right) dT$$

One way of avoiding the entropy crisis is for the liquid to form an ideal glass of unique configuration at T_K . This is the thermodynamic view of the glass transition, according to which the observable glass transition is a manifestation, masked by kinetics, of an underlying second-order phase transition⁵⁰ occurring at T_K .

Because the glass transition intervenes before the entropy crisis occurs ($T_g > T_K$), estimates of the Kauzmann temperature involve an extrapolation of liquid properties below T_g . The validity of such extrapolations, and hence of the very possibility of an entropy crisis, has been questioned by Stillinger¹⁰³, owing to the apparent necessity for configurational excitations out of the ideal glass state to require an unbounded energy. Furthermore, recent computer simulations of polydisperse hard disks found no evidence of a thermodynamic basis underlying the glass transition¹⁰⁴. Although the notion of an ideal glass remains controversial¹⁰³, this does not undermine the usefulness of T_K as an empirical classification parameter for glass-formers.

Experimentally, there are substances with known Kauzmann points. ⁴He is a liquid at 0 K and 1 bar (liquid He-II). Its equilibrium freezing pressure at 0 K is 26 bar. At this point, the entropies of the liquid and the crystal are equal, and this is therefore a Kauzmann point, although not an entropy crisis as both phases have zero entropy. The melting curves of ³He and ⁴He exhibit pressure minima: these occur at about 0.32 K and 0.8 K, respectively¹⁰⁵. These are also equal-entropy (Kauzmann) points. Experiments indicate that poly(4-methylpentene-1) exhibits a pressure maximum along its melting curve¹⁰⁶. Although the appearance of an additional phase complicates the interpretation of this observation, the implication would be that the pressure maximum is a Kauzmann point, and the continuation of the melting curve to lower temperatures and pressures beyond this point corresponds to endothermic freezing of a stable liquid into a crystal possessing higher entropy¹⁰⁷. How this liquid would avoid conflict with the third law is not understood, but may hinge on the vibrational anharmonicity of the two phases with changing temperature.

response to an imposed deformation) can often be described by the stretched exponential, or Kohlrausch–Williams–Watts (KWW) function^{26,27}

$$F(t) = \exp[-(t/\tau)^\beta] \quad (\beta < 1) \quad (2)$$

where $F(t) = [\sigma(t) - \sigma(\infty)]/[\sigma(0) - \sigma(\infty)]$ and σ is the measured quantity (for example, the instantaneous stress following a step change in deformation). τ in equation (2) is a characteristic relaxation time, whose temperature dependence is often non-Arrhenius (exhibiting fragile behaviour). The slowing down of long-time relaxation embodied in equation (2) contrasts with the behaviour of liquids above the melting point, which is characterized by simple exponential relaxation. Experimental and computational evidence indicates that this slow-down is related to the growth of distinct relaxing domains^{28–39} (spatial heterogeneity). Whether each of these spatially heterogeneous domains relaxes exponentially or not is a matter of considerable current interest^{38,39}.

Decouplings

In supercooled liquids below approximately $1.2T_g$ there occurs a decoupling between translational diffusion and viscosity, and between rotational and translational diffusion^{30,39,40}. At higher temperatures, both the translational and the rotational diffusion coefficients are inversely proportional to the viscosity, in agreement with the Stokes–Einstein and Debye equations, respectively. Below approximately $1.2T_g$, the inverse relationship between translational motion and viscosity breaks down, whereas that between rotational motion and viscosity does not. Near T_g , it is found that molecules translate faster than expected based on their viscosity, by as much as two orders of magnitude. This therefore means that, as the temperature is lowered, molecules on average translate progressively more for every rotation they execute. Yet another decoupling occurs in the moderately supercooled range. At sufficiently high temperature the liquid shows a single peak relaxation frequency (Fig. 3), indicative of one relaxation mechanism. In the moderately supercooled regime, however, the peak splits into slow (α) and fast (β) relaxations^{41–43}. The former exhibit non-Arrhenius behaviour and disappear at T_g ; the latter continue below T_g and display Arrhenius behaviour⁴⁴.

Thermodynamics

The entropy of a liquid at its melting temperature is higher than that of the corresponding crystal. Because the heat capacity of a liquid is higher than that of the crystal, this entropy difference decreases upon supercooling (Box 1). Figure 4 shows the temperature dependence of the entropy difference between several supercooled liquids and their stable crystals⁴⁵. For lactic acid this entropic surplus is consumed so

fast that a modest extrapolation of experimental data predicts its impending vanishing. In practice, the glass transition intervenes, and ΔS does not vanish. If the glass transition did not intervene, the liquid entropy would equal the crystal's entropy at a nonzero temperature T_K (the Kauzmann temperature.) Because the entropy of the crystal approaches zero as T tends to zero, the entropy of the liquid would eventually become negative upon cooling if this trend were to continue. Because entropy is an inherently non-negative quantity (Box 1), the state of affairs to which liquids such as lactic acid are tending when the glass transition intervenes is an entropy crisis^{46–48}. The extrapolation needed to provoke conflict with the third law is quite modest for many fragile liquids⁴⁹, and the imminent crisis is thwarted by a kinetic phenomenon, the glass transition. This suggests a connection between the kinetics and the thermodynamics of glasses⁴⁷. The thermodynamic viewpoint that emerges from this analysis⁵⁰ considers the laboratory glass transition as a kinetically controlled manifestation of an underlying thermodynamic transition to an ideal glass with a unique configuration.

A formula of Adam and Gibbs⁵¹ provides a suggestive connection between kinetics and thermodynamics:

$$t = A \exp(B/Ts_c) \quad (3)$$

In this equation, t is a relaxation time (or, equivalently, the viscosity) and A and B are constants. s_c , the configurational entropy, is related to the number of minima of the system's multidimensional potential energy surface (Box 2). According to the Adam–Gibbs picture, the origin of viscous slow-down close to T_g is the decrease in the number of configurations that the system is able to sample. At the Kauzmann temperature the liquid would have attained a unique, non-crystalline state of lowest energy, the ideal glass. Because there is no configurational entropy associated with confinement in such a state, the Adam–Gibbs theory predicts structural arrest to occur at T_K . In their derivation of equation (3), Adam and Gibbs invoked the concept of a cooperatively rearranging region (CRR)⁵¹. A weakness of their treatment is the fact that it provides no information on the size of such regions. The fact that the CRRs are indistinguishable from each other is also problematic, in light of the heterogeneity that is believed to underlie stretched exponential behaviour⁸.

Figure 1 Temperature dependence of a liquid's volume v or enthalpy h at constant pressure. T_m is the melting temperature.

A slow cooling rate produces a glass transition at T_{ga} ; a faster cooling rate leads to a glass transition at T_{gb} . The thermal expansion coefficient $\alpha_p = (\partial \ln v / \partial T)_p$ and the isobaric heat capacity $c_p = (\partial h / \partial T)_p$ change abruptly but continuously at T_g .

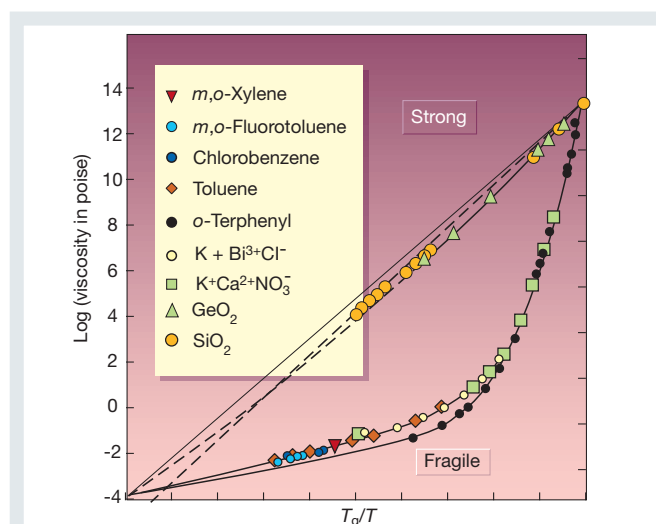
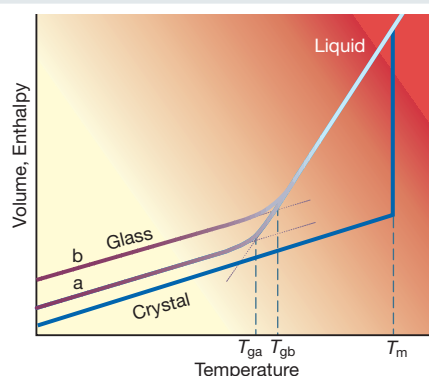


Figure 2 T_g -scaled Arrhenius representation of liquid viscosities showing Angell's strong–fragile pattern. Strong liquids exhibit approximate linearity (Arrhenius behaviour), indicative of a temperature-independent activation energy $E = \ln \eta / d(1/T) \approx \text{const}$. Fragile liquids exhibit super-Arrhenius behaviour, their effective activation energy increasing as temperature decreases. (Adapted from refs 9 and 11.)

Nevertheless, equation (3) describes the relaxation behaviour of deeply supercooled liquids remarkably well. If the difference in heat capacities between a supercooled liquid and its stable crystalline form is inversely proportional to temperature⁵², the Adam–Gibbs

relation yields the VTF equation, which is mathematically equivalent to the Williams–Landel–Ferry equation for the temperature dependence of viscosity in polymers⁵³. This transformation is predicated on the assumption that the vibrational entropies of the

Box 2

Statistics of landscapes

The complexity of many-body landscapes makes a statistical description inevitable. The quantity of interest is the number of minima of given depth, which is given by¹⁰⁸

$$\frac{d\Omega}{d\phi} = C \exp[N\sigma(\phi)]$$

Here, $d\Omega$ denotes the number of potential energy minima with depth per particle ($\phi = \Phi/N$) between ϕ and $\phi \pm d\phi/2$. C is an N -independent factor with units of inverse energy, and $\sigma(\phi)$, also an N -independent quantity, is a so-called basin enumeration function. Taking the logarithm of the above expression and comparing with Boltzmann's entropy formula (Box 1), we see that $\sigma(\phi)$ is the entropy per particle arising from the existence of multiple minima of depth ϕ , or, in other words, the configurational entropy.

At low temperatures, it is possible to separate the configurational contribution to thermophysical properties, which arises from the exploration of different basins, from the vibrational component, which arises from thermal motions confined to a given basin^{75,76}. The Helmholtz free energy A is then given by

$$\frac{A}{NkT} = \frac{\bar{\phi}}{kT} - \sigma(\bar{\phi}) + \frac{a^v}{k_B T}$$

where $\bar{\phi}$ is the depth of the basins preferentially sampled at the given temperature, and a^v is the vibrational free energy per particle. Thus, the free energy consists of an energetic component that reflects the depth of landscape basins sampled preferentially at the given temperature, an entropic component that accounts for the number of existing basins of a given depth, and a vibrational component. The statistical description of a landscape consists of the basin enumeration function $\sigma(\phi)$, from which the excitation profile $\phi(T)$ is obtained through the free-energy minimization condition

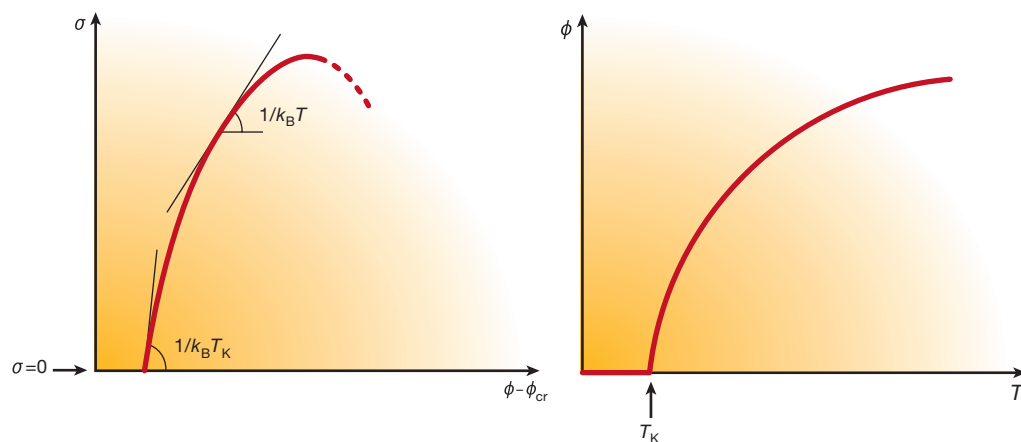
$$\frac{d\sigma}{d\phi} = \frac{1}{k_B T}$$

The above equation assumes that a^v depends on T , but not on ϕ — that is, all basins have the same mean curvature at their respective minima.

The shape of a given system's landscape is determined by the density (number of molecules per unit volume, N/V). Temperature governs the manner in which the landscape is sampled. A different basin enumeration function and excitation profile corresponds to each density. Temperature dictates the point along the enumeration curve and the excitation profile sampled by the system at fixed density (see figure below).

It is possible to construct the basin enumeration function and excitation profile of a system from experimental heat capacity data for the crystal and the supercooled liquid¹⁰⁹, and by computer simulation^{77,78}. In the latter case, the calculations involve determining the probability distribution of inherent structure energies sampled as a function of temperature. These calculations are at the limit of what is presently feasible with available computational power. The enumeration function is often well represented by a parabola, indicative of a gaussian distribution of basins^{4,77,97}. At present it is not understood how the enumeration function deforms with density for a given system (but see ref. 96 for a recent example of such a calculation), or how it depends on molecular architecture. Understanding such questions would provide a direct link between landscape statistics and physical properties. The success of the Adam–Gibbs equation indicates that this link applies also to transport properties such as diffusion and viscosity.

Box 2 Figure Schematic representation of the basin enumeration function (left) and the excitation profile (right). ϕ is the potential energy per particle in mechanically stable potential energy minima. ϕ_{cr} is the corresponding quantity in the stable crystal. The number of potential energy minima with depth between ϕ and $\phi \pm d\phi/2$ is proportional to $\exp[N\sigma(\phi)]$. In the thermodynamic limit (N of the order of Avogadro's number, 6.02×10^{23}), basins that possess larger (less negative) potential energies (shallow basins) are overwhelmingly more numerous than deeper basins possessing very negative ϕ -values. The slope of the enumeration function is inversely proportional to the temperature. The excitation profile gives the depth of the inherent structures sampled preferentially at a given temperature. At the Kauzmann temperature T_K the system attains the state of a configurationally unique ideal glass ($\sigma=0$), corresponding to the deepest amorphous basin (see Figs 5 and 8) and its inherent structure energy does not therefore change upon further cooling.



supercooled liquid and its stable crystal are equal⁹. For many fragile glass-formers the VTF temperature of structural arrest, T_o , is very close to T_K obtained from calorimetric measurements (typically⁴⁹ $0.9 < T_K/T_o < 1.1$). This again indicates a connection between dynamics and thermodynamics not present at higher temperatures. Equally suggestive is the correspondence between kinetic fragilities based on the temperature dependence of the viscosity (see Fig. 2) and thermodynamic fragilities⁵⁴, based on the temperature dependence of the entropy surplus of the supercooled liquid with respect to its stable crystal.

The energy landscape

A convenient framework for interpreting the complex phenomenology just described is provided by the energy landscape⁴⁴. This is the name generally given to the potential energy function of an N -body system $\Phi(r_1, \dots, r_N)$, where the vectors r_i comprise position, orientation and vibration coordinates. In condensed phases, whether liquid or solid, every molecule experiences simultaneous interactions with numerous neighbours. Under these conditions it is convenient to consider the full N -body Φ -function. The landscape is a multidimensional surface. For the simplest case of N structureless particles possessing no internal orientational and vibrational degrees of freedom, the landscape is a $(3N + 1)$ -dimensional object. Figure 5 is a schematic illustration of an energy landscape. The quantities of interest are the number of potential energy minima (also called inherent structures) of a given depth (Box 2), and the nature of the saddle points separating neighbouring minima. More than 30 years ago, Goldstein articulated a topographic viewpoint of condensed phases⁵⁵ that has come to be known as the energy landscape paradigm. His seminal ideas have since been applied to protein folding^{56–64}, the mechanical properties of glasses^{65–67}, shear-enhanced diffusion⁶⁸ and the dynamics of supercooled liquids^{69–71}.

Landscape sampling

For an N -body material system in a volume V , the landscape is fixed. The manner in which a material system samples its landscape as a

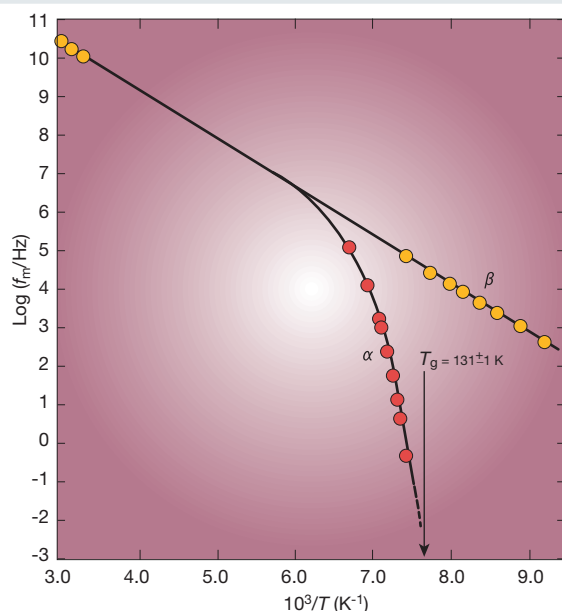


Figure 3 Temperature dependence of the peak dielectric relaxation frequency of the glass-forming mixture chlorobenzene/*cis*-decalin (molar ratio 17.2/82.8%). At high enough temperature there is a single relaxation mechanism. In the moderately supercooled regime the peak splits into slow (α) and fast (β) relaxations, of which α -processes exhibit non-Arrhenius temperature dependence and vanish at T_g . (Adapted from refs 9 and 41.)

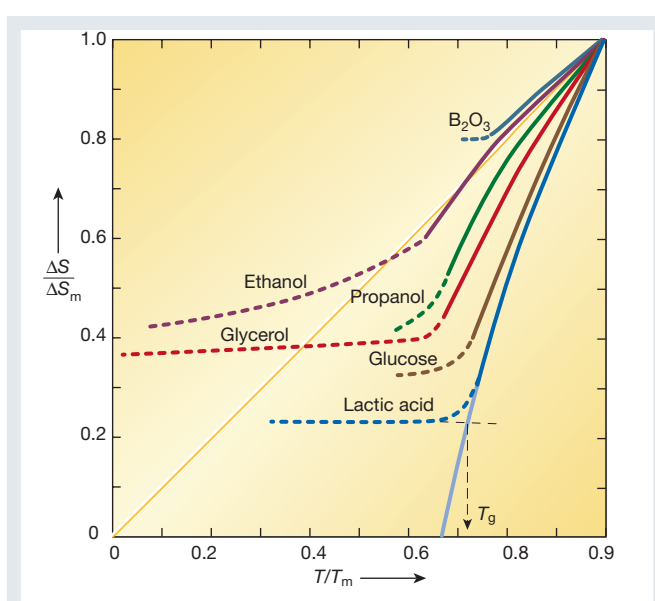


Figure 4 Temperature dependence of the entropy difference between several supercooled liquids and their stable crystals at atmospheric pressure. ΔS_m is the melting entropy and T_m is the melting temperature. The glass transition always intervenes before the vanishing of the entropy surplus. For fragile liquids such as lactic acid, however, an entropy crisis is imminent. (Adapted from ref. 45.)

function of temperature provides information on its dynamic behaviour⁷⁰. The way that a landscape deforms as a result of changes in density provides information on the mechanical properties of a material system⁷². Figure 6 shows the average inherent structure energy for a mixture of unequal-sized atoms, as a function of the temperature of the equilibrated liquid^{70,73}. In these calculations, molecular dynamics simulations of the binary mixture were performed to generate configurations. Periodically, the system's energy was minimized, yielding mechanically stable inherent structures, the average energy of which is reported in the figure. At high temperatures the inherent structure energy is virtually temperature-independent, and appears to have reached a plateau. When the system has sufficient kinetic energy to sample its entire energy landscape, the overwhelming number of minima that it samples are shallow, reflecting the fact that deep minima are very rare (Box 2). But as the reduced temperature decreases below about $T = 1$, the system is unable to surmount the highest energy barriers, and is therefore forced to sample the much rarer deeper minima (Box 2). When this happens, the kinetics of structural relaxation changes from exponential to stretched exponential, and the activation energy (and entropy) associated with structural relaxation become super-Arrhenius, that is to say they increase with decreasing temperature⁷⁰.

These calculations established a connection between changes in dynamics and the manner in which the static and thermodynamic energy landscape is sampled as a function of temperature. Figure 6 also shows that at a low enough temperature the system becomes stuck in a single minimum, the depth of which increases as the cooling rate decreases. This corresponds to the glass transition. Another important observation of this study was the existence of a temperature $T \approx 0.45$, below which the height of the barriers separating sampled inherent structures increases abruptly. This temperature was found to correspond closely to the crossover temperature predicted by mode-coupling theory (MCT; see below) for this system. Here again, it is the manner in which the system samples its landscape, not the landscape itself, that changes with temperature. (See ref. 74 for a recent, different interpretation of landscape sampling at this temperature.)

The landscape picture provides a natural separation of low-temperature molecular motion into sampling distinct potential

energy minima, and vibration within a minimum. It is possible to separate formally the corresponding configurational and vibrational contributions to a liquid's properties^{75,76}. In two important computational studies, the configurational entropy was calculated by probing systematically the statistics governing the sampling of potential energy minima^{77,78} (Box 2). Using this technique, a remarkable connection between configurational entropy and diffusion was identified in liquid water⁷⁹. One of water's distinguishing anomalies is the fact that, at sufficiently low temperature, its diffusivity increases upon compression⁸⁰. As shown in Fig. 7, diffusivity maxima are correlated strongly with configurational entropy maxima, the respective loci coinciding within numerical error.

The results shown in Fig. 7 and the success of the Adam–Gibbs equation in describing experimental data on relaxation in a wide variety of systems⁵² indicate that there exists a scaling relationship between the depth distribution of basins and the height of the saddle points along paths connecting neighbouring basins. Such scaling is not a mathematical necessity, but arises from the nature of real molecular interactions. The topographic nature of this statistical scaling relationship between minima and saddle points is poorly understood (but see the recent computational investigation of saddle points⁷⁴). Its elucidation will explain the origin of the connection between the dynamics and thermodynamics of glass-forming liquids, and constitutes the principal theoretical challenge in this field.

Strong versus fragile behaviour

The extent to which the shear viscosity η deviates from Arrhenius behaviour, $\eta = \eta_0 \exp(E/k_B T)$, constitutes the basis of the classification of liquids as either strong or fragile (Fig. 2). Molten SiO₂, often considered as the prototypical strong glass-former, displays an almost constant activation energy of 180 kcal mol⁻¹ (ref. 81). This constancy indicates that the underlying mechanism, presumably breaking and reformation of Si–O bonds, applies throughout the entire landscape⁴. In contrast, the viscosity of OTP — the canonical fragile glass-former — deviates markedly from Arrhenius behaviour⁸², showing an effective activation energy ($d \ln \eta / d(1/T)$) that increases 20-fold, from one-quarter of the heat of vaporization for the liquid above its melting point to roughly five times the heat of vaporization near T_g . This means that OTP's landscape is very heterogeneous. The basins sampled at high temperature allow relaxation by surmounting low barriers involving the rearrangement of a small number of molecules. The very large activation energy at $T \approx T_g$, on the other hand, corresponds to the cooperative rearrangement of many molecules. These differences between strong and fragile behaviour imply a corresponding topographic distinction between the two

archetypal landscapes. Aside from multiplicity due to permutational symmetry, strong landscapes may consist of a single 'megabasin', whereas fragile ones display a proliferation of well-separated 'megabasins' (Fig. 8).

Cooperative rearrangements such as those that must occur in OTP are unlikely to consist of elementary transitions between adjacent basins. Rather, the likely scenario involves a complicated sequence of elementary transitions. At low temperatures, these rearrangements should be rare and long-lived on the molecular timescale. Furthermore, the diversity of deep landscape traps and of the pathways of configuration space that connect them should result in a broad spectrum of relaxation times, as required for the stretched exponential function in equation (2). This in turn suggests that supercooled fragile liquids are dynamically heterogeneous, probably consisting at any instant of mostly non-diffusing molecules with a few 'hot spots' of mobile molecules. This dynamic heterogeneity³⁹ has both experimental^{29,30,36} and computational^{31–35} support.

The inverse relation between the self-diffusion coefficient and viscosity embodied in the Stokes–Einstein equation is based on macroscopic hydrodynamics that treats the liquid as a continuum.

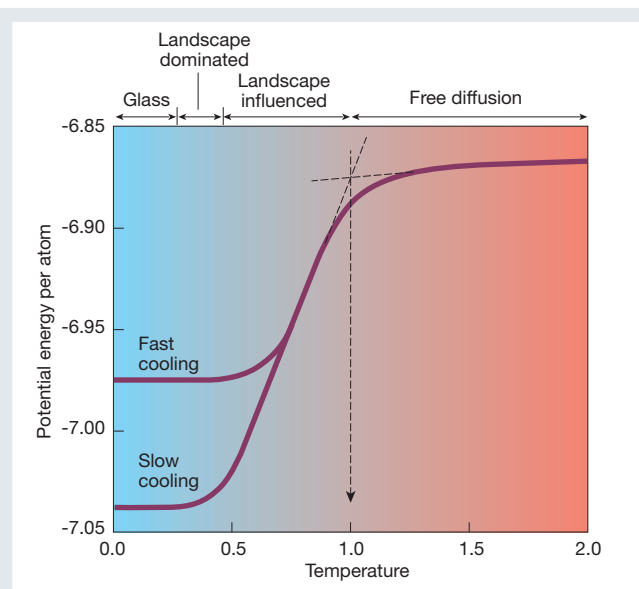


Figure 6 Mean inherent structure energy per particle of a binary mixture of unequal-sized Lennard–Jones atoms, as a function of the temperature of the equilibrated liquid from which the inherent structures were generated by energy minimization. Molecular dynamics simulations at constant energy and density were performed over a range of temperatures for 256 Lennard–Jones atoms, of which 20% are of type A and 80% are of type B. The Lennard–Jones size and energy parameters are $\sigma_{AA} = 1$, $\sigma_{BB} = 0.88$, $\sigma_{AB} = 0.8$, and $\epsilon_{AA} = 1$, $\epsilon_{BB} = 0.5$, $\epsilon_{AB} = 1.5$, respectively. Length, temperature, energy and time are expressed in units of σ_{AA} , ϵ_{AA}/k_B , ϵ_{AA} and $\sigma_{AA}(m/\epsilon_{AA})^{1/2}$, respectively, with m representing the mass of the particles. Simulations were performed at a density of 1.2. The fast and slow cooling rates are 1.08×10^{-3} and 3.33×10^{-6} . When $T > 1$, the system has sufficient kinetic energy to sample the entire energy landscape, and the overwhelming number of sampled energy minima are shallow. Under these conditions, the system exhibits a temperature-independent activation energy for structural relaxation (calculations not shown). Between $T = 1$ and $T \approx 0.45$, the activation energy increases upon cooling, the dynamics become 'landscape-influenced', and the mechanically stable configurations sampled are strongly temperature-dependent. Below $T \approx 0.45$, the height of the barriers separating sampled adjacent energy minima seems to increase abruptly (calculations not shown). This is the 'landscape-dominated' regime. In it, particles execute rare jumps over distances roughly equal to interparticle separations. The crossover between landscape-influenced and landscape-dominated behaviour corresponds closely with the mode-coupling transition temperature^{70,82}. (Adapted from refs 70 and 72.)

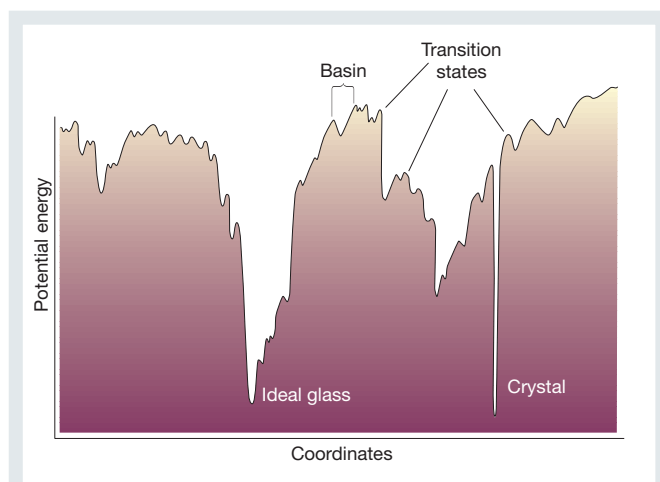


Figure 5 Schematic illustration of an energy landscape. The x-axis represents all configurational coordinates. (Adapted from ref. 44.)

This picture must clearly break down in supercooled fragile liquids, which are dynamically heterogeneous. The failure of the Stokes–Einstein equation, referred to above as one of the distinguishing characteristics of fragile supercooled liquids, is therefore qualitatively understandable. Plausible models for the low-temperature enhancement of diffusive motion relative to hydrodynamic expectations based on the viscosity have been proposed^{83–85}, but an accurate predictive theory is missing. The landscape viewpoint also provides a plausible interpretation for the α/β -relaxation decoupling shown in Fig. 3 — α -relaxations correspond to configurational sampling of neighbouring megabins (Fig. 8), whereas β -processes are thought to correspond to elementary relaxations between contiguous basins⁴⁴. Direct computational evidence of this interpretation is not available.

Avoided singularities

Alternative viewpoints to the landscape perspective have also contributed to current understanding of some aspects of supercooling and the glass transition. Two such interpretations invoke a narrowly avoided singularity above T_g .

According to MCT⁸⁶, structural arrest occurs as a result of the following feedback mechanism: (i) shear-stress relaxation occurs primarily through diffusive motion; (ii) diffusion and viscosity are inversely related; and (iii) viscosity is proportional to shear-stress relaxation time. These facts lead to a viscosity feedback whereby

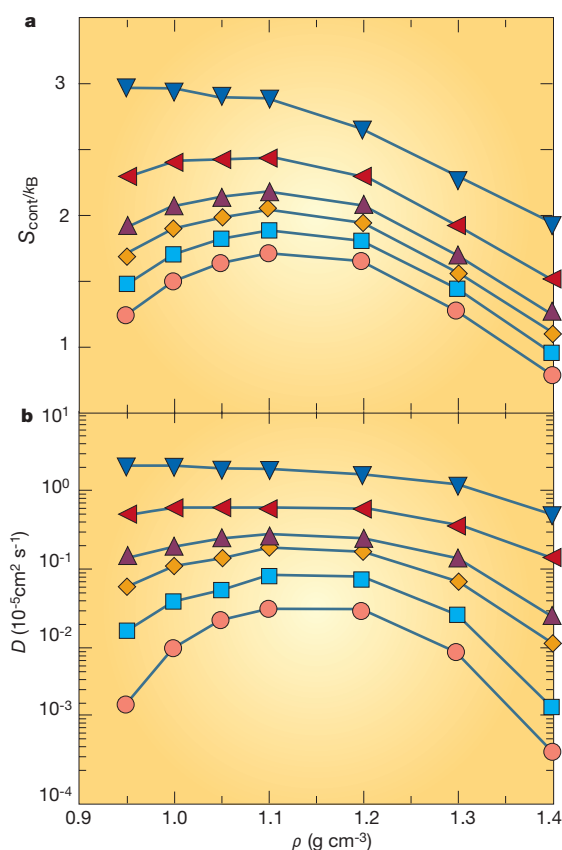
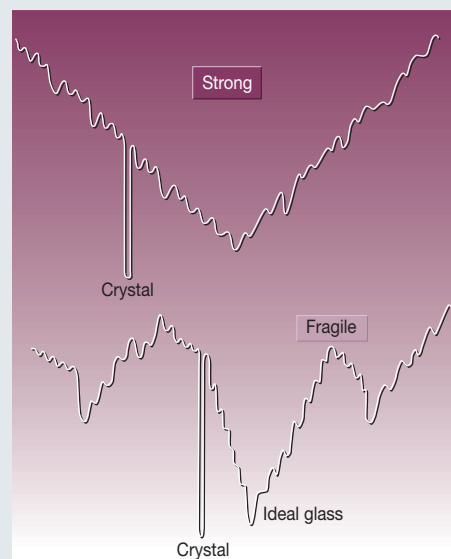


Figure 7 Relationship between diffusivity (D) and configurational entropy (S_{conf}) of supercooled water⁷⁹ at six different temperatures. Filled and open symbols from top to bottom represent the following temperatures: 300 K, 260 K, 240 K, 230 K, 220 K and 210 K; ρ is density. The configurational entropy, which is related to the number of potential energy minima of a given depth (Box 2), was calculated by subtracting the vibrational contribution from the total entropy. The calculations involved performing molecular dynamics simulations of the extended simple point charge (SPC/E) model of water¹⁰² over a range of temperatures and densities. (Adapted from ref. 79.)

Figure 8 Schematic representation of the energy landscapes of strong and fragile substances. The potential energy increases vertically, and the horizontal direction represents collective configurational coordinates.



structural arrest occurs as a purely dynamic singularity, that is to say it is not accompanied by thermodynamic signatures such as a diverging correlation length. What is now known as the idealized MCT^{87,88} predicts structural arrest to occur at a temperature T_x . Initially, therefore, it was thought that MCT was a useful theory for the laboratory-generated glass transition. It is now widely understood that this is not the case, as one finds that $T_x > T_g$, and the MCT-predicted singularity does not occur. In subsequent modifications of the theory⁸⁹, additional relaxation mechanisms occur, often referred to as ‘hopping’ or activated motions, which restore ergodicity (the system’s ability to sample all configurations) below T_x , thereby avoiding a kinetic singularity. These additional relaxation modes arise as a result of a coupling between fluctuations in density and momentum.

Although not a theory of the glass transition, MCT accurately describes many important aspects of relaxation dynamics in liquids above or moderately below their melting temperatures. In particular, the theory makes detailed predictions about the behaviour of the intermediate scattering function F , an experimentally observable quantity that measures the decay of density fluctuations. After a fast initial decay due to microscopic intermolecular collisions, MCT predicts that the decay of F obeys the following sequence (Fig. 9): (i) power-law decay towards a plateau, according to $F = f + At^{-a}$; (ii) a second power-law decay away from the plateau value $F = f - Bt^b$; and (iii) slow relaxation at longer times, which can be fitted by the KWW function $F = \exp[-(t/\tau)^\beta]$. Here, f is the plateau value of the scattering function, which only appears at sufficiently low temperature; t is time; A , B , a and b are constants; τ is the characteristic, temperature-dependent relaxation time; and $\beta < 1$ is the KWW stretch exponent. The basic accuracy of these detailed predictions has been verified experimentally and in computer simulations^{90–92}.

Kivelson and co-workers have proposed a theory of supercooled liquids that is based also on an avoided singularity^{24,93–95}. According to this viewpoint, the liquid has an energetically preferred local structure that differs from the structure in the actual crystalline phase. The system is prevented from crystallizing into a reference crystal with the preferred local structure because of geometric frustration owing to the fact that the latter does not tile space. An example of such energetically favoured but non-space-tiling local structure is the icosahedral packing seen in computer simulations of the supercooled Lennard–Jones liquid⁷³. At a temperature T^* the system would, but for frustration, crystallize into the reference crystal. Instead, strain build-up causes the system to break up into frustration-limited domains, thereby avoiding a phase transition (singularity) at T^* . The avoided transition temperature T^* acts as a critical point, below which two length scales emerge, both of which are large compared to

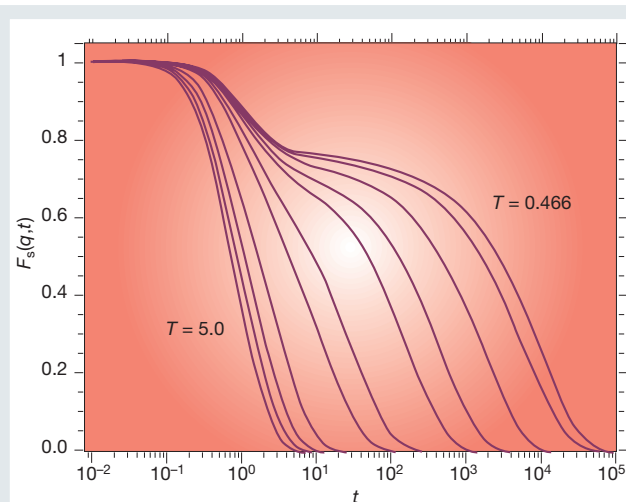


Figure 9 Evolution of the self-intermediate scattering function for A-type atoms for the same supercooled Lennard–Jones mixture as in Fig. 6, at $q\sigma_{AA} = 7.251$, corresponding to the first peak of the static structure factor of species A (ref. 92). Here q is the magnitude of the wave vector. Temperature and time are in units of ϵ_{AA}/k_B and $\sigma_{AA}(m/48\epsilon_{AA})^{1/2}$, respectively. Temperatures from left to right are 5, 4, 3, 2, 1, 0.8, 0.6, 0.55, 0.5, 0.475, and 0.466. The self-intermediate scattering function is the space Fourier transform of the van Hove function $G_s(r, t)$, which is proportional to the probability of observing a particle at $r \pm dr$ at time t given that the same particle was at the origin at $t = 0$. Note the two-step relaxation behaviour upon decreasing T . Molecular dynamics simulations of 1,000 atoms. (Adapted from refs 9 and 92.)

molecular dimensions. One is the critical correlation length, which governs density fluctuations in the absence of frustration. The second is the frustration-limited domain size. From these considerations there emerge predictions on the temperature dependence of the viscosity. Experimental data analysed according to the theory display universality²⁴, but at the expense of introducing a number of fitting parameters. The improvement with respect to competing interpretations is a matter of controversy²⁵.

Challenges and open questions

Important aspects of the complex behaviour of viscous liquids close to the glass transition can be explained qualitatively from the energy landscape perspective. Making this descriptive picture quantitative and predictive is a major challenge. This will require investigating how basic landscape features such as the basin enumeration function depend on molecular architecture and, for a given substance or mixture, on density (see ref. 96 for an example of such a calculation). Equally important is the translation of qualitative pictures such as Fig. 8 into precise measures of strength and fragility based on the basin enumeration function. Uncovering the topographic nature of the scaling relationship between basin minima and saddle points holds the key to understanding the relationship between kinetics and thermodynamics in deeply supercooled liquids. All of these calculations are in principle straightforward, but computationally at the very limit of what is currently feasible. The development of theoretical models⁹⁷ is therefore of paramount importance.

MCT and the landscape perspective offer complementary viewpoints of the same phenomena. So far, however, not enough effort has been devoted to bridging the gap that separates these two approaches. Recent calculations^{70,74} offer the promise of establishing a clearer connection between the static landscape viewpoint and the dynamic perspective of MCT. At the least, what is required is a precise landscape-based explanation of what ‘hopping’ and ‘activated processes’ really mean. Additional theoretical viewpoints of supercooling and the glass transition include the instantaneous normal-mode perspective on liquid dynamics⁹⁸ and thermodynamic

treatments of the vitreous state based on invoking analogies to spin glasses^{99–101}. Establishing a coherent theoretical perspective on supercooled liquids and glasses is important. We believe that the landscape formalism offers the natural technical tools for accomplishing this task.

- Angell, C. A. Formation of glasses from liquids and biopolymers. *Science* **267**, 1924–1935 (1995).
- Blanshard, J. M. V. & Lillford, P. (eds) *The Glassy State in Foods* (Nottingham Univ. Press, Nottingham, 1993).
- Crowe, J. H., Carpenter, J. F. & Crowe, L. M. The role of vitrification in anhydrobiosis. *Annu. Rev. Physiol.* **60**, 73–103 (1998).
- Debenedetti, P. G., Stillinger, F. H., Truskett, T. M. & Lewis, C. P. Theory of supercooled liquids and glasses: energy landscape and statistical geometry perspectives. *Adv. Chem. Eng.* (in the press).
- Greer, A. L. Metallic glasses. *Science* **267**, 1947–1953 (1995).
- Jenniskens, P. & Blake, D. F. Structural transitions in amorphous water ice and astrophysical implications. *Science* **265**, 753–756 (1994).
- Anderson, P. W. Through a glass lightly. *Science* **267**, 1615 (1995).
- Angell, C. A., Ngai, K. L., McKenna, G. B., McMillan, P. F. & Martin, S. W. Relaxation in glass-forming liquids and amorphous solids. *J. Appl. Phys.* **88**, 3113–3157 (2000).
- Debenedetti, P. G. *Metastable Liquids. Concepts and Principles* (Princeton Univ. Press, Princeton, 1996).
- Turnbull, D. Under what conditions can a glass be formed? *Contemp. Phys.* **10**, 473–488 (1969).
- Angell, C. A. Structural instability and relaxation in liquid and glassy phases near the fragile liquid limit. *J. Non-Cryst. Solids* **102**, 205–221 (1988).
- Moynihan, C. T. et al. in *The Glass Transition and the Nature of the Glassy State* (eds Goldstein, M. & Simha, R.) *Ann. NY Acad. Sci.* **279**, 15–36 (1976).
- Brüning, R. & Samwer, K. Glass transition on long time scales. *Phys. Rev. B* **46**, 318–322 (1992).
- Ediger, M. D., Angell, C. A. & Nagel, S. R. Supercooled liquids and glasses. *J. Phys. Chem.* **100**, 13200–13212 (1996).
- Vogel, H. Das temperatur-abhängigkeitsgesetz der viskosität von flüssigkeiten. *Phys. Zeit.* **22**, 645–646 (1921).
- Tammann, G. & Hesse, W. Die abhängigkeit der viskosität von der temperatur bei unterkühlten flüssigkeiten. *Z. Anorg. Allg. Chem.* **156**, 245–257 (1926).
- Fulcher, G. S. Analysis of recent measurements of the viscosity of glasses. *J. Am. Ceram. Soc.* **8**, 339 (1925).
- Laughlin, W. T. & Uhlmann, D. R. Viscous flow in simple organic liquids. *J. Phys. Chem.* **76**, 2317–2325 (1972).
- Angell, C. A. in *Relaxations in Complex Systems* (eds Ngai, K. & Wright, G. B.) **1** (Nat'l Technol. Inform. Ser., US Dept. of Commerce, Springfield, VA, 1985).
- Angell, C. A. Relaxation in liquids, polymers and plastic crystals—strong/fragile patterns and problems. *J. Non-Cryst. Solids* **131–133**, 13–31 (1991).
- Green, J. L., Ito, K., Xu, K. & Angell, C. A. Fragility in liquids and polymers: new, simple quantifications and interpretations. *J. Phys. Chem. B* **103**, 3991–3996 (1999).
- Novikov, V. N., Rössler, E., Malinovsky, V. K. & Surovstev, N. V. Strong and fragile liquids in percolation approach to the glass transition. *Europhys. Lett.* **35**, 289–294 (1996).
- Fujimori, H. & Oguni, M. Correlation index $(T_g - T_{\infty})/T_{\infty}$ and activation energy ratio $\Delta E_{\infty}/\Delta E_{\infty}$ as parameters characterizing the structure of liquid and glass. *Solid State Commun.* **94**, 157–162 (1995).
- Kivelson, D., Tarjus, G., Zhao, X. & Kivelson, S. A. Fitting of viscosity: distinguishing the temperature dependencies predicted by various models of supercooled liquids. *Phys. Rev. E* **53**, 751–758 (1996).
- Cummins, H. Z. Comment on “Fitting of viscosity: distinguishing the temperature dependencies predicted by various models of supercooled liquids”. *Phys. Rev. E* **54**, 5870–5872 (1996).
- Kohlrausch, R. Theorie des elektrischen rückstandes in der leidener flasche. *Ann. Phys. Chem. (Leipzig)* **91**, 179–214 (1874).
- Williams, G. & Watts, D. C. Non-symmetrical dielectric relaxation behavior arising from a simple empirical decay function. *Trans. Faraday Soc.* **66**, 80–85 (1970).
- Richert, R. & Blumen, A. in *Disorder Effects on Relaxational Processes* (eds Richert, R. & Blumen, A.) **1–7** (Springer, Berlin, 1994).
- Cicerone, M. T. & Ediger, M. D. Relaxation of spatially heterogeneous dynamic domains in supercooled ortho-terphenyl. *J. Chem. Phys.* **103**, 5684–5692 (1995).
- Cicerone, M. T. & Ediger, M. D. Enhanced translation of probe molecules in supercooled o-terphenyl: signature of spatially heterogeneous dynamics? *J. Chem. Phys.* **104**, 7210–7218 (1996).
- Me’cuk, A. I., Ramos, R. A., Gould, H., Klein, W. & Mountain, R. D. Long-lived structures in fragile glass-forming liquids. *Phys. Rev. Lett.* **75**, 2522–2525 (1995).
- Hurley, M. M. & Harrowell, P. Non-gaussian behavior and the dynamical complexity of particle motion in a dense two-dimensional liquid. *J. Chem. Phys.* **105**, 10521–10526 (1996).
- Perera, D. N. & Harrowell, P. Measuring diffusion in supercooled liquids: the effect of kinetic inhomogeneities. *J. Chem. Phys.* **104**, 2369–2375 (1996).
- Perera, D. N. & Harrowell, P. Consequence of kinetic inhomogeneities in glasses. *Phys. Rev. E* **54**, 1652–1662 (1996).
- Donati, C., Glotzer, S. C., Poole, P. H., Kob, W. & Plimpton, S. J. Spatial correlations of mobility and immobility in a glass-forming Lennard–Jones liquid. *Phys. Rev. E* **60**, 3107–3119 (1999).
- Böhmer, R., Hinze, G., Diezemann, G., Geil, B. & Sillescu, H. Dynamic heterogeneity on supercooled ortho-terphenyl studied by multidimensional deuterium NMR. *Europhys. Lett.* **36**, 55–60 (1996).
- Wang, C.-Y. & Ediger, M. D. How long do regions of different dynamics persist in supercooled o-terphenyl? *J. Phys. Chem. B* **103**, 4177–4184 (1999).
- Vidal Russell, E. & Israeloff, N. E. Direct observation of molecular cooperativity near the glass transition. *Nature* **408**, 695–698 (2000).
- Ediger, M. D. Spatially heterogeneous dynamics in supercooled liquids. *Annu. Rev. Phys. Chem.* **51**, 99–128 (2000).
- Fujara, F., Geil, B., Sillescu, H. H. & Fleischer, G. Translational and rotational diffusion in supercooled ortho-terphenyl close to the glass transition. *Z. Phys. B Cond. Matt.* **88**, 195–204 (1992).
- Johari, G. P. Intrinsic mobility of molecular glasses. *J. Chem. Phys.* **58**, 1766–1770 (1973).
- Johari, G. P. & Goldstein, M. Viscous liquids and the glass transition. II. Secondary relaxations in glasses of rigid molecules. *J. Chem. Phys.* **53**, 2372–2388 (1970).

43. Rössler, E., Warschewski, U., Eiermann, P., Sokolov, A. P. & Quitmann, D. Indications for a change of transport mechanism in supercooled liquids and the dynamics close and below T_g . *J. Non-Cryst. Solids* **172–174**, 113–125 (1994).
44. Stillinger, F. H. A topographic view of supercooled liquids and glass formation. *Science* **267**, 1935–1939 (1995).
45. Kauzmann, W. The nature of the glassy state and the behavior of liquids at low temperatures. *Chem. Rev.* **43**, 219–256 (1948).
46. Simon, F. Über den Zustand der unterkühlten Flüssigkeiten und Gläser. *Z. Anorg. Allg. Chem.* **203**, 219–227 (1931).
47. Wolynes, P. G. Aperiodic crystals: biology, chemistry and physics in a fugue with stretto. *AIP Conf. Proc.* **180**, 39–65 (1988).
48. Wolynes, P. G. Entropy crises in glasses and random heteropolymers. *J. Res. Natl. Inst. Standards Technol.* **102**, 187–194 (1997).
49. Angell, C. A. Landscapes with megabasins: polyamorphism in liquids and biopolymers and the role of nucleation in folding and folding diseases. *Physica D* **107**, 122–142 (1997).
50. Gibbs, J. H. & DiMarzio, E. A. Nature of the glass transition and the glassy state. *J. Chem. Phys.* **28**, 373–383 (1958).
51. Adam, G. & Gibbs, J. H. On the temperature dependence of cooperative relaxation properties in glass-forming liquids. *J. Chem. Phys.* **43**, 139–146 (1965).
52. Richert, R. & Angell, C. A. Dynamics of glass-forming liquids. V. On the link between molecular dynamics and configurational entropy. *J. Chem. Phys.* **108**, 9016–9026 (1998).
53. Williams, M. L., Landel, R. F. & Ferry, J. D. The temperature dependence of the relaxation mechanisms in amorphous polymers and other glass-forming liquids. *J. Am. Chem. Soc.* **77**, 3701–3707 (1955).
54. Ito, K., Moynihan, C. T. & Angell, C. A. Thermodynamic determination of fragility in liquids and a fragile-to-strong liquid transition in water. *Nature* **398**, 492–495 (1999).
55. Goldstein, M. Viscous liquids and the glass transition: a potential energy barrier picture. *J. Chem. Phys.* **51**, 3728–3739 (1969).
56. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. The energy landscapes and motions of proteins. *Science* **254**, 1598–1603 (1991).
57. Nienhaus, G. U., Müller, J. D., McMahon, B. H. & Frauenfelder, H. Exploring the conformational energy landscape of proteins. *Physica D* **107**, 297–311 (1997).
58. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. Free energy landscape for protein folding kinetics: intermediates, traps, and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.* **101**, 6052–6062 (1994).
59. Saven, J. G., Wang, J. & Wolynes, P. G. Kinetics of protein folding: the dynamics of globally connected rough energy landscapes with biases. *J. Chem. Phys.* **101**, 11037–11043 (1994).
60. Wang, J., Onuchic, J. & Wolynes, P. Statistics of kinetic pathways on biased rough energy landscapes with applications to protein folding. *Phys. Rev. Lett.* **76**, 4861–4864 (1996).
61. Plotkin, S. S., Wang, J. & Wolynes, P. G. Correlated energy landscape model for finite, random heteropolymers. *Phys. Rev. E* **53**, 6271–6296 (1996).
62. Becker, O. M. & Karplus, M. The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics. *J. Chem. Phys.* **106**, 1495–1517 (1997).
63. Dill, K. A. & Chan, H. S. From Levinthal to pathways and funnels. *Nature Struct. Biol.* **4**, 10–19 (1997).
64. Klepeis, J. L., Floudas, C. A., Morikis, D. & Lambris, J. D. Predicting peptide structure using NMR data and deterministic global optimization. *J. Comp. Chem.* **20**, 1354–1370 (1999).
65. Lacks, D. J. Localized mechanical instabilities and structural transformations in silica glass under high pressure. *Phys. Rev. Lett.* **80**, 5385–5388 (1998).
66. Malandro, D. L. & Lacks, D. J. Volume dependence of potential energy landscapes in glasses. *J. Chem. Phys.* **107**, 5804–5810 (1997).
67. Malandro, D. L. & Lacks, D. J. Relationships of shear-induced changes in the potential energy landscape to the mechanical properties of ductile glasses. *J. Chem. Phys.* **110**, 4593–4601 (1999).
68. Malandro, D. L. & Lacks, D. J. Molecular-level instabilities and enhanced self-diffusion in flowing liquids. *Phys. Rev. Lett.* **81**, 5576–5579 (1998).
69. Schulz, M. Energy landscape, minimum points, and non-Arrhenius behavior of supercooled liquids. *Phys. Rev. B* **57**, 11319–11333 (1998).
70. Sastry, S., Debenedetti, P. G. & Stillinger, F. H. Signatures of distinct dynamical regimes in the energy landscape of a glass-forming liquid. *Nature* **393**, 554–557 (1998).
71. Keyes, T. Dependence of supercooled liquid dynamics on elevation in the energy landscape. *Phys. Rev. E* **59**, 3207–3211 (1999).
72. Debenedetti, P. G., Stillinger, F. H., Truskett, T. M. & Roberts, C. J. The equation of state of an energy landscape. *J. Phys. Chem. B* **103**, 7390–7397 (1999).
73. Jonsson, H. & Andersen, H. C. Icosahedral ordering in the Lennard-Jones crystal and glass. *Phys. Rev. Lett.* **60**, 2295–2298 (1988).
74. Angelani, L., Di Leonardo, R., Ruocco, G., Scala, A. & Sciortino, F. Saddles in the energy landscape probed by supercooled liquids. *Phys. Rev. Lett.* **85**, 5356–5359 (2000).
75. Stillinger, F. H., Debenedetti, P. G. & Sastry, S. Resolving vibrational and structural contributions to isothermal compressibility. *J. Chem. Phys.* **109**, 3983–3988 (1998).
76. Stillinger, F. H. & Debenedetti, P. G. Distinguishing vibrational and structural equilibration contributions to thermal expansion. *J. Phys. Chem. B* **103**, 4052–4059 (1999).
77. Sciortino, F., Kob, W. & Tartaglia, P. Inherent structure entropy of supercooled liquids. *Phys. Rev. Lett.* **83**, 3214–3217 (1999).
78. Büchner, S. & Heuer, A. Potential energy landscape of a model glass former: thermodynamics, anharmonicities, and finite size effects. *Phys. Rev. E* **60**, 6507–6518 (1999).
79. Scala, A., Starr, F. W., La Nave, E., Sciortino, F. & Stanley, H. E. Configurational entropy and diffusivity in supercooled water. *Nature* **406**, 166–169 (2000).
80. Prielmeier, F. X., Lang, E. W., Speedy, R. J. & Lüdemann, H.-D. Diffusion in supercooled water to 300 Mpa. *Phys. Rev. Lett.* **59**, 1128–1131 (1987).
81. Mackenzie, J. D. Viscosity-temperature relationship for network liquids. *J. Am. Ceram. Soc.* **44**, 598–601 (1961).
82. Greet, R. J. & Turnbull, D. Glass transition in o-terphenyl. *J. Chem. Phys.* **46**, 1243–1251 (1967).
83. Stillinger, F. H. & Hodgdon, J. A. Translation-rotation paradox for diffusion in fragile glass-forming liquids. *Phys. Rev. E* **50**, 2064–2068 (1994).
84. Tarjus, G. & Kivelson, D. Breakdown of the Stokes-Einstein relation in supercooled liquids. *J. Chem. Phys.* **103**, 3071–3073 (1995).
85. Liu, C. Z.-W. & Openheim, I. Enhanced diffusion upon approaching the kinetic glass transition. *Phys. Rev. E* **53**, 799–802 (1996).
86. Gesztzi, T. Pre-vitrification by viscosity feedback. *J. Phys. C* **16**, 5805–5814 (1983).
87. Bengtzelius, U., Götz, W. & Sjölander, A. Dynamics of supercooled liquids and the glass transition. *J. Phys. C* **17**, 5915–5934 (1984).
88. Götz, W. & Sjögren, L. Relaxation processes in supercooled liquids. *Rep. Prog. Phys.* **55**, 241–376 (1992).
89. Götz, W. & Sjögren, L. The mode coupling theory of structural relaxations. *Transp. Theory Stat. Phys.* **24**, 801–853 (1995).
90. Götz, W. Recent tests of the mode-coupling theory for glassy dynamics. *J. Phys. Cond. Matt.* **11**, A1–A45 (1999).
91. Kob, W. Computer simulations of supercooled liquids and glasses. *J. Phys. Cond. Matt.* **11**, R85–R115 (1999).
92. Kob, W. & Andersen, H. C. Testing mode-coupling theory for a supercooled binary Lennard-Jones mixture: the van Hove correlation function. *Phys. Rev. E* **51**, 4626–4641 (1995).
93. Kivelson, D., Kivelson, S. A., Zhao, X., Nussinov, Z. & Tarjus, G. A thermodynamic theory of supercooled liquids. *Physica A* **219**, 27–38 (1995).
94. Kivelson, D. & Tarjus, G. SuperArrhenius character of supercooled glass-forming liquids. *J. Non-Cryst. Solids* **235–237**, 86–100 (1998).
95. Kivelson, D. & Tarjus, G. The Kauzmann paradox interpreted via the theory of frustration-limited domains. *J. Chem. Phys.* **109**, 5481–5486 (1998).
96. Sastry, S. The relationship between fragility, configurational entropy and the potential energy landscape of glass-forming liquids. *Nature* **409**, 164–167 (2001).
97. Speedy, R. J. Relations between a liquid and its glasses. *J. Phys. Chem. B* **103**, 4060–4065 (1999).
98. Keyes, T. Instantaneous normal mode approach to liquid state dynamics. *J. Phys. Chem. A* **101**, 2921–2930 (1997).
99. Kirkpatrick, T. R. & Wolynes, P. G. Stable and metastable states in mean-field Potts and structural glasses. *Phys. Rev. B* **36**, 8552–8564 (1987).
100. Kirkpatrick, T. R., Thirumalai, D. & Wolynes, P. G. Scaling concepts for the dynamics of viscous liquids near an ideal glassy state. *Phys. Rev. A* **40**, 1045–1054 (1989).
101. Mézard, M. & Parisi, G. Thermodynamics of glasses: a first principles computation. *Phys. Rev. Lett.* **82**, 747–750 (1999).
102. Berendsen, H. J., Grigera, J. R. & Stroatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **91**, 6269–6271 (1987).
103. Stillinger, F. H. Supercooled liquids, glass transitions, and the Kauzmann paradox. *J. Chem. Phys.* **88**, 7818–7825 (1988).
104. Santen, L. & Krauth, W. Absence of thermodynamic phase transition in a model glass former. *Nature* **405**, 550–551 (2000).
105. Wilks, J. *The Properties of Liquid and Solid Helium* (Clarendon, Oxford, 1967).
106. Rastogi, S., Höhne, G. W. H. & Keller, A. Unusual pressure-induced phase behavior in crystalline Poly(4-methylpentene-1): calorimetric and spectroscopic results and further implications. *Macromolecules* **32**, 8897–8909 (1999).
107. Greer, A. L. Too hot to melt. *Nature* **404**, 134–135 (2000).
108. Stillinger, F. H. Exponential multiplicity of inherent structures. *Phys. Rev. E* **59**, 48–51 (1999).
109. Stillinger, F. H. Enumeration of isobaric inherent structures for the fragile glass former o-terphenyl. *J. Phys. Chem. B* **102**, 2807–2810 (1998).

Acknowledgements

P.G.D.'s work is supported by the US Department of Energy.

Supercooled liquids and the glass transition

Pablo G. Debenedetti* & Frank H. Stillinger†‡

*Department of Chemical Engineering and ‡Princeton Materials Institute, Princeton University, Princeton, New Jersey 08544, USA (e-mail: pdebene@princeton.edu)

†Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA

Glasses are disordered materials that lack the periodicity of crystals but behave mechanically like solids. The most common way of making a glass is by cooling a viscous liquid fast enough to avoid crystallization. Although this route to the vitreous state — supercooling — has been known for millennia, the molecular processes by which liquids acquire amorphous rigidity upon cooling are not fully understood. Here we discuss current theoretical knowledge of the manner in which intermolecular forces give rise to complex behaviour in supercooled liquids and glasses. An intriguing aspect of this behaviour is the apparent connection between dynamics and thermodynamics. The multidimensional potential energy surface as a function of particle coordinates (the energy landscape) offers a convenient viewpoint for the analysis and interpretation of supercooling and glass-formation phenomena. That much of this analysis is at present largely qualitative reflects the fact that precise computations of how viscous liquids sample their landscape have become possible only recently.

The glassy state is ubiquitous in nature and technology¹. It is crucial in the processing of foods², the commercial stabilization of labile biochemicals³, and the preservation of insect life under extremes of cold or dehydration³.

Window glass, composed mostly of sand, lime and soda, is the best-known example of an engineered amorphous solid⁴. Optical fibres are made of very pure amorphous silica, occasionally carefully doped. Most engineering plastics are amorphous solids, as are some metallic glasses and alloys of interest because of their soft magnetism and corrosion resistance⁵. The silicon used in many photovoltaic cells is amorphous, and it is possible that most water in the Universe may be glassy⁶. Most of these examples entail supercooling of a liquid to take advantage of viscous retardation of nucleation and crystallization. Understanding quantitatively the extraordinary viscous slow-down that accompanies supercooling and glass formation is a major scientific challenge⁷.

We begin by reviewing the phenomenology of vitrification and supercooling. A useful approach for relating this complex phenomenology to molecular-scale events is to focus attention on the liquid's energy landscape, that is, the multidimensional surface generated by the system's potential energy as a function of molecular coordinates. Accordingly, basic landscape concepts and a discussion of the important theoretical and computational progress currently being made in this area are presented next. This is followed by a discussion of alternative viewpoints, in which narrowly avoided singularities are assumed to occur well above the glass-transition temperature. We then close with a summary of the important open questions.

It is impossible to do justice to the entire field of supercooled liquids and amorphous solids in an article of this length. We have therefore limited the scope to the dynamics and thermodynamics of viscous liquids above and close to the glass-transition temperature T_g — in other words, to the glass transition viewed 'from the liquid'. The view 'from the solid', including such topics as relaxation both relatively near

(for example, during annealing or ageing) and far below T_g , is not discussed. The reader is referred to an excellent recent review⁸ for a thorough coverage of these and other topics.

Phenomenology of supercooling and glass formation

Figure 1 illustrates the temperature dependence of a liquid's volume (or enthalpy) at constant pressure^{4,9}. Upon cooling below the freezing point T_m , molecular motion slows down. If the liquid is cooled sufficiently fast, crystallization can be avoided^{10,11}. Eventually molecules will rearrange so slowly that they cannot adequately sample configurations in the available time allowed by the cooling rate. The liquid's structure therefore appears 'frozen' on the laboratory timescale (for example, minutes). This falling out of equilibrium occurs across a narrow transformation range where the characteristic molecular relaxation time becomes of the order of 100 seconds, and the rate of change of volume or enthalpy with respect to temperature decreases abruptly (but continuously) to a value comparable to that of a crystalline solid. The resulting material is a glass. The intersection of the liquid and vitreous portions of the volume versus temperature curve provides one definition of T_g , which usually occurs around $2T_m/3$. The behaviour depicted in Fig. 1 is not a true phase transition, as it does not involve discontinuous changes in any physical property.

The slower a liquid is cooled, the longer the time available for configurational sampling at each temperature, and hence the colder it can become before falling out of liquid-state equilibrium. Consequently, T_g increases with cooling rate^{12,13}. The properties of a glass, therefore, depend on the process by which it is formed. In practice, the dependence of T_g on the cooling rate is weak (T_g changes by 3–5 °C when the cooling rate changes by an order of magnitude¹⁴), and the transformation range is narrow, so that T_g is an important material characteristic.

Slowing down

Another definition of T_g is the temperature at which the shear viscosity reaches 10^{13} poise. Close to T_g the viscosity η

is extraordinarily sensitive to temperature. For silica this dependence is reasonably well described by the Arrhenius functionality, $\eta = A \exp(E/k_B T)$, where A and E are temperature-independent and k_B is Boltzmann's constant. Other liquids exhibit an even more pronounced viscous slow-down close to the glass transition, which is reasonably well represented, over 2–4 orders of magnitude in viscosity⁸, by the Vogel–Tammann–Fulcher (VTF) equation^{15–17}

$$\eta = A \exp[B/(T - T_0)] \quad (1)$$

where A and B are temperature-independent constants. Understanding the origin of this extraordinary slow-down of relaxation processes is one of the main challenges in the physics of glasses.

Figure 2 shows a T_g -scaled Arrhenius representation of liquid viscosities^{18–21}. Angell has proposed a useful classification of liquids

along a 'strong' to 'fragile' scale. The viscosity and relaxation times (for example, dielectric relaxation) of the former behave in nearly Arrhenius fashion, whereas fragile liquids show marked deviations from Arrhenius behaviour. Silica (SiO_2) is the prototypical strong liquid, whereas *o*-terphenyl (OTP) is the canonical fragile glass-former. Strong liquids, such as the network oxides SiO_2 and germanium dioxide (GeO_2), have tetrahedrally coordinated structures, whereas the molecules of fragile liquids exert largely non-directional, dispersive forces on each other. Alternative scaling descriptions that attempt to extract universal aspects of viscous slow-down have been proposed^{22–24}. Their relative merits are still being assessed^{8,25}.

Viscous liquids close to T_g exhibit non-exponential relaxation. The temporal behaviour of the response function $F(t)$ (for example, the polarization in response to an applied electric field, the strain (deformation) resulting from an applied stress, or the stress in

Box 1

Entropy crises

Boltzmann's entropy formula establishes the connection between the microscopic world of atoms and molecules and the bulk properties of matter:

$$S(N, V, E) = k_B \ln \Omega$$

In this equation, S is the entropy, k_B is Boltzmann's constant and Ω is the number of quantum states accessible to N particles with fixed energy E in a volume V . Because Ω cannot be less than one, the entropy cannot be negative. When a crystal is cooled sufficiently slowly, it approaches a unique state of lowest energy, and hence its entropy approaches zero as $T \rightarrow 0$. If the entropy of a supercooled liquid were to become smaller than that of the stable crystal at the Kauzmann temperature, its entropy would eventually become negative upon further cooling. This impossible scenario constitutes an entropy crisis^{46–48}.

The Kauzmann temperature T_K is given by⁹

$$\Delta s_m = \int_{T_K}^{T_m} \frac{\Delta C_p}{T} dT$$

where Δs_m is the melting entropy (the difference between liquid and crystal entropies at the melting temperature), T_m is the melting temperature at the given pressure, and ΔC_p is the temperature-dependent difference between the heat capacity of the liquid and the crystal at the given pressure. The rate of change of entropy with temperature at constant pressure is given by

$$\left(\frac{\partial S}{\partial T}\right)_P = \frac{C_p}{T}$$

The entropy crisis arises because the heat capacity of a liquid is greater than that of the stable crystal. The entropy of fusion is therefore consumed upon supercooling, and vanishes at T_K . The entropy crisis entails no conflict with the second law of thermodynamics, as the difference in chemical potential $\Delta\mu$ between the supercooled liquid and the stable crystal at T_K is a positive quantity. Because the chemical potential is the Gibbs free energy per unit mass, this means that the system can reduce its Gibbs free energy by freezing, in accord with experience. The chemical potential difference at T_K is given by⁹

$$\Delta\mu(T_K) = \int_{T_K}^{T_m} \Delta C_p \left(\frac{T_m}{T} - 1\right) dT$$

One way of avoiding the entropy crisis is for the liquid to form an ideal glass of unique configuration at T_K . This is the thermodynamic view of the glass transition, according to which the observable glass transition is a manifestation, masked by kinetics, of an underlying second-order phase transition⁵⁰ occurring at T_K .

Because the glass transition intervenes before the entropy crisis occurs ($T_g > T_K$), estimates of the Kauzmann temperature involve an extrapolation of liquid properties below T_g . The validity of such extrapolations, and hence of the very possibility of an entropy crisis, has been questioned by Stillinger¹⁰³, owing to the apparent necessity for configurational excitations out of the ideal glass state to require an unbounded energy. Furthermore, recent computer simulations of polydisperse hard disks found no evidence of a thermodynamic basis underlying the glass transition¹⁰⁴. Although the notion of an ideal glass remains controversial¹⁰³, this does not undermine the usefulness of T_K as an empirical classification parameter for glass-formers.

Experimentally, there are substances with known Kauzmann points. ⁴He is a liquid at 0 K and 1 bar (liquid He-II). Its equilibrium freezing pressure at 0 K is 26 bar. At this point, the entropies of the liquid and the crystal are equal, and this is therefore a Kauzmann point, although not an entropy crisis as both phases have zero entropy. The melting curves of ³He and ⁴He exhibit pressure minima: these occur at about 0.32 K and 0.8 K, respectively¹⁰⁵. These are also equal-entropy (Kauzmann) points. Experiments indicate that poly(4-methylpentene-1) exhibits a pressure maximum along its melting curve¹⁰⁶. Although the appearance of an additional phase complicates the interpretation of this observation, the implication would be that the pressure maximum is a Kauzmann point, and the continuation of the melting curve to lower temperatures and pressures beyond this point corresponds to endothermic freezing of a stable liquid into a crystal possessing higher entropy¹⁰⁷. How this liquid would avoid conflict with the third law is not understood, but may hinge on the vibrational anharmonicity of the two phases with changing temperature.

response to an imposed deformation) can often be described by the stretched exponential, or Kohlrausch–Williams–Watts (KWW) function^{26,27}

$$F(t) = \exp[-(t/\tau)^\beta] \quad (\beta < 1) \quad (2)$$

where $F(t) = [\sigma(t) - \sigma(\infty)]/[\sigma(0) - \sigma(\infty)]$ and σ is the measured quantity (for example, the instantaneous stress following a step change in deformation). τ in equation (2) is a characteristic relaxation time, whose temperature dependence is often non-Arrhenius (exhibiting fragile behaviour). The slowing down of long-time relaxation embodied in equation (2) contrasts with the behaviour of liquids above the melting point, which is characterized by simple exponential relaxation. Experimental and computational evidence indicates that this slow-down is related to the growth of distinct relaxing domains^{28–39} (spatial heterogeneity). Whether each of these spatially heterogeneous domains relaxes exponentially or not is a matter of considerable current interest^{38,39}.

Decouplings

In supercooled liquids below approximately $1.2T_g$ there occurs a decoupling between translational diffusion and viscosity, and between rotational and translational diffusion^{30,39,40}. At higher temperatures, both the translational and the rotational diffusion coefficients are inversely proportional to the viscosity, in agreement with the Stokes–Einstein and Debye equations, respectively. Below approximately $1.2T_g$, the inverse relationship between translational motion and viscosity breaks down, whereas that between rotational motion and viscosity does not. Near T_g , it is found that molecules translate faster than expected based on their viscosity, by as much as two orders of magnitude. This therefore means that, as the temperature is lowered, molecules on average translate progressively more for every rotation they execute. Yet another decoupling occurs in the moderately supercooled range. At sufficiently high temperature the liquid shows a single peak relaxation frequency (Fig. 3), indicative of one relaxation mechanism. In the moderately supercooled regime, however, the peak splits into slow (α) and fast (β) relaxations^{41–43}. The former exhibit non-Arrhenius behaviour and disappear at T_g ; the latter continue below T_g and display Arrhenius behaviour⁴⁴.

Thermodynamics

The entropy of a liquid at its melting temperature is higher than that of the corresponding crystal. Because the heat capacity of a liquid is higher than that of the crystal, this entropy difference decreases upon supercooling (Box 1). Figure 4 shows the temperature dependence of the entropy difference between several supercooled liquids and their stable crystals⁴⁵. For lactic acid this entropic surplus is consumed so

fast that a modest extrapolation of experimental data predicts its impending vanishing. In practice, the glass transition intervenes, and ΔS does not vanish. If the glass transition did not intervene, the liquid entropy would equal the crystal's entropy at a nonzero temperature T_K (the Kauzmann temperature.) Because the entropy of the crystal approaches zero as T tends to zero, the entropy of the liquid would eventually become negative upon cooling if this trend were to continue. Because entropy is an inherently non-negative quantity (Box 1), the state of affairs to which liquids such as lactic acid are tending when the glass transition intervenes is an entropy crisis^{46–48}. The extrapolation needed to provoke conflict with the third law is quite modest for many fragile liquids⁴⁹, and the imminent crisis is thwarted by a kinetic phenomenon, the glass transition. This suggests a connection between the kinetics and the thermodynamics of glasses⁴⁷. The thermodynamic viewpoint that emerges from this analysis⁵⁰ considers the laboratory glass transition as a kinetically controlled manifestation of an underlying thermodynamic transition to an ideal glass with a unique configuration.

A formula of Adam and Gibbs⁵¹ provides a suggestive connection between kinetics and thermodynamics:

$$t = A \exp(B/Ts_c) \quad (3)$$

In this equation, t is a relaxation time (or, equivalently, the viscosity) and A and B are constants. s_c , the configurational entropy, is related to the number of minima of the system's multidimensional potential energy surface (Box 2). According to the Adam–Gibbs picture, the origin of viscous slow-down close to T_g is the decrease in the number of configurations that the system is able to sample. At the Kauzmann temperature the liquid would have attained a unique, non-crystalline state of lowest energy, the ideal glass. Because there is no configurational entropy associated with confinement in such a state, the Adam–Gibbs theory predicts structural arrest to occur at T_K . In their derivation of equation (3), Adam and Gibbs invoked the concept of a cooperatively rearranging region (CRR)⁵¹. A weakness of their treatment is the fact that it provides no information on the size of such regions. The fact that the CRRs are indistinguishable from each other is also problematic, in light of the heterogeneity that is believed to underlie stretched exponential behaviour⁸.

Figure 1 Temperature dependence of a liquid's volume v or enthalpy h at constant pressure. T_m is the melting temperature.

A slow cooling rate produces a glass transition at T_{ga} ; a faster cooling rate leads to a glass transition at T_{gb} . The thermal expansion coefficient $\alpha_p = (\partial \ln v / \partial T)_p$ and the isobaric heat capacity $c_p = (\partial h / \partial T)_p$ change abruptly but continuously at T_g .

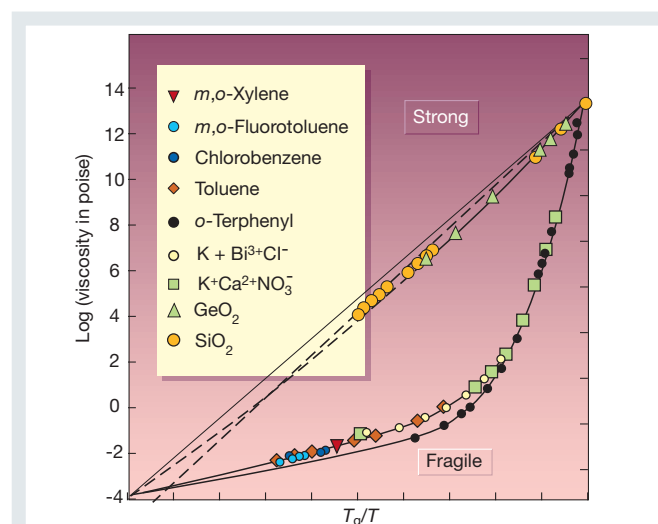
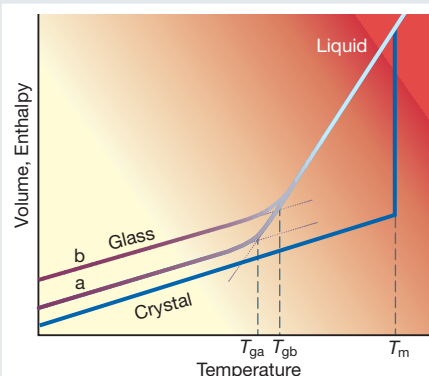


Figure 2 T_g -scaled Arrhenius representation of liquid viscosities showing Angell's strong–fragile pattern. Strong liquids exhibit approximate linearity (Arrhenius behaviour), indicative of a temperature-independent activation energy $E = \ln \eta / d(1/T) \approx \text{const}$. Fragile liquids exhibit super-Arrhenius behaviour, their effective activation energy increasing as temperature decreases. (Adapted from refs 9 and 11.)

Nevertheless, equation (3) describes the relaxation behaviour of deeply supercooled liquids remarkably well. If the difference in heat capacities between a supercooled liquid and its stable crystalline form is inversely proportional to temperature⁵², the Adam–Gibbs

relation yields the VTF equation, which is mathematically equivalent to the Williams–Landel–Ferry equation for the temperature dependence of viscosity in polymers⁵³. This transformation is predicated on the assumption that the vibrational entropies of the

Box 2

Statistics of landscapes

The complexity of many-body landscapes makes a statistical description inevitable. The quantity of interest is the number of minima of given depth, which is given by¹⁰⁸

$$\frac{d\Omega}{d\phi} = C \exp[N\sigma(\phi)]$$

Here, $d\Omega$ denotes the number of potential energy minima with depth per particle ($\phi = \Phi/N$) between ϕ and $\phi \pm d\phi/2$. C is an N -independent factor with units of inverse energy, and $\sigma(\phi)$, also an N -independent quantity, is a so-called basin enumeration function. Taking the logarithm of the above expression and comparing with Boltzmann's entropy formula (Box 1), we see that $\sigma(\phi)$ is the entropy per particle arising from the existence of multiple minima of depth ϕ , or, in other words, the configurational entropy.

At low temperatures, it is possible to separate the configurational contribution to thermophysical properties, which arises from the exploration of different basins, from the vibrational component, which arises from thermal motions confined to a given basin^{75,76}. The Helmholtz free energy A is then given by

$$\frac{A}{NkT} = \frac{\bar{\phi}}{kT} - \sigma(\bar{\phi}) + \frac{a^v}{k_B T}$$

where $\bar{\phi}$ is the depth of the basins preferentially sampled at the given temperature, and a^v is the vibrational free energy per particle. Thus, the free energy consists of an energetic component that reflects the depth of landscape basins sampled preferentially at the given temperature, an entropic component that accounts for the number of existing basins of a given depth, and a vibrational component. The statistical description of a landscape consists of the basin enumeration function $\sigma(\phi)$, from which the excitation profile $\phi(T)$ is obtained through the free-energy minimization condition

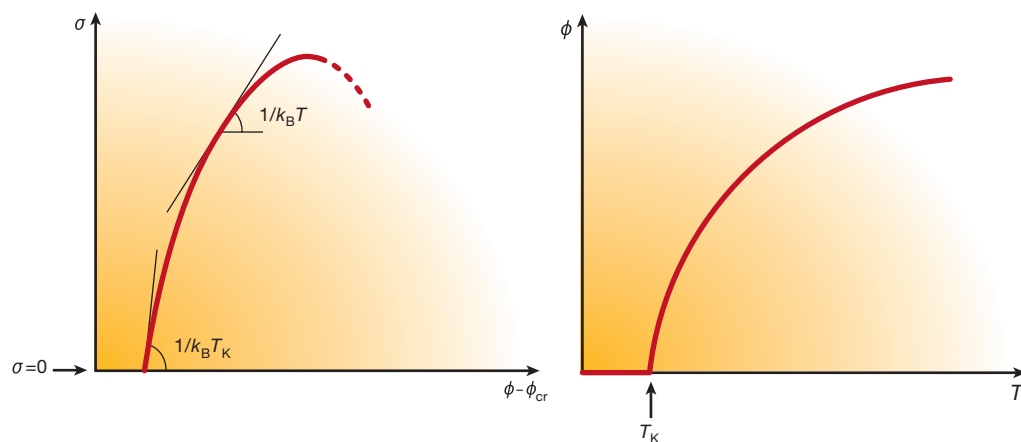
$$\frac{d\sigma}{d\phi} = \frac{1}{k_B T}$$

The above equation assumes that a^v depends on T , but not on ϕ — that is, all basins have the same mean curvature at their respective minima.

The shape of a given system's landscape is determined by the density (number of molecules per unit volume, N/V). Temperature governs the manner in which the landscape is sampled. A different basin enumeration function and excitation profile corresponds to each density. Temperature dictates the point along the enumeration curve and the excitation profile sampled by the system at fixed density (see figure below).

It is possible to construct the basin enumeration function and excitation profile of a system from experimental heat capacity data for the crystal and the supercooled liquid¹⁰⁹, and by computer simulation^{77,78}. In the latter case, the calculations involve determining the probability distribution of inherent structure energies sampled as a function of temperature. These calculations are at the limit of what is presently feasible with available computational power. The enumeration function is often well represented by a parabola, indicative of a gaussian distribution of basins^{4,77,97}. At present it is not understood how the enumeration function deforms with density for a given system (but see ref. 96 for a recent example of such a calculation), or how it depends on molecular architecture. Understanding such questions would provide a direct link between landscape statistics and physical properties. The success of the Adam–Gibbs equation indicates that this link applies also to transport properties such as diffusion and viscosity.

Box 2 Figure Schematic representation of the basin enumeration function (left) and the excitation profile (right). ϕ is the potential energy per particle in mechanically stable potential energy minima. ϕ_{cr} is the corresponding quantity in the stable crystal. The number of potential energy minima with depth between ϕ and $\phi \pm d\phi/2$ is proportional to $\exp[N\sigma(\phi)]$. In the thermodynamic limit (N of



the order of Avogadro's number, 6.02×10^{23}), basins that possess larger (less negative) potential energies (shallow basins) are overwhelmingly more numerous than deeper basins possessing very negative ϕ -values. The slope of the enumeration function is inversely proportional to the temperature. The excitation profile gives the depth of the inherent structures sampled preferentially at a given temperature. At the Kauzmann temperature T_K the system attains the state of a configurationally unique ideal glass ($\sigma=0$), corresponding to the deepest amorphous basin (see Figs 5 and 8) and its inherent structure energy does not therefore change upon further cooling.

supercooled liquid and its stable crystal are equal⁹. For many fragile glass-formers the VTF temperature of structural arrest, T_o , is very close to T_K obtained from calorimetric measurements (typically⁴⁹ $0.9 < T_K/T_o < 1.1$). This again indicates a connection between dynamics and thermodynamics not present at higher temperatures. Equally suggestive is the correspondence between kinetic fragilities based on the temperature dependence of the viscosity (see Fig. 2) and thermodynamic fragilities⁵⁴, based on the temperature dependence of the entropy surplus of the supercooled liquid with respect to its stable crystal.

The energy landscape

A convenient framework for interpreting the complex phenomenology just described is provided by the energy landscape⁴⁴. This is the name generally given to the potential energy function of an N -body system $\Phi(r_1, \dots, r_N)$, where the vectors r_i comprise position, orientation and vibration coordinates. In condensed phases, whether liquid or solid, every molecule experiences simultaneous interactions with numerous neighbours. Under these conditions it is convenient to consider the full N -body Φ -function. The landscape is a multidimensional surface. For the simplest case of N structureless particles possessing no internal orientational and vibrational degrees of freedom, the landscape is a $(3N + 1)$ -dimensional object. Figure 5 is a schematic illustration of an energy landscape. The quantities of interest are the number of potential energy minima (also called inherent structures) of a given depth (Box 2), and the nature of the saddle points separating neighbouring minima. More than 30 years ago, Goldstein articulated a topographic viewpoint of condensed phases⁵⁵ that has come to be known as the energy landscape paradigm. His seminal ideas have since been applied to protein folding^{56–64}, the mechanical properties of glasses^{65–67}, shear-enhanced diffusion⁶⁸ and the dynamics of supercooled liquids^{69–71}.

Landscape sampling

For an N -body material system in a volume V , the landscape is fixed. The manner in which a material system samples its landscape as a

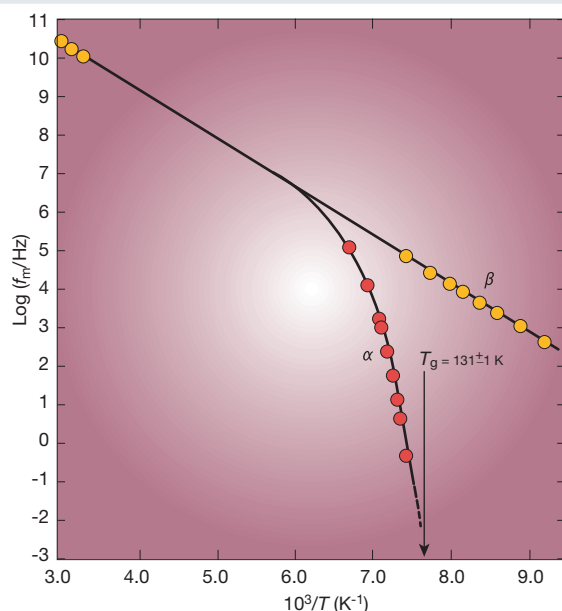


Figure 3 Temperature dependence of the peak dielectric relaxation frequency of the glass-forming mixture chlorobenzene/*cis*-decalin (molar ratio 17.2/82.8%). At high enough temperature there is a single relaxation mechanism. In the moderately supercooled regime the peak splits into slow (α) and fast (β) relaxations, of which α -processes exhibit non-Arrhenius temperature dependence and vanish at T_g . (Adapted from refs 9 and 41.)

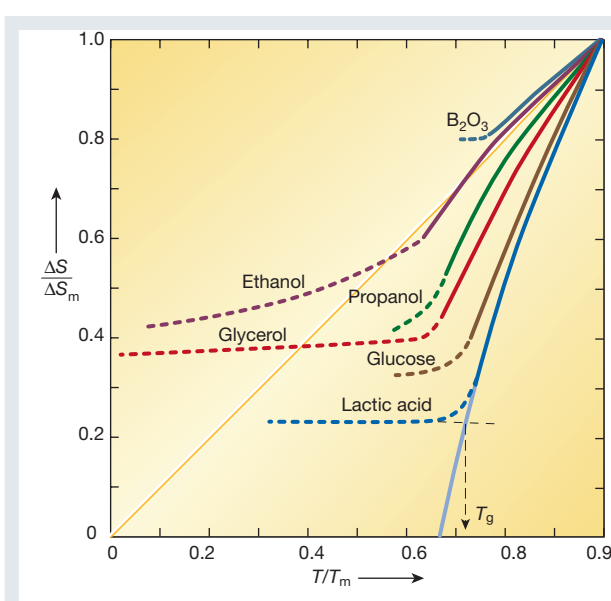


Figure 4 Temperature dependence of the entropy difference between several supercooled liquids and their stable crystals at atmospheric pressure. ΔS_m is the melting entropy and T_m is the melting temperature. The glass transition always intervenes before the vanishing of the entropy surplus. For fragile liquids such as lactic acid, however, an entropy crisis is imminent. (Adapted from ref. 45.)

function of temperature provides information on its dynamic behaviour⁷⁰. The way that a landscape deforms as a result of changes in density provides information on the mechanical properties of a material system⁷². Figure 6 shows the average inherent structure energy for a mixture of unequal-sized atoms, as a function of the temperature of the equilibrated liquid^{70,73}. In these calculations, molecular dynamics simulations of the binary mixture were performed to generate configurations. Periodically, the system's energy was minimized, yielding mechanically stable inherent structures, the average energy of which is reported in the figure. At high temperatures the inherent structure energy is virtually temperature-independent, and appears to have reached a plateau. When the system has sufficient kinetic energy to sample its entire energy landscape, the overwhelming number of minima that it samples are shallow, reflecting the fact that deep minima are very rare (Box 2). But as the reduced temperature decreases below about $T = 1$, the system is unable to surmount the highest energy barriers, and is therefore forced to sample the much rarer deeper minima (Box 2). When this happens, the kinetics of structural relaxation changes from exponential to stretched exponential, and the activation energy (and entropy) associated with structural relaxation become super-Arrhenius, that is to say they increase with decreasing temperature⁷⁰.

These calculations established a connection between changes in dynamics and the manner in which the static and thermodynamic energy landscape is sampled as a function of temperature. Figure 6 also shows that at a low enough temperature the system becomes stuck in a single minimum, the depth of which increases as the cooling rate decreases. This corresponds to the glass transition. Another important observation of this study was the existence of a temperature $T \approx 0.45$, below which the height of the barriers separating sampled inherent structures increases abruptly. This temperature was found to correspond closely to the crossover temperature predicted by mode-coupling theory (MCT; see below) for this system. Here again, it is the manner in which the system samples its landscape, not the landscape itself, that changes with temperature. (See ref. 74 for a recent, different interpretation of landscape sampling at this temperature.)

The landscape picture provides a natural separation of low-temperature molecular motion into sampling distinct potential

energy minima, and vibration within a minimum. It is possible to separate formally the corresponding configurational and vibrational contributions to a liquid's properties^{75,76}. In two important computational studies, the configurational entropy was calculated by probing systematically the statistics governing the sampling of potential energy minima^{77,78} (Box 2). Using this technique, a remarkable connection between configurational entropy and diffusion was identified in liquid water⁷⁹. One of water's distinguishing anomalies is the fact that, at sufficiently low temperature, its diffusivity increases upon compression⁸⁰. As shown in Fig. 7, diffusivity maxima are correlated strongly with configurational entropy maxima, the respective loci coinciding within numerical error.

The results shown in Fig. 7 and the success of the Adam–Gibbs equation in describing experimental data on relaxation in a wide variety of systems⁵² indicate that there exists a scaling relationship between the depth distribution of basins and the height of the saddle points along paths connecting neighbouring basins. Such scaling is not a mathematical necessity, but arises from the nature of real molecular interactions. The topographic nature of this statistical scaling relationship between minima and saddle points is poorly understood (but see the recent computational investigation of saddle points⁷⁴). Its elucidation will explain the origin of the connection between the dynamics and thermodynamics of glass-forming liquids, and constitutes the principal theoretical challenge in this field.

Strong versus fragile behaviour

The extent to which the shear viscosity η deviates from Arrhenius behaviour, $\eta = \eta_0 \exp(E/k_B T)$, constitutes the basis of the classification of liquids as either strong or fragile (Fig. 2). Molten SiO₂, often considered as the prototypical strong glass-former, displays an almost constant activation energy of 180 kcal mol⁻¹ (ref. 81). This constancy indicates that the underlying mechanism, presumably breaking and reformation of Si–O bonds, applies throughout the entire landscape⁴. In contrast, the viscosity of OTP — the canonical fragile glass-former — deviates markedly from Arrhenius behaviour⁸², showing an effective activation energy ($d \ln \eta / d(1/T)$) that increases 20-fold, from one-quarter of the heat of vaporization for the liquid above its melting point to roughly five times the heat of vaporization near T_g . This means that OTP's landscape is very heterogeneous. The basins sampled at high temperature allow relaxation by surmounting low barriers involving the rearrangement of a small number of molecules. The very large activation energy at $T \approx T_g$, on the other hand, corresponds to the cooperative rearrangement of many molecules. These differences between strong and fragile behaviour imply a corresponding topographic distinction between the two

archetypal landscapes. Aside from multiplicity due to permutational symmetry, strong landscapes may consist of a single 'megabasin', whereas fragile ones display a proliferation of well-separated 'megabasins' (Fig. 8).

Cooperative rearrangements such as those that must occur in OTP are unlikely to consist of elementary transitions between adjacent basins. Rather, the likely scenario involves a complicated sequence of elementary transitions. At low temperatures, these rearrangements should be rare and long-lived on the molecular timescale. Furthermore, the diversity of deep landscape traps and of the pathways of configuration space that connect them should result in a broad spectrum of relaxation times, as required for the stretched exponential function in equation (2). This in turn suggests that supercooled fragile liquids are dynamically heterogeneous, probably consisting at any instant of mostly non-diffusing molecules with a few 'hot spots' of mobile molecules. This dynamic heterogeneity³⁹ has both experimental^{29,30,36} and computational^{31–35} support.

The inverse relation between the self-diffusion coefficient and viscosity embodied in the Stokes–Einstein equation is based on macroscopic hydrodynamics that treats the liquid as a continuum.

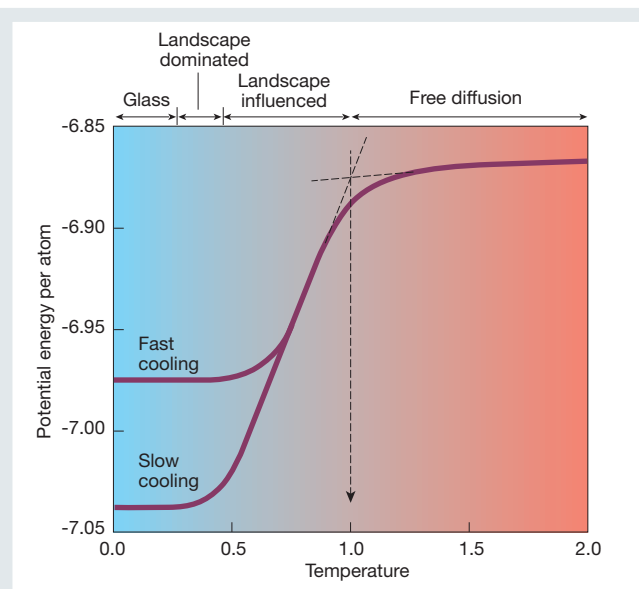


Figure 6 Mean inherent structure energy per particle of a binary mixture of unequal-sized Lennard–Jones atoms, as a function of the temperature of the equilibrated liquid from which the inherent structures were generated by energy minimization. Molecular dynamics simulations at constant energy and density were performed over a range of temperatures for 256 Lennard–Jones atoms, of which 20% are of type A and 80% are of type B. The Lennard–Jones size and energy parameters are $\sigma_{AA} = 1$, $\sigma_{BB} = 0.88$, $\sigma_{AB} = 0.8$, and $\epsilon_{AA} = 1$, $\epsilon_{BB} = 0.5$, $\epsilon_{AB} = 1.5$, respectively. Length, temperature, energy and time are expressed in units of σ_{AA} , ϵ_{AA}/k_B , ϵ_{AA} and $\sigma_{AA}(m/\epsilon_{AA})^{1/2}$, respectively, with m representing the mass of the particles. Simulations were performed at a density of 1.2. The fast and slow cooling rates are 1.08×10^{-3} and 3.33×10^{-6} . When $T > 1$, the system has sufficient kinetic energy to sample the entire energy landscape, and the overwhelming number of sampled energy minima are shallow. Under these conditions, the system exhibits a temperature-independent activation energy for structural relaxation (calculations not shown). Between $T = 1$ and $T \approx 0.45$, the activation energy increases upon cooling, the dynamics become 'landscape-influenced', and the mechanically stable configurations sampled are strongly temperature-dependent. Below $T \approx 0.45$, the height of the barriers separating sampled adjacent energy minima seems to increase abruptly (calculations not shown). This is the 'landscape-dominated' regime. In it, particles execute rare jumps over distances roughly equal to interparticle separations. The crossover between landscape-influenced and landscape-dominated behaviour corresponds closely with the mode-coupling transition temperature^{70,82}. (Adapted from refs 70 and 72.)

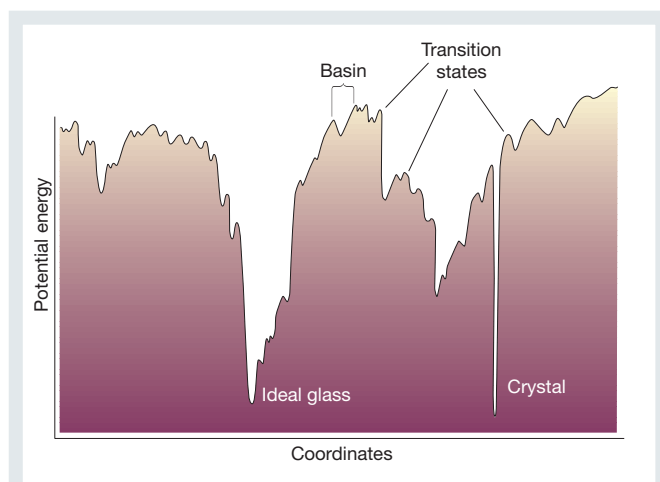


Figure 5 Schematic illustration of an energy landscape. The x-axis represents all configurational coordinates. (Adapted from ref. 44.)

This picture must clearly break down in supercooled fragile liquids, which are dynamically heterogeneous. The failure of the Stokes–Einstein equation, referred to above as one of the distinguishing characteristics of fragile supercooled liquids, is therefore qualitatively understandable. Plausible models for the low-temperature enhancement of diffusive motion relative to hydrodynamic expectations based on the viscosity have been proposed^{83–85}, but an accurate predictive theory is missing. The landscape viewpoint also provides a plausible interpretation for the α/β -relaxation decoupling shown in Fig. 3 — α -relaxations correspond to configurational sampling of neighbouring megabins (Fig. 8), whereas β -processes are thought to correspond to elementary relaxations between contiguous basins⁴⁴. Direct computational evidence of this interpretation is not available.

Avoided singularities

Alternative viewpoints to the landscape perspective have also contributed to current understanding of some aspects of supercooling and the glass transition. Two such interpretations invoke a narrowly avoided singularity above T_g .

According to MCT⁸⁶, structural arrest occurs as a result of the following feedback mechanism: (i) shear-stress relaxation occurs primarily through diffusive motion; (ii) diffusion and viscosity are inversely related; and (iii) viscosity is proportional to shear-stress relaxation time. These facts lead to a viscosity feedback whereby

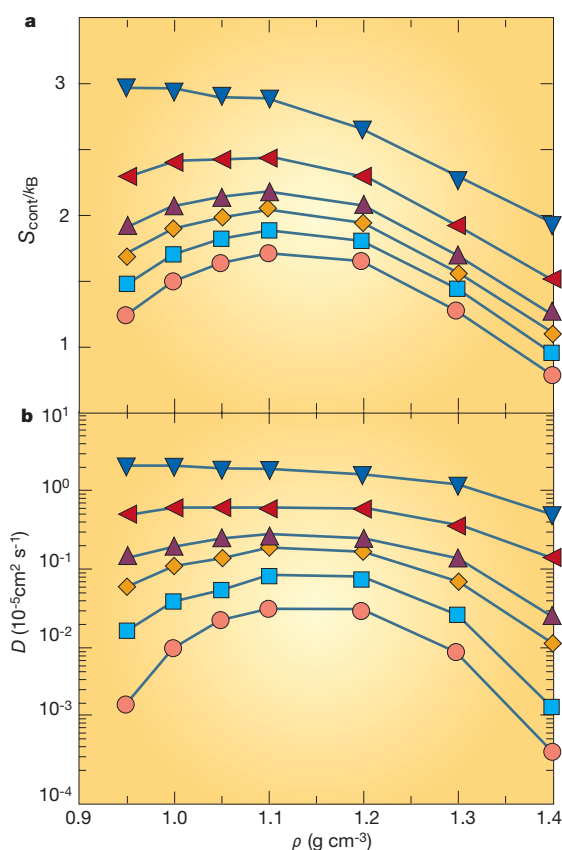
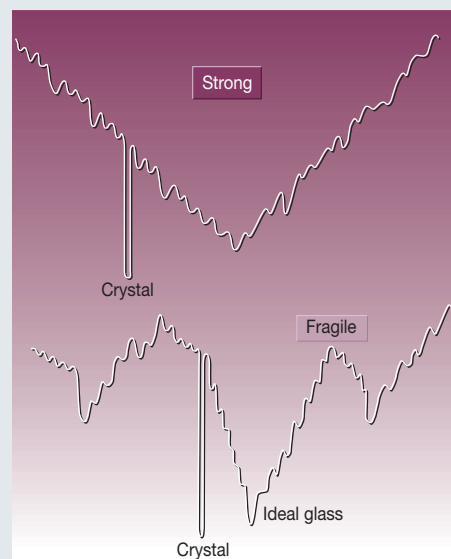


Figure 7 Relationship between diffusivity (D) and configurational entropy (S_{conf}) of supercooled water⁷⁹ at six different temperatures. Filled and open symbols from top to bottom represent the following temperatures: 300 K, 260 K, 240 K, 230 K, 220 K and 210 K; ρ is density. The configurational entropy, which is related to the number of potential energy minima of a given depth (Box 2), was calculated by subtracting the vibrational contribution from the total entropy. The calculations involved performing molecular dynamics simulations of the extended simple point charge (SPC/E) model of water¹⁰² over a range of temperatures and densities. (Adapted from ref. 79.)

Figure 8 Schematic representation of the energy landscapes of strong and fragile substances. The potential energy increases vertically, and the horizontal direction represents collective configurational coordinates.



structural arrest occurs as a purely dynamic singularity, that is to say it is not accompanied by thermodynamic signatures such as a diverging correlation length. What is now known as the idealized MCT^{87,88} predicts structural arrest to occur at a temperature T_x . Initially, therefore, it was thought that MCT was a useful theory for the laboratory-generated glass transition. It is now widely understood that this is not the case, as one finds that $T_x > T_g$, and the MCT-predicted singularity does not occur. In subsequent modifications of the theory⁸⁹, additional relaxation mechanisms occur, often referred to as ‘hopping’ or activated motions, which restore ergodicity (the system’s ability to sample all configurations) below T_x , thereby avoiding a kinetic singularity. These additional relaxation modes arise as a result of a coupling between fluctuations in density and momentum.

Although not a theory of the glass transition, MCT accurately describes many important aspects of relaxation dynamics in liquids above or moderately below their melting temperatures. In particular, the theory makes detailed predictions about the behaviour of the intermediate scattering function F , an experimentally observable quantity that measures the decay of density fluctuations. After a fast initial decay due to microscopic intermolecular collisions, MCT predicts that the decay of F obeys the following sequence (Fig. 9): (i) power-law decay towards a plateau, according to $F = f + At^{-a}$; (ii) a second power-law decay away from the plateau value $F = f - Bt^b$; and (iii) slow relaxation at longer times, which can be fitted by the KWW function $F = \exp[-(t/\tau)^\beta]$. Here, f is the plateau value of the scattering function, which only appears at sufficiently low temperature; t is time; A , B , a and b are constants; τ is the characteristic, temperature-dependent relaxation time; and $\beta < 1$ is the KWW stretch exponent. The basic accuracy of these detailed predictions has been verified experimentally and in computer simulations^{90–92}.

Kivelson and co-workers have proposed a theory of supercooled liquids that is based also on an avoided singularity^{24,93–95}. According to this viewpoint, the liquid has an energetically preferred local structure that differs from the structure in the actual crystalline phase. The system is prevented from crystallizing into a reference crystal with the preferred local structure because of geometric frustration owing to the fact that the latter does not tile space. An example of such energetically favoured but non-space-tiling local structure is the icosahedral packing seen in computer simulations of the supercooled Lennard–Jones liquid⁷³. At a temperature T^* the system would, but for frustration, crystallize into the reference crystal. Instead, strain build-up causes the system to break up into frustration-limited domains, thereby avoiding a phase transition (singularity) at T^* . The avoided transition temperature T^* acts as a critical point, below which two length scales emerge, both of which are large compared to

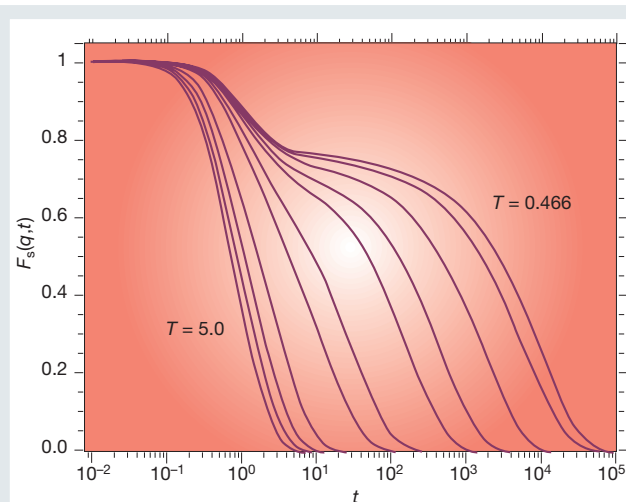


Figure 9 Evolution of the self-intermediate scattering function for A-type atoms for the same supercooled Lennard–Jones mixture as in Fig. 6, at $q\sigma_{AA} = 7.251$, corresponding to the first peak of the static structure factor of species A (ref. 92). Here q is the magnitude of the wave vector. Temperature and time are in units of ϵ_{AA}/k_B and $\sigma_{AA}(m/48\epsilon_{AA})^{1/2}$, respectively. Temperatures from left to right are 5, 4, 3, 2, 1, 0.8, 0.6, 0.55, 0.5, 0.475, and 0.466. The self-intermediate scattering function is the space Fourier transform of the van Hove function $G_s(r, t)$, which is proportional to the probability of observing a particle at $r \pm dr$ at time t given that the same particle was at the origin at $t = 0$. Note the two-step relaxation behaviour upon decreasing T . Molecular dynamics simulations of 1,000 atoms. (Adapted from refs 9 and 92.)

molecular dimensions. One is the critical correlation length, which governs density fluctuations in the absence of frustration. The second is the frustration-limited domain size. From these considerations there emerge predictions on the temperature dependence of the viscosity. Experimental data analysed according to the theory display universality²⁴, but at the expense of introducing a number of fitting parameters. The improvement with respect to competing interpretations is a matter of controversy²⁵.

Challenges and open questions

Important aspects of the complex behaviour of viscous liquids close to the glass transition can be explained qualitatively from the energy landscape perspective. Making this descriptive picture quantitative and predictive is a major challenge. This will require investigating how basic landscape features such as the basin enumeration function depend on molecular architecture and, for a given substance or mixture, on density (see ref. 96 for an example of such a calculation). Equally important is the translation of qualitative pictures such as Fig. 8 into precise measures of strength and fragility based on the basin enumeration function. Uncovering the topographic nature of the scaling relationship between basin minima and saddle points holds the key to understanding the relationship between kinetics and thermodynamics in deeply supercooled liquids. All of these calculations are in principle straightforward, but computationally at the very limit of what is currently feasible. The development of theoretical models⁹⁷ is therefore of paramount importance.

MCT and the landscape perspective offer complementary viewpoints of the same phenomena. So far, however, not enough effort has been devoted to bridging the gap that separates these two approaches. Recent calculations^{70,74} offer the promise of establishing a clearer connection between the static landscape viewpoint and the dynamic perspective of MCT. At the least, what is required is a precise landscape-based explanation of what ‘hopping’ and ‘activated processes’ really mean. Additional theoretical viewpoints of supercooling and the glass transition include the instantaneous normal-mode perspective on liquid dynamics⁹⁸ and thermodynamic

treatments of the vitreous state based on invoking analogies to spin glasses^{99–101}. Establishing a coherent theoretical perspective on supercooled liquids and glasses is important. We believe that the landscape formalism offers the natural technical tools for accomplishing this task.

- Angell, C. A. Formation of glasses from liquids and biopolymers. *Science* **267**, 1924–1935 (1995).
- Blanshard, J. M. V. & Lillford, P. (eds) *The Glassy State in Foods* (Nottingham Univ. Press, Nottingham, 1993).
- Crowe, J. H., Carpenter, J. F. & Crowe, L. M. The role of vitrification in anhydrobiosis. *Annu. Rev. Physiol.* **60**, 73–103 (1998).
- Debenedetti, P. G., Stillinger, F. H., Truskett, T. M. & Lewis, C. P. Theory of supercooled liquids and glasses: energy landscape and statistical geometry perspectives. *Adv. Chem. Eng.* (in the press).
- Greer, A. L. Metallic glasses. *Science* **267**, 1947–1953 (1995).
- Jenniskens, P. & Blake, D. F. Structural transitions in amorphous water ice and astrophysical implications. *Science* **265**, 753–756 (1994).
- Anderson, P. W. Through a glass lightly. *Science* **267**, 1615 (1995).
- Angell, C. A., Ngai, K. L., McKenna, G. B., McMillan, P. F. & Martin, S. W. Relaxation in glass-forming liquids and amorphous solids. *J. Appl. Phys.* **88**, 3113–3157 (2000).
- Debenedetti, P. G. *Metastable Liquids. Concepts and Principles* (Princeton Univ. Press, Princeton, 1996).
- Turnbull, D. Under what conditions can a glass be formed? *Contemp. Phys.* **10**, 473–488 (1969).
- Angell, C. A. Structural instability and relaxation in liquid and glassy phases near the fragile liquid limit. *J. Non-Cryst. Solids* **102**, 205–221 (1988).
- Moynihan, C. T. et al. in *The Glass Transition and the Nature of the Glassy State* (eds Goldstein, M. & Simha, R.) *Ann. NY Acad. Sci.* **279**, 15–36 (1976).
- Brünig, R. & Samwer, K. Glass transition on long time scales. *Phys. Rev. B* **46**, 318–322 (1992).
- Ediger, M. D., Angell, C. A. & Nagel, S. R. Supercooled liquids and glasses. *J. Phys. Chem.* **100**, 13200–13212 (1996).
- Vogel, H. Das temperatur-abhängigkeitsgesetz der viskosität von flüssigkeiten. *Phys. Zeit.* **22**, 645–646 (1921).
- Tammann, G. & Hesse, W. Die abhängigkeit der viskosität von der temperatur bei unterkühlten flüssigkeiten. *Z. Anorg. Allg. Chem.* **156**, 245–257 (1926).
- Fulcher, G. S. Analysis of recent measurements of the viscosity of glasses. *J. Am. Ceram. Soc.* **8**, 339 (1925).
- Laughlin, W. T. & Uhlmann, D. R. Viscous flow in simple organic liquids. *J. Phys. Chem.* **76**, 2317–2325 (1972).
- Angell, C. A. in *Relaxations in Complex Systems* (eds Ngai, K. & Wright, G. B.) **1** (Nat'l Technol. Inform. Ser., US Dept. of Commerce, Springfield, VA, 1985).
- Angell, C. A. Relaxation in liquids, polymers and plastic crystals—strong/fragile patterns and problems. *J. Non-Cryst. Solids* **131–133**, 13–31 (1991).
- Green, J. L., Ito, K., Xu, K. & Angell, C. A. Fragility in liquids and polymers: new, simple quantifications and interpretations. *J. Phys. Chem. B* **103**, 3991–3996 (1999).
- Novikov, V. N., Rössler, E., Malinovsky, V. K. & Surovstev, N. V. Strong and fragile liquids in percolation approach to the glass transition. *Europhys. Lett.* **35**, 289–294 (1996).
- Fujimori, H. & Oguni, M. Correlation index $(T_g - T_{\infty})/T_g$ and activation energy ratio $\Delta E_{\infty}/\Delta E_{\text{eff}}$ as parameters characterizing the structure of liquid and glass. *Solid State Commun.* **94**, 157–162 (1995).
- Kivelson, D., Tarjus, G., Zhao, X. & Kivelson, S. A. Fitting of viscosity: distinguishing the temperature dependencies predicted by various models of supercooled liquids. *Phys. Rev. E* **53**, 751–758 (1996).
- Cummins, H. Z. Comment on “Fitting of viscosity: distinguishing the temperature dependencies predicted by various models of supercooled liquids”. *Phys. Rev. E* **54**, 5870–5872 (1996).
- Kohlrausch, R. Theorie des elektrischen rückstandes in der leidener flasche. *Ann. Phys. Chem. (Leipzig)* **91**, 179–214 (1874).
- Williams, G. & Watts, D. C. Non-symmetrical dielectric relaxation behavior arising from a simple empirical decay function. *Trans. Faraday Soc.* **66**, 80–85 (1970).
- Richert, R. & Blumen, A. in *Disorder Effects on Relaxational Processes* (eds Richert, R. & Blumen, A.) **1–7** (Springer, Berlin, 1994).
- Cicerone, M. T. & Ediger, M. D. Relaxation of spatially heterogeneous dynamic domains in supercooled ortho-terphenyl. *J. Chem. Phys.* **103**, 5684–5692 (1995).
- Cicerone, M. T. & Ediger, M. D. Enhanced translation of probe molecules in supercooled o-terphenyl: signature of spatially heterogeneous dynamics? *J. Chem. Phys.* **104**, 7210–7218 (1996).
- Me’cuk, A. I., Ramos, R. A., Gould, H., Klein, W. & Mountain, R. D. Long-lived structures in fragile glass-forming liquids. *Phys. Rev. Lett.* **75**, 2522–2525 (1995).
- Hurley, M. M. & Harrowell, P. Non-gaussian behavior and the dynamical complexity of particle motion in a dense two-dimensional liquid. *J. Chem. Phys.* **105**, 10521–10526 (1996).
- Perera, D. N. & Harrowell, P. Measuring diffusion in supercooled liquids: the effect of kinetic inhomogeneities. *J. Chem. Phys.* **104**, 2369–2375 (1996).
- Perera, D. N. & Harrowell, P. Consequence of kinetic inhomogeneities in glasses. *Phys. Rev. E* **54**, 1652–1662 (1996).
- Donati, C., Glotzer, S. C., Poole, P. H., Kob, W. & Plimpton, S. J. Spatial correlations of mobility and immobility in a glass-forming Lennard–Jones liquid. *Phys. Rev. E* **60**, 3107–3119 (1999).
- Böhmer, R., Hinze, G., Diezemann, G., Geil, B. & Sillescu, H. Dynamic heterogeneity on supercooled ortho-terphenyl studied by multidimensional deuterium NMR. *Europhys. Lett.* **36**, 55–60 (1996).
- Wang, C.-Y. & Ediger, M. D. How long do regions of different dynamics persist in supercooled o-terphenyl? *J. Phys. Chem. B* **103**, 4177–4184 (1999).
- Vidal Russell, E. & Israeloff, N. E. Direct observation of molecular cooperativity near the glass transition. *Nature* **408**, 695–698 (2000).
- Ediger, M. D. Spatially heterogeneous dynamics in supercooled liquids. *Annu. Rev. Phys. Chem.* **51**, 99–128 (2000).
- Fujara, F., Geil, B., Sillescu, H. H. & Fleischer, G. Translational and rotational diffusion in supercooled ortho-terphenyl close to the glass transition. *Z. Phys. B Cond. Matt.* **88**, 195–204 (1992).
- Johari, G. P. Intrinsic mobility of molecular glasses. *J. Chem. Phys.* **58**, 1766–1770 (1973).
- Johari, G. P. & Goldstein, M. Viscous liquids and the glass transition. II. Secondary relaxations in glasses of rigid molecules. *J. Chem. Phys.* **53**, 2372–2388 (1970).

43. Rössler, E., Warschewski, U., Eiermann, P., Sokolov, A. P. & Quitmann, D. Indications for a change of transport mechanism in supercooled liquids and the dynamics close and below T_g . *J. Non-Cryst. Solids* **172–174**, 113–125 (1994).
44. Stillinger, F. H. A topographic view of supercooled liquids and glass formation. *Science* **267**, 1935–1939 (1995).
45. Kauzmann, W. The nature of the glassy state and the behavior of liquids at low temperatures. *Chem. Rev.* **43**, 219–256 (1948).
46. Simon, F. Über den Zustand der unterkühlten Flüssigkeiten und Gläser. *Z. Anorg. Allg. Chem.* **203**, 219–227 (1931).
47. Wolynes, P. G. Aperiodic crystals: biology, chemistry and physics in a fugue with stretto. *AIP Conf. Proc.* **180**, 39–65 (1988).
48. Wolynes, P. G. Entropy crises in glasses and random heteropolymers. *J. Res. Natl. Inst. Standards Technol.* **102**, 187–194 (1997).
49. Angell, C. A. Landscapes with megabasins: polyamorphism in liquids and biopolymers and the role of nucleation in folding and folding diseases. *Physica D* **107**, 122–142 (1997).
50. Gibbs, J. H. & DiMarzio, E. A. Nature of the glass transition and the glassy state. *J. Chem. Phys.* **28**, 373–383 (1958).
51. Adam, G. & Gibbs, J. H. On the temperature dependence of cooperative relaxation properties in glass-forming liquids. *J. Chem. Phys.* **43**, 139–146 (1965).
52. Richert, R. & Angell, C. A. Dynamics of glass-forming liquids. V. On the link between molecular dynamics and configurational entropy. *J. Chem. Phys.* **108**, 9016–9026 (1998).
53. Williams, M. L., Landel, R. F. & Ferry, J. D. The temperature dependence of the relaxation mechanisms in amorphous polymers and other glass-forming liquids. *J. Am. Chem. Soc.* **77**, 3701–3707 (1955).
54. Ito, K., Moynihan, C. T. & Angell, C. A. Thermodynamic determination of fragility in liquids and a fragile-to-strong liquid transition in water. *Nature* **398**, 492–495 (1999).
55. Goldstein, M. Viscous liquids and the glass transition: a potential energy barrier picture. *J. Chem. Phys.* **51**, 3728–3739 (1969).
56. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. The energy landscapes and motions of proteins. *Science* **254**, 1598–1603 (1991).
57. Nienhaus, G. U., Müller, J. D., McMahon, B. H. & Frauenfelder, H. Exploring the conformational energy landscape of proteins. *Physica D* **107**, 297–311 (1997).
58. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. Free energy landscape for protein folding kinetics: intermediates, traps, and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.* **101**, 6052–6062 (1994).
59. Saven, J. G., Wang, J. & Wolynes, P. G. Kinetics of protein folding: the dynamics of globally connected rough energy landscapes with biases. *J. Chem. Phys.* **101**, 11037–11043 (1994).
60. Wang, J., Onuchic, J. & Wolynes, P. Statistics of kinetic pathways on biased rough energy landscapes with applications to protein folding. *Phys. Rev. Lett.* **76**, 4861–4864 (1996).
61. Plotkin, S. S., Wang, J. & Wolynes, P. G. Correlated energy landscape model for finite, random heteropolymers. *Phys. Rev. E* **53**, 6271–6296 (1996).
62. Becker, O. M. & Karplus, M. The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics. *J. Chem. Phys.* **106**, 1495–1517 (1997).
63. Dill, K. A. & Chan, H. S. From Levinthal to pathways and funnels. *Nature Struct. Biol.* **4**, 10–19 (1997).
64. Klepeis, J. L., Floudas, C. A., Morikis, D. & Lambris, J. D. Predicting peptide structure using NMR data and deterministic global optimization. *J. Comp. Chem.* **20**, 1354–1370 (1999).
65. Lacks, D. J. Localized mechanical instabilities and structural transformations in silica glass under high pressure. *Phys. Rev. Lett.* **80**, 5385–5388 (1998).
66. Malandro, D. L. & Lacks, D. J. Volume dependence of potential energy landscapes in glasses. *J. Chem. Phys.* **107**, 5804–5810 (1997).
67. Malandro, D. L. & Lacks, D. J. Relationships of shear-induced changes in the potential energy landscape to the mechanical properties of ductile glasses. *J. Chem. Phys.* **110**, 4593–4601 (1999).
68. Malandro, D. L. & Lacks, D. J. Molecular-level instabilities and enhanced self-diffusion in flowing liquids. *Phys. Rev. Lett.* **81**, 5576–5579 (1998).
69. Schulz, M. Energy landscape, minimum points, and non-Arrhenius behavior of supercooled liquids. *Phys. Rev. B* **57**, 11319–11333 (1998).
70. Sastry, S., Debenedetti, P. G. & Stillinger, F. H. Signatures of distinct dynamical regimes in the energy landscape of a glass-forming liquid. *Nature* **393**, 554–557 (1998).
71. Keyes, T. Dependence of supercooled liquid dynamics on elevation in the energy landscape. *Phys. Rev. E* **59**, 3207–3211 (1999).
72. Debenedetti, P. G., Stillinger, F. H., Truskett, T. M. & Roberts, C. J. The equation of state of an energy landscape. *J. Phys. Chem. B* **103**, 7390–7397 (1999).
73. Jonsson, H. & Andersen, H. C. Icosahedral ordering in the Lennard-Jones crystal and glass. *Phys. Rev. Lett.* **60**, 2295–2298 (1988).
74. Angelani, L., Di Leonardo, R., Ruocco, G., Scala, A. & Sciortino, F. Saddles in the energy landscape probed by supercooled liquids. *Phys. Rev. Lett.* **85**, 5356–5359 (2000).
75. Stillinger, F. H., Debenedetti, P. G. & Sastry, S. Resolving vibrational and structural contributions to isothermal compressibility. *J. Chem. Phys.* **109**, 3983–3988 (1998).
76. Stillinger, F. H. & Debenedetti, P. G. Distinguishing vibrational and structural equilibration contributions to thermal expansion. *J. Phys. Chem. B* **103**, 4052–4059 (1999).
77. Sciortino, F., Kob, W. & Tartaglia, P. Inherent structure entropy of supercooled liquids. *Phys. Rev. Lett.* **83**, 3214–3217 (1999).
78. Büchner, S. & Heuer, A. Potential energy landscape of a model glass former: thermodynamics, anharmonicities, and finite size effects. *Phys. Rev. E* **60**, 6507–6518 (1999).
79. Scala, A., Starr, F. W., La Nave, E., Sciortino, F. & Stanley, H. E. Configurational entropy and diffusivity in supercooled water. *Nature* **406**, 166–169 (2000).
80. Prielmeier, F. X., Lang, E. W., Speedy, R. J. & Lüdemann, H.-D. Diffusion in supercooled water to 300 Mpa. *Phys. Rev. Lett.* **59**, 1128–1131 (1987).
81. Mackenzie, J. D. Viscosity-temperature relationship for network liquids. *J. Am. Ceram. Soc.* **44**, 598–601 (1961).
82. Greet, R. J. & Turnbull, D. Glass transition in o-terphenyl. *J. Chem. Phys.* **46**, 1243–1251 (1967).
83. Stillinger, F. H. & Hodgdon, J. A. Translation-rotation paradox for diffusion in fragile glass-forming liquids. *Phys. Rev. E* **50**, 2064–2068 (1994).
84. Tarjus, G. & Kivelson, D. Breakdown of the Stokes-Einstein relation in supercooled liquids. *J. Chem. Phys.* **103**, 3071–3073 (1995).
85. Liu, C. Z.-W. & Openheim, I. Enhanced diffusion upon approaching the kinetic glass transition. *Phys. Rev. E* **53**, 799–802 (1996).
86. Gesztzi, T. Pre-vitrification by viscosity feedback. *J. Phys. C* **16**, 5805–5814 (1983).
87. Bengtzelius, U., Götz, W. & Sjölander, A. Dynamics of supercooled liquids and the glass transition. *J. Phys. C* **17**, 5915–5934 (1984).
88. Götz, W. & Sjögren, L. Relaxation processes in supercooled liquids. *Rep. Prog. Phys.* **55**, 241–376 (1992).
89. Götz, W. & Sjögren, L. The mode coupling theory of structural relaxations. *Transp. Theory Stat. Phys.* **24**, 801–853 (1995).
90. Götz, W. Recent tests of the mode-coupling theory for glassy dynamics. *J. Phys. Cond. Matt.* **11**, A1–A45 (1999).
91. Kob, W. Computer simulations of supercooled liquids and glasses. *J. Phys. Cond. Matt.* **11**, R85–R115 (1999).
92. Kob, W. & Andersen, H. C. Testing mode-coupling theory for a supercooled binary Lennard-Jones mixture: the van Hove correlation function. *Phys. Rev. E* **51**, 4626–4641 (1995).
93. Kivelson, D., Kivelson, S. A., Zhao, X., Nussinov, Z. & Tarjus, G. A thermodynamic theory of supercooled liquids. *Physica A* **219**, 27–38 (1995).
94. Kivelson, D. & Tarjus, G. SuperArrhenius character of supercooled glass-forming liquids. *J. Non-Cryst. Solids* **235–237**, 86–100 (1998).
95. Kivelson, D. & Tarjus, G. The Kauzmann paradox interpreted via the theory of frustration-limited domains. *J. Chem. Phys.* **109**, 5481–5486 (1998).
96. Sastry, S. The relationship between fragility, configurational entropy and the potential energy landscape of glass-forming liquids. *Nature* **409**, 164–167 (2001).
97. Speedy, R. J. Relations between a liquid and its glasses. *J. Phys. Chem. B* **103**, 4060–4065 (1999).
98. Keyes, T. Instantaneous normal mode approach to liquid state dynamics. *J. Phys. Chem. A* **101**, 2921–2930 (1997).
99. Kirkpatrick, T. R. & Wolynes, P. G. Stable and metastable states in mean-field Potts and structural glasses. *Phys. Rev. B* **36**, 8552–8564 (1987).
100. Kirkpatrick, T. R., Thirumalai, D. & Wolynes, P. G. Scaling concepts for the dynamics of viscous liquids near an ideal glassy state. *Phys. Rev. A* **40**, 1045–1054 (1989).
101. Mézard, M. & Parisi, G. Thermodynamics of glasses: a first principles computation. *Phys. Rev. Lett.* **82**, 747–750 (1999).
102. Berendsen, H. J., Grigera, J. R. & Stroatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **91**, 6269–6271 (1987).
103. Stillinger, F. H. Supercooled liquids, glass transitions, and the Kauzmann paradox. *J. Chem. Phys.* **88**, 7818–7825 (1988).
104. Santen, L. & Krauth, W. Absence of thermodynamic phase transition in a model glass former. *Nature* **405**, 550–551 (2000).
105. Wilks, J. *The Properties of Liquid and Solid Helium* (Clarendon, Oxford, 1967).
106. Rastogi, S., Höhne, G. W. H. & Keller, A. Unusual pressure-induced phase behavior in crystalline Poly(4-methylpentene-1): calorimetric and spectroscopic results and further implications. *Macromolecules* **32**, 8897–8909 (1999).
107. Greer, A. L. Too hot to melt. *Nature* **404**, 134–135 (2000).
108. Stillinger, F. H. Exponential multiplicity of inherent structures. *Phys. Rev. E* **59**, 48–51 (1999).
109. Stillinger, F. H. Enumeration of isobaric inherent structures for the fragile glass former o-terphenyl. *J. Phys. Chem. B* **102**, 2807–2810 (1998).

Acknowledgements

P.G.D.'s work is supported by the US Department of Energy.

Exploring complex networks

Steven H. Strogatz

Department of Theoretical and Applied Mechanics and Center for Applied Mathematics, 212 Kimball Hall, Cornell University, Ithaca, New York 14853-1503, USA (e-mail: strogatz@cornell.edu)

The study of networks pervades all of science, from neurobiology to statistical physics. The most basic issues are structural: how does one characterize the wiring diagram of a food web or the Internet or the metabolic network of the bacterium *Escherichia coli*? Are there any unifying principles underlying their topology? From the perspective of nonlinear dynamics, we would also like to understand how an enormous network of interacting dynamical systems — be they neurons, power stations or lasers — will behave collectively, given their individual dynamics and coupling architecture. Researchers are only now beginning to unravel the structure and dynamics of complex networks.

Networks are on our minds nowadays. Sometimes we fear their power — and with good reason. On 10 August 1996, a fault in two power lines in Oregon led, through a cascading series of failures, to blackouts in 11 US states and two Canadian provinces, leaving about 7 million customers without power for up to 16 hours¹. The Love Bug worm, the worst computer attack to date, spread over the Internet on 4 May 2000 and inflicted billions of dollars of damage worldwide.

In our lighter moments we play parlour games about connectivity. 'Six degrees of Marlon Brando' broke out as a nationwide fad in Germany, as readers of *Die Zeit* tried to connect a falafel vendor in Berlin with his favourite actor through the shortest possible chain of acquaintances². And during the height of the Lewinsky scandal, the *New York Times* printed a diagram³ of the famous people within 'six degrees of Monica'.

Meanwhile scientists have been thinking about networks too. Empirical studies have shed light on the topology of food webs^{4,5}, electrical power grids, cellular and metabolic networks^{6–9}, the World-Wide Web¹⁰, the Internet backbone¹¹, the neural network of the nematode worm *Caenorhabditis elegans*¹², telephone call graphs¹³, coauthorship and citation networks of scientists^{14–16}, and the quintessential 'old-boy' network, the overlapping boards of directors of the largest companies in the United States¹⁷ (Fig. 1). These databases are now easily accessible, courtesy of the Internet. Moreover, the availability of powerful computers has made it feasible to probe their structure; until recently, computations involving million-node networks would have been impossible without specialized facilities.

Why is network anatomy so important to characterize? Because structure always affects function. For instance, the topology of social networks affects the spread of information and disease, and the topology of the power grid affects the robustness and stability of power transmission.

From this perspective, the current interest in networks is part of a broader movement towards research on complex systems. In the words of E. O. Wilson¹⁸, "The greatest challenge today, not just in cell biology and ecology but in all of science, is the accurate and complete description of complex systems. Scientists have broken down many kinds of systems. They think they know most of the elements and forces. The next task is to reassemble them, at least in mathematical models that capture the key properties of the entire ensembles."

But networks are inherently difficult to understand, as the following list of possible complications illustrates.

1. Structural complexity: the wiring diagram could be an intricate tangle (Fig. 1).
2. Network evolution: the wiring diagram could change over time. On the World-Wide Web, pages and links are created and lost every minute.
3. Connection diversity: the links between nodes could have different weights, directions and signs. Synapses in

Box 1

Nonlinear dynamics: terminology and concepts⁹⁷

Dynamical systems can often be modelled by differential equations $d\mathbf{x}/dt = \mathbf{v}(\mathbf{x})$, where $\mathbf{x}(t) = (x_1(t), \dots, x_n(t))$ is a vector of state variables, t is time, and $\mathbf{v}(\mathbf{x}) = (v_1(\mathbf{x}), \dots, v_n(\mathbf{x}))$ is a vector of functions that encode the dynamics. For example, in a chemical reaction, the state variables represent concentrations. The differential equations represent the kinetic rate laws, which usually involve nonlinear functions of the concentrations.

Such nonlinear equations are typically impossible to solve analytically, but one can gain qualitative insight by imagining an abstract n -dimensional state space with axes x_1, \dots, x_n . As the system evolves, $\mathbf{x}(t)$ flows through state space, guided by the 'velocity' field $d\mathbf{x}/dt = \mathbf{v}(\mathbf{x})$ like a speck carried along in a steady, viscous fluid.

Suppose $\mathbf{x}(t)$ eventually comes to rest at some point \mathbf{x}^* . Then the velocity must be zero there, so we call \mathbf{x}^* a fixed point. It corresponds to an equilibrium state of the physical system being modelled. If all small disturbances away from \mathbf{x}^* damp out, \mathbf{x}^* is called a stable fixed point — it acts as an attractor for states in its vicinity.

Another long-term possibility is that $\mathbf{x}(t)$ flows towards a closed loop and eventually circulates around it forever. Such a loop is called a limit cycle. It represents a self-sustained oscillation of the physical system.

A third possibility is that $\mathbf{x}(t)$ might settle onto a strange attractor, a set of states on which it wanders forever, never stopping or repeating. Such erratic, aperiodic motion is considered chaotic if two nearby states flow away from each other exponentially fast. Long-term prediction is impossible in a real chaotic system because of this exponential amplification of small uncertainties or measurement errors

the nervous system can be strong or weak, inhibitory or excitatory.

4. Dynamical complexity: the nodes could be nonlinear dynamical systems. In a gene network or a Josephson junction array, the state of each node can vary in time in complicated ways.
5. Node diversity: there could be many different kinds of nodes. The biochemical network that controls cell division in mammals consists of a bewildering variety of substrates and enzymes⁶, only a few of which are shown in Fig. 1c.
6. Meta-complication: the various complications can influence each other. For example, the present layout of a power grid depends on how it has grown over the years — a case where network evolution (2) affects topology (1). When coupled neurons fire together repeatedly, the connection between them is strengthened; this is the basis of memory and learning. Here nodal dynamics (4) affect connection weights (3).

To make progress, different fields have suppressed certain complications while highlighting others. For instance, in nonlinear dynamics we have tended to favour simple, nearly identical

dynamical systems coupled together in simple, geometrically regular ways. Furthermore we usually assume that the network architecture is static. These simplifications allow us to sidestep any issues of structural complexity and to concentrate instead on the system's potentially formidable dynamics.

Laser arrays provide a concrete example^{19–24}. In the single-mode approximation, each laser is characterized by its time-dependent gain, polarization, and the phase and amplitude of its electric field. These evolve according to four coupled, nonlinear differential equations. We usually hope the laser will settle down to a stable state, corresponding to steady emission of light, but periodic pulsations and even chaotic intensity fluctuations can occur in some cases¹⁹. Now suppose that many identical lasers are arranged side by side in a regular chain²⁰ or ring²¹, interacting with their neighbours by evanescent coupling or by overlap of their electric fields²². Will the lasers lock their phases together spontaneously, or break up into a standing wave pattern, or beat each other into incoherence? From a technological standpoint, self-synchronization would be the most desirable outcome, because a perfectly coherent array of N lasers would

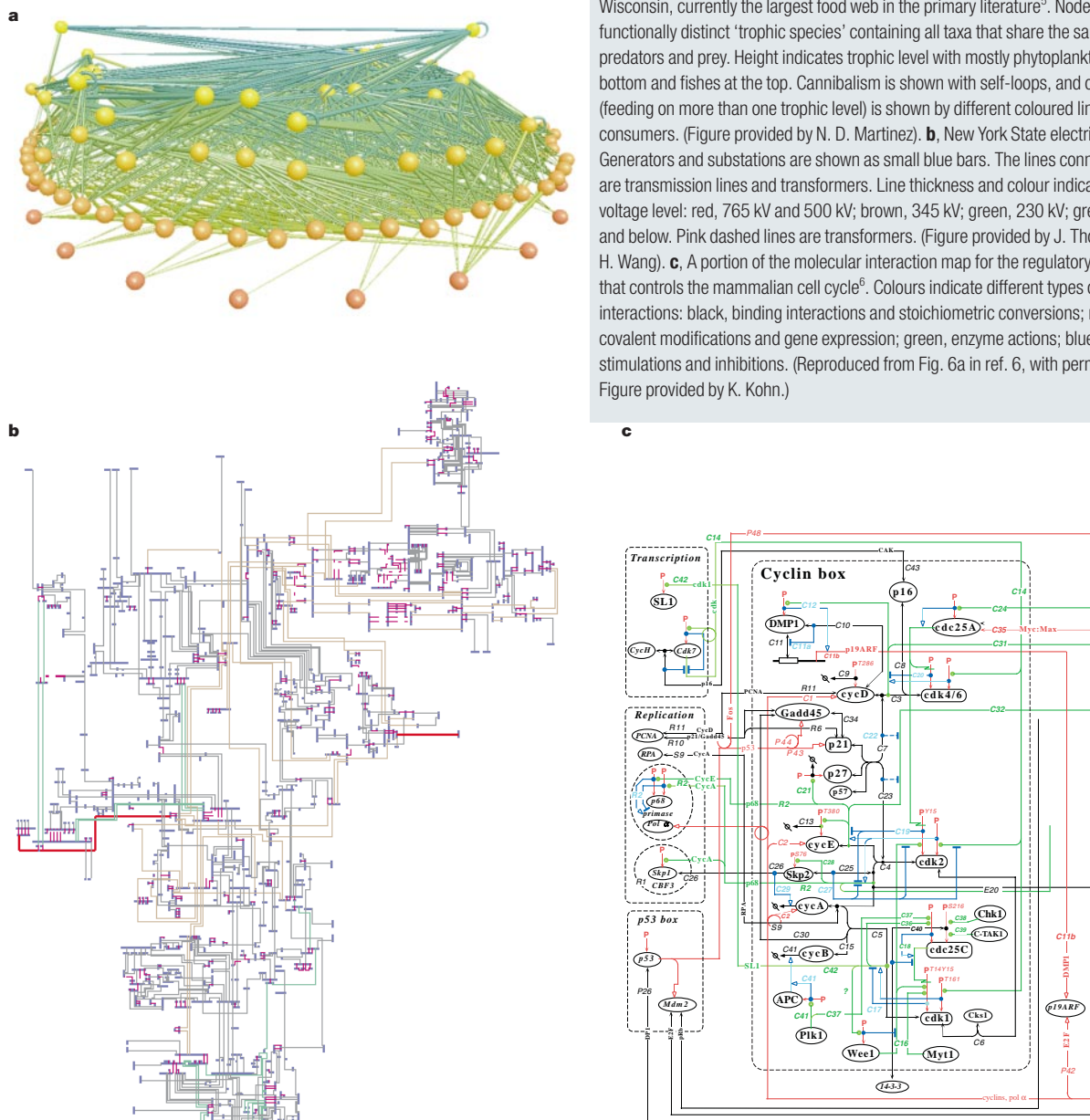


Figure 1 Wiring diagrams for complex networks. **a**, Food web of Little Rock Lake, Wisconsin, currently the largest food web in the primary literature⁵. Nodes are functionally distinct 'trophic species' containing all taxa that share the same set of predators and prey. Height indicates trophic level with mostly phytoplankton at the bottom and fishes at the top. Cannibalism is shown with self-loops, and omnivory (feeding on more than one trophic level) is shown by different coloured links to consumers. (Figure provided by N. D. Martinez). **b**, New York State electric power grid. Generators and substations are shown as small blue bars. The lines connecting them are transmission lines and transformers. Line thickness and colour indicate the voltage level: red, 765 kV and 500 kV; brown, 345 kV; green, 230 kV; grey, 138 kV and below. Pink dashed lines are transformers. (Figure provided by J. Thorp and H. Wang). **c**, A portion of the molecular interaction map for the regulatory network that controls the mammalian cell cycle⁶. Colours indicate different types of interactions: black, binding interactions and stoichiometric conversions; red, covalent modifications and gene expression; green, enzyme actions; blue, stimulations and inhibitions. (Reproduced from Fig. 6a in ref. 6, with permission. Figure provided by K. Kohn.)

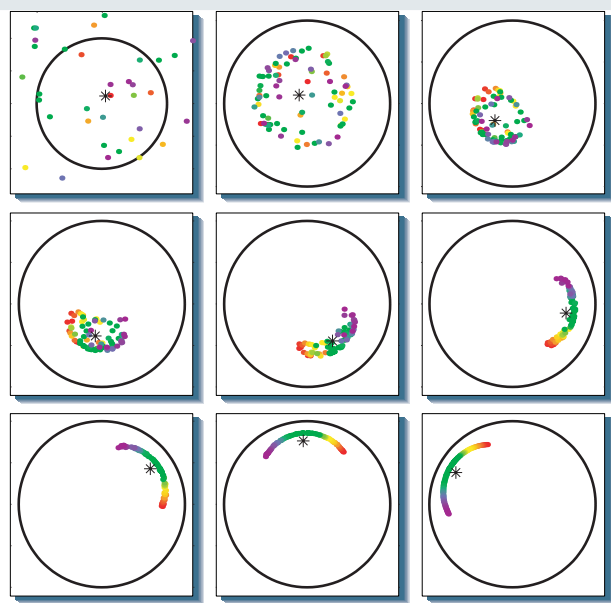


Figure 2 Spontaneous synchronization in a network of limit-cycle oscillators with distributed natural frequencies. The state of each oscillator is represented geometrically as a dot in the complex plane. The amplitude and phase of the oscillation correspond to the radius and angle of the dot in polar coordinates. Colours code the oscillators' natural frequencies, running from slowest (red) to fastest (violet). In the absence of coupling, each oscillator would settle onto its limit cycle (circle) and rotate at its natural frequency. However, here all the oscillators are also pulled towards the mean field that they generate collectively (shown as an asterisk at the centre of the population). Time increases from left to right, and from top to bottom. Starting from a random initial condition, the oscillators self-organize by collapsing their amplitudes; then they sort their phases so that the fastest oscillators are in the lead. Ultimately they all rotate as a synchronized pack, with locked amplitudes and phases. The governing equations describe a mean-field model of a laser array²³. (Simulation provided by R. Oliva.)

produce N^2 times as much power as a single one. But in practice, semiconductor laser arrays are notoriously prone to both spatial and temporal instabilities^{20,21}. Even for a simple ring geometry, this problem is dynamically complex.

The first part of this article reviews what is known about dynamical complexity in regular networks of nonlinear systems. I offer a few rules of thumb about the impact of network structure on collective dynamics, especially for arrays of coupled limit-cycle oscillators.

The logical next step would be to tackle networks that combine dynamical and structural complexity, such as power grids or ecological webs. Unfortunately they lie beyond our mathematical reach — we do not even know how to characterize their wiring diagrams. So we have to begin with network topology.

By a happy coincidence, such architectural questions are being pursued in other branches of science, thanks to the excitement about the Internet, functional genomics, financial networks, and so on. The second part of this article uses graph theory to explore the structure of complex networks, an approach that has recently led to some encouraging progress, especially when combined with the tools of statistical mechanics and computer simulations.

Needless to say, many other topics within network science deserve coverage here. The subject is amazingly rich, and apologies are offered to those readers whose favourite topics are omitted.

Regular networks of coupled dynamical systems

Networks of dynamical systems have been used to model everything from earthquakes to ecosystems, neurons to neutrinos^{25–32}. To impose some order on this list, consider the dynamics that each node would exhibit if it were isolated. Assuming it is a generic dynamical

system, its long-term behaviour is given by stable fixed points, limit cycles or chaotic attractors (Box 1).

If we now couple many such systems together, what can be said about their collective behaviour? The answer is not much — the details matter. But I will propose some rough generalizations anyway.

If the dynamical system at each node has stable fixed points and no other attractors, the network tends to lock into a static pattern. Many such patterns may coexist, especially if the nodes have competing interactions. In that case the network may become frustrated and display enormous numbers of locally stable equilibria. This kind of complex static behaviour is seen in models of spin glasses, associative memory neural networks and combinatorial optimization problems³³.

At the opposite extreme, suppose each node has a chaotic attractor. Few rules have emerged about the effect of coupling architecture on dynamics in this case. It is known that networks of identical chaotic systems can synchronize their erratic fluctuations, a curious phenomenon with possible applications to private communications^{34,35}. For a wide range of network topologies, synchronized chaos requires that the coupling be neither too weak nor too strong; otherwise spatial instabilities are triggered³⁴. Related lines of research deal with networks of identical chaotic maps, lattice dynamical systems and cellular automata. These systems have been used mainly as testbeds for exploring spatiotemporal chaos and pattern formation in the simplest mathematical settings, rather than as models of real physical systems.

Identical oscillators

The intermediate case where each node has a stable limit cycle has turned out to be particularly fruitful. Much of the research has been inspired by biological examples, ranging from the mutual synchronization of cardiac pacemaker cells, to rhythmically flashing fireflies and chorusing crickets, to wave propagation in the heart, brain, intestine and nervous system²⁵.

Arrays of identical oscillators often synchronize, or else form patterns that depend on the symmetry of the underlying network³⁶. Other common modes of organization are travelling waves in one spatial dimension, rotating spirals in two dimensions and scroll waves in three dimensions^{25,26}. For fully connected networks where each node is coupled equally to all the others, the completely synchronized state becomes likely.

These heuristics apply to systems coupled by smooth interactions akin to diffusion. But many biological oscillators communicate by sudden impulses: a neuron fires, a firefly flashes, a cricket chirps. Hence the recent interest in pulse-coupled oscillators³⁷. This thread began with Peskin's model of the sinoatrial node, the heart's natural pacemaker, as a collection of N identical integrate-and-fire oscillators³⁸. For the simple case where each oscillator is connected to all the others, Peskin conjectured that they would all end up firing in unison, no matter how they started. He gave a proof for $N=2$ oscillators; it was later demonstrated³⁹ that the conjecture holds for all N . Peskin also conjectured that synchronization would occur even if the oscillators were not quite identical, but that problem remains unproven.

Peskin's model has been used as a caricature of coupled neurons^{40–42} by including synaptic delays, refractory periods, inhibition and local coupling; these realistic features also remove some of the undesirable discontinuities in the mathematics. In an example of scientific cross-fertilization, Hopfield⁴³ pointed out that the locally coupled version of the model is closely related to slider-block models of earthquakes and should therefore display self-organized criticality. That observation suggested intriguing links among neurobiology, geophysics, synchronization and self-organized criticality, and sparked a burst of research activity, as reviewed in ref. 37.

Non-identical oscillators

While modelling populations of biological oscillators, Winfree discovered a new kind of cooperative phenomenon, the temporal

analogue of a phase transition⁴⁴. He proposed a mean-field model of nearly identical, weakly coupled limit-cycle oscillators and showed that when the coupling is small compared to the spread of natural frequencies, the system behaves incoherently, with each oscillator running at its natural frequency. As the coupling is increased, the

incoherence persists until a certain threshold is crossed — then a small cluster of oscillators suddenly ‘freezes’ into synchrony. For still greater coupling, all the oscillators become locked in phase and amplitude (Fig. 2).

Kuramoto²⁶ refined this connection between nonlinear dynamics and statistical physics. He proposed an exactly solvable model of collective synchronization, given by

$$\frac{d\theta_i}{dt} = \omega_i + \frac{K}{N} \sum_{j=1}^N \sin(\theta_j - \theta_i), \quad i = 1, \dots, N$$

where $\theta_i(t)$ is the phase of the i th oscillator and ω_i is its natural frequency, chosen at random from a lorentzian probability density

$$g(\omega) = \frac{\gamma}{\pi[\gamma^2 + (\omega - \omega_0)^2]}$$

of width γ and mean ω_0 . Using an ingenious self-consistency argument, Kuramoto solved for the order parameter

$$r(t) = \left| \frac{1}{N} \sum_{j=1}^N e^{i\theta_j(t)} \right|$$

(a convenient measure of the extent of synchronization) in the limit $N \rightarrow \infty$ and $t \rightarrow \infty$. He found that

$$r = \begin{cases} 0, & K < K_c \\ \sqrt{1 - (K_c/K)}, & K \geq K_c \end{cases}$$

where $K_c = 2\gamma$. In other words, the oscillators are desynchronized completely until the coupling strength K exceeds a critical value K_c . After that, the population splits into a partially synchronized state

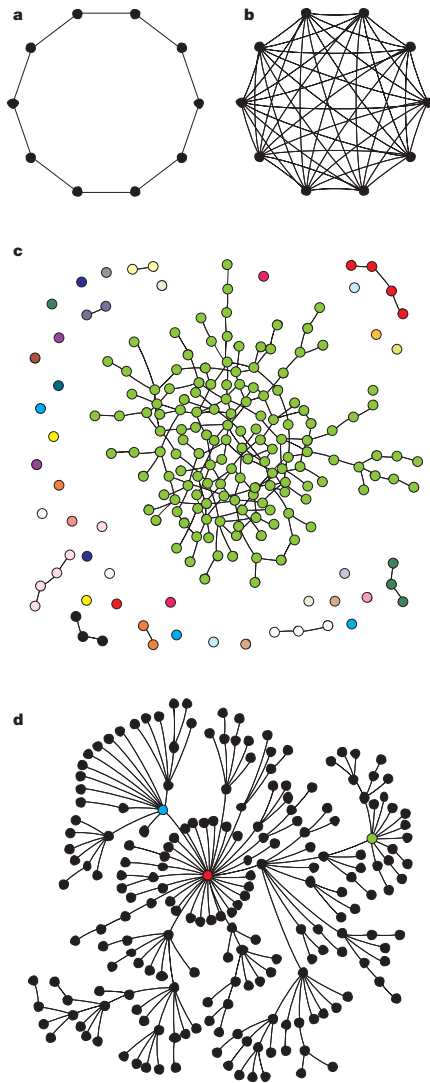


Figure 3 Schematic illustration of regular and random network architectures. **a**, Ring of ten nodes connected to their nearest neighbours. **b**, Fully connected network of ten nodes. **c**, Random graph constructed by placing n nodes on a plane, then joining pairs of them together at random until m links are used. Nodes may be chosen more than once, or not at all. The resulting wiring diagram (not shown) would be a snarl of criss-crossed lines; to clarify it, I have segregated the different connected components, coloured them, and eliminated as many spurious crossings as possible. The main topological features are the presence of a single giant component, as expected^{51–53} for a random graph with $m > n/2$ (here $n = 200$, $m = 193$), and the absence of any dominant hubs. The degree, or number of neighbours, is Poisson distributed across the nodes; most nodes have between one and four neighbours, and all have between zero and six. **d**, Scale-free graph, grown by attaching new nodes at random to previously existing nodes. The probability of attachment is proportional to the degree of the target node; thus richly connected nodes tend to get richer, leading to the formation of hubs and a skewed degree distribution with a heavy tail. Colours indicate the three nodes with the most links (red, $k = 33$ links; blue, $k = 12$; green, $k = 11$). Here $n = 200$ nodes, $m = 199$ links. Figure provided by D. Callaway. Network visualization was done using the Pajek program for large network analysis (<http://vlado.fmf.uni-lj.si/pub/networks/pajek/pajekman.htm>).

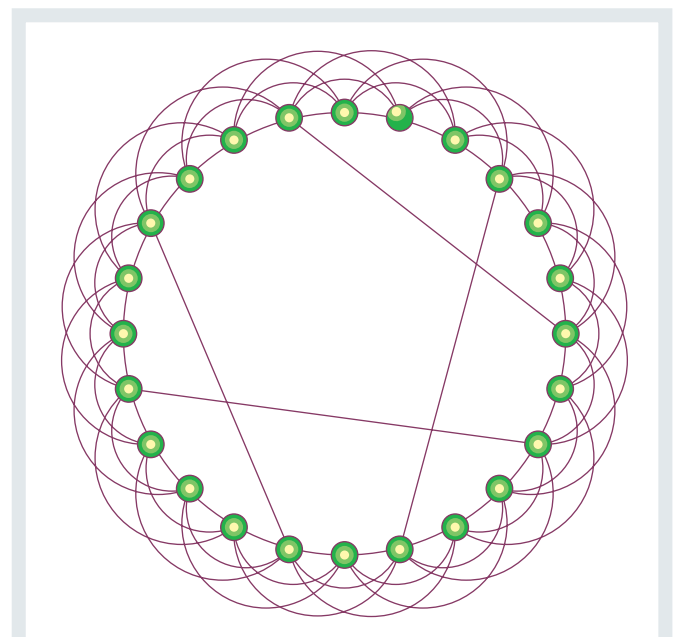


Figure 4 Solvable model of a small-world network. The model starts with a ring lattice of n nodes, each connected to its neighbours out to some range k (here $n = 24$ and $k = 3$). Shortcut links are added between random pairs of nodes, with probability ϕ per link on the underlying lattice. In the limit $n \gg 1$, the average path length between nodes can be approximated analytically. (Adapted from ref. 75.)

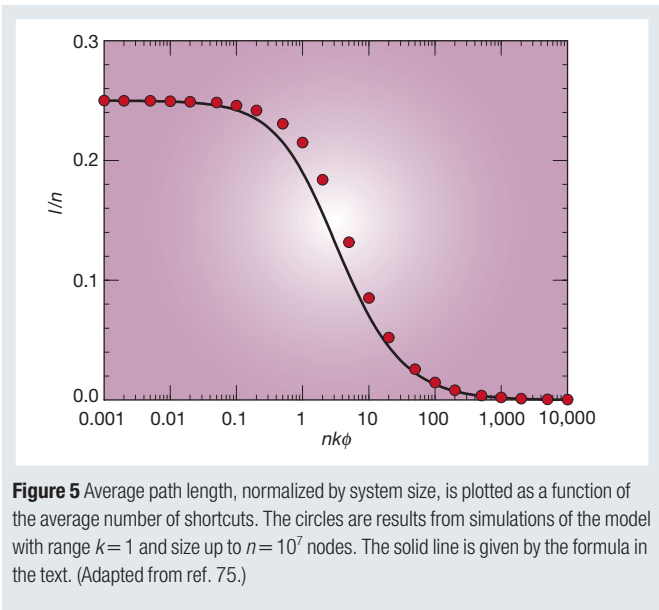


Figure 5 Average path length, normalized by system size, is plotted as a function of the average number of shortcuts. The circles are results from simulations of the model with range $k = 1$ and size up to $n = 10^7$ nodes. The solid line is given by the formula in the text. (Adapted from ref. 75.)

consisting of two groups of oscillators: a synchronized group that contributes to the order parameter r , and a desynchronized group whose natural frequencies lie in the tails of the distribution $g(\omega)$ and are too extreme to be entrained. With further increases in K , more and more oscillators are recruited into the synchronized group, and r grows accordingly.

Twenty-five years later, the Kuramoto model continues to surprise us (see ref. 45 for a review). First, the incoherent state with $r = 0$ was found to be neutrally stable below threshold, despite its apparent stability in simulations; the analysis reveals a connection to Landau damping in plasmas. Second, the square-root critical behaviour of r , almost a cliché for mean-field models in statistical mechanics, turns out to be non-generic; if the sinusoidal coupling is replaced by a periodic function with second harmonics, the scaling changes to $r \sim K - K_c$. Third, although the model was motivated originally by biological oscillators, it has appeared in such far-flung settings as the flavour evolution of neutrinos³², and arrays of Josephson junctions²⁷ and semiconductor lasers²⁴.

The main unsolved problem is the stability of the partially synchronized state for $K > K_c$. Numerical simulations indicate that it is globally stable, in the sense that it attracts almost all solutions, but even the linear stability problem has yet to be solved. Another issue concerns the extension of the model to nearest-neighbour coupling on a d -dimensional cubic lattice. Simulations⁴⁶ and renormalization arguments⁴⁷ indicate that the synchronization phase transition persists for $d \geq 3$ and vanishes for $d = 1$; the results are ambiguous for $d = 2$. All of this awaits a mathematical resolution.

In contrast to the mean-field models of Winfree and Kuramoto, Ermentrout and Kopell's classic work deals with one-dimensional chains of oscillators, first in connection with neuromuscular rhythms in the mammalian intestine⁴⁸, and later in their model of the central pattern generator for the lamprey eel^{49,50}. The main phenomena here involve travelling waves, rather than the synchrony found in mean-field models. This is not accidental, as wave propagation is essential for the generation of peristalsis in the intestine, and for the creation of the swimming rhythm in lamprey.

Ermentrout and Kopell introduced several deep mathematical innovations, but perhaps their most impressive result is a counterintuitive biological prediction. Their lamprey model suggested that the tail-to-head neural connections along the spinal cord would be stronger than those running from head to tail, despite the fact that the wave associated with swimming travels from head to tail. To

everyone's delight, that prediction was later confirmed by their experimental collaborators⁵⁰.

Complex network architectures

All the network topologies discussed so far — chains, grids, lattices and fully-connected graphs — have been completely regular (Fig. 3a, b). Those simple architectures allowed us to focus on the complexity caused by the nonlinear dynamics of the nodes, without being burdened by any additional complexity in the network structure itself. Now I take the complementary approach, setting dynamics aside and turning to more complex architectures. A natural place to start is at the opposite end of the spectrum from regular networks, with graphs that are completely random.

Random graphs

Imagine $n \gg 1$ buttons strewn across the floor⁵¹. Pick two buttons at random and tie them together with thread. Repeat this process m times, always choosing pairs of buttons at random. (If m is large, you might eventually select buttons that already have threads attached. That is certainly allowed; it merely creates clusters of connected buttons.) The result is a physical example of a random graph with n nodes and m links (Fig. 3c). Now slowly lift a random button off the floor. If it is tied to other buttons, either directly or indirectly, those are dragged up too. So what happens? Are you likely to pull up an isolated button, a small cluster or a vast meshwork?

Erdős and Rényi⁵² studied how the expected topology of this random graph changes as a function of m . When m is small, the graph is likely to be fragmented into many small clusters of nodes, called components. As m increases, the components grow, at first by linking to isolated nodes and later by coalescing with other components. A phase transition occurs at $m = n/2$, where many clusters crosslink spontaneously to form a single giant component. For $m > n/2$, this giant component contains on the order of n nodes (more precisely, its size scales linearly with n , as $n \rightarrow \infty$), while its closest rival contains only about $\log n$ nodes. Furthermore, all nodes in the giant component are connected to each other by short paths: the maximum number of 'degrees of separation' between any two nodes grows slowly, like $\log n$.

In the decades since this pioneering work, random graphs have been studied deeply within pure mathematics⁵³. They have also served as idealized coupling architectures for dynamical models of gene networks, ecosystems and the spread of infectious diseases and computer viruses^{29,51,54,55}.

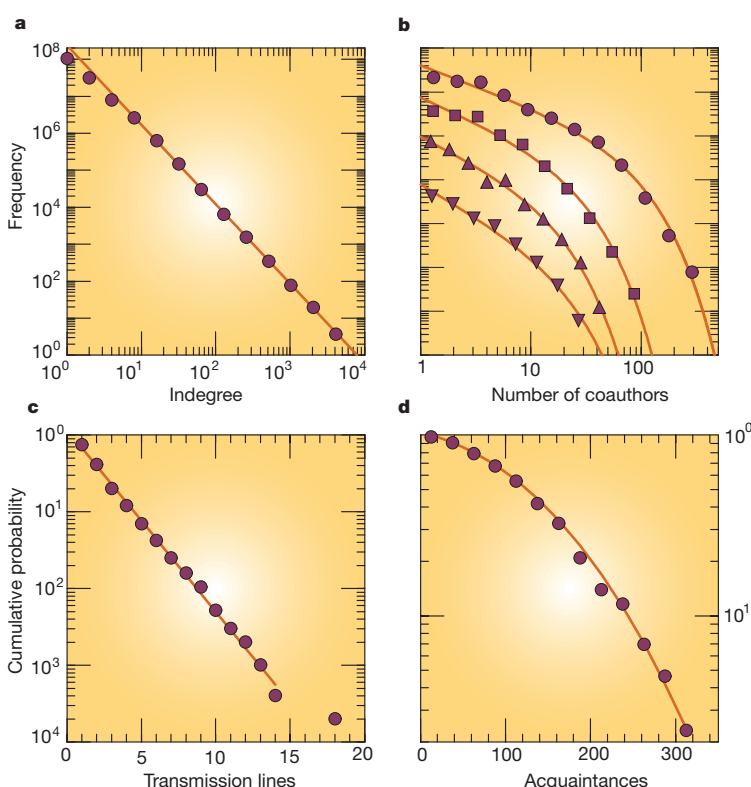
Small-world networks

Although regular networks and random graphs are both useful idealizations, many real networks lie somewhere between the extremes of order and randomness. Watts and Strogatz^{56,57} studied a simple model that can be tuned through this middle ground: a regular lattice

Table 1 Clustering for three affiliation networks		
Network	Clustering C	
	Theory	Actual
Company directors	0.590	0.588
Movie actors	0.084	0.199
Biomedical authors	0.042	0.088

US corporate directors: 7,673 company directors linked by joint membership on 914 boards of the Fortune 1,000 companies for 1999. Movie actors: 449,913 actors linked by mutual appearances in 151,261 feature films, as specified by the Internet Movie Database (www.imdb.com) as of 1 May 2000. Biomedical collaborations: 1,388,989 scientists linked by coauthorship of at least one of 2,156,769 biomedical journal articles published between 1995 and 1999 inclusive, as listed in the MEDLINE database. The clustering coefficient C is defined as the probability that a connected triple of nodes is actually a triangle; here nodes correspond to people, as in the unipartite representation shown in Fig. 7b. Intuitively, C measures the likelihood that two people who have a mutual collaborator are also collaborators of each other. The results show that the random model accurately predicts C for the corporate director network, given the network's bipartite structure and its degree distributions; no additional social forces need to be invoked. For the networks of actors and scientists, the model accounts for about half of the observed clustering. The remaining portion depends on social mechanisms at work in these communities (see text). (Adapted from ref. 91.)

Figure 6 Degree distributions for real networks. **a**, World-Wide Web. Nodes are web pages; links are directed URL hyperlinks from one page to another. The log–log plot shows the number of web pages that have a given in-degree (number of incoming links). The raw data have been logarithmically binned, with bin ratio 2. The tail of the distribution follows an approximate power law with exponent $\gamma \approx 2.2$. (Adapted from ref. 10.) **b**, Coauthorship networks. Nodes represent scientists; an undirected link between two people means that they have written a paper together. The data are for the years 1995–1999 inclusive and come from three databases: arxiv.org (preprints mainly in theoretical physics), SPIRES (preprints in high-energy physics) and NCSTRL (preprints in computer science). Symbols denote different scientific communities: astrophysics (circles), condensed-matter physics (squares), high-energy physics (upright triangles) and computer science (inverted triangles). The log–log plot shows that the probability of having a total of z coauthors is well approximated by a truncated power law (solid line) of the form $p_z \sim z^{-\gamma} \exp(-z/z_c)$, where z_c is the cutoff. The curves have been displaced vertically for clarity. (Adapted from ref. 14.) **c**, Power grid of the western United States and Canada. Nodes represent generators, transformers and substations; undirected links represent high-voltage transmission lines between them. The data are plotted as a cumulative distribution to reduce the effects of noise on this smaller data set. The linear–log plot shows that the proportion of nodes with at least k transmission lines decays exponentially in k . The negative derivative of this cumulative distribution is the degree distribution, also an exponential. (Adapted from ref. 62.) **d**, Social network. Nodes are 43 Mormons in Utah; undirected links represent acquaintances with other Mormons⁷⁹. The linear–log plot of the cumulative distribution is well fit by an error function (solid line), so the degree distribution is a gaussian. (Adapted from ref. 62.)



where the original links are replaced by random ones with some probability $0 \leq \phi \leq 1$. They found that the slightest bit of rewiring transforms the network into a ‘small world’, with short paths between any two nodes, just as in the giant component of a random graph. Yet the network is much more highly clustered than a random graph, in the sense that if A is linked to B and B is linked to C , there is a greatly increased probability that A will also be linked to C (a property that sociologists⁵⁸ call ‘transitivity’).

Watts and Strogatz conjectured that the same two properties — short paths and high clustering — would hold also for many natural and technological networks. Furthermore, they conjectured that dynamical systems coupled in this way would display enhanced signal propagation speed, synchronizability and computational power, as compared with regular lattices of the same size. The intuition is that the short paths could provide high-speed communication channels between distant parts of the system, thereby facilitating any dynamical process (like synchronization or computation) that requires global coordination and information flow.

Research has proceeded along several fronts. Many empirical examples of small-world networks have been documented, in fields ranging from cell biology to business^{9,14,59–64}. On the theoretical side, small-world networks are turning out to be a Rorschach test — different scientists see different problems here, depending on their disciplines.

Computer scientists see questions about algorithms and their complexity. Walsh⁶⁵ showed that graphs associated with many difficult search problems have a small-world topology. Kleinberg⁶⁶ introduced an elegant model of the algorithmic challenge posed by Milgram’s original sociological experiment⁶⁷ — how to actually find a short chain of acquaintances linking yourself to a random target person, using only local information — and he proved that the problem is easily solvable for some kinds of small worlds, and essentially intractable for others.

Epidemiologists have asked how local clustering and global

contacts together influence the spread of infectious disease, with implications for vaccination strategies and the evolution of virulence^{68–71}. Neurobiologists have wondered about the possible evolutionary significance of small-world neural architecture. They have argued that this kind of topology combines fast signal processing with coherent oscillations⁷², unlike either regular or random architectures, and that it may be selected by adaptation to rich sensory environments and motor demands⁶⁴.

Perhaps the strongest response to the Rorschach test comes from the statistical physicists, who sensed immediately⁷³ that the toy model of Watts and Strogatz⁵⁶ would yield to their techniques (see ref. 74 for a review). In its improved form the model starts with a ring of n nodes, each connected by undirected links to its nearest and next-nearest neighbours out to some range k . Shortcut links are then added — rather than rewired — between randomly selected pairs of nodes, with probability ϕ per link on the underlying lattice; thus there are typically $nk\phi$ shortcuts in the graph (Fig. 4). The question is: on average, how many steps are required to go from one node to another along the shortest route? If ℓ denotes that average separation, we find that ℓ drops sharply near $\phi=0$, confirming that a few shortcuts do indeed shrink the world dramatically. One of the most striking results is the following formula derived by Newman, Moore and Watts⁷⁵:

$$\ell = \frac{n}{k} f(nk\phi)$$

where

$$f(x) = \frac{1}{2\sqrt{x^2 + 2x}} \tanh^{-1} \frac{x}{\sqrt{x^2 + 2x}}$$

This solution is asymptotically exact in the limits $n \rightarrow \infty$ (large system size) and either $nk\phi \rightarrow \infty$ or $nk\phi \rightarrow 0$ (large or small number of

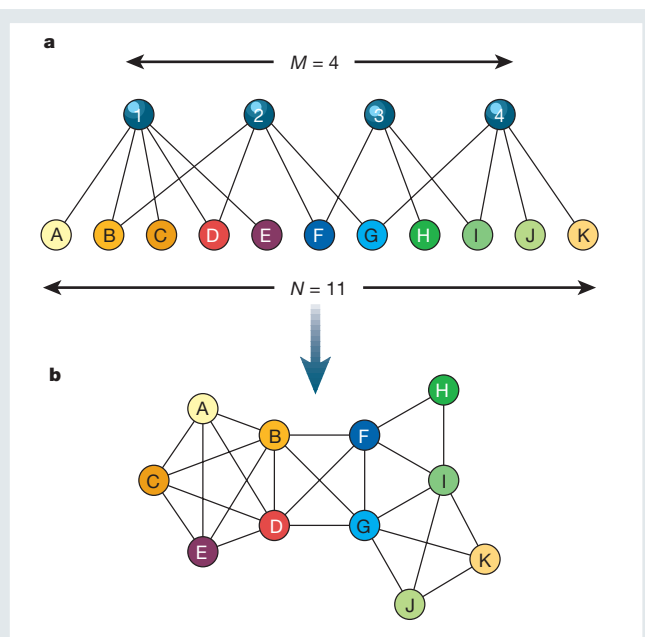


Figure 7 Bipartite versus unipartite representations of the corporate director network. **a**, In the bipartite approach, directors and boards are treated as distinct kinds of nodes. The schematic example shows 11 directors and 4 boards. Links indicate which people sit on which boards. By definition there are no links between pairs of people, or between pairs of boards. **b**, The more familiar unipartite representation depicts the people as nodes, with links between those on the same board, forming cliques. This approach discards important information and can conflate different structures. For example, the triangle FHI corresponds to board number 3, as seen in **a**, whereas the similar-looking triangle FGI does not correspond to any single board. Another confusing effect is the large number of cliques that occur automatically in this projection of the full bipartite graph. Such cliques account for much of the high clustering observed in real affiliation networks⁵⁹. The random graph model teases out this generic source of clustering from that indicative of more interesting social interactions. (Adapted from ref. 91.)

shortcuts). Figure 5 shows that it also gives the correct qualitative behaviour for $nk\phi \approx 1$. Barbour and Reinert⁷⁶ improved this result by proving a rigorous distributional approximation for ℓ , along with a bound on the error.

Scale-free networks

In any real network, some nodes are more highly connected than others are. To quantify this effect, let p_k denote the fraction of nodes that have k links. Here k is called the degree and p_k is the degree distribution.

The simplest random graph models^{52,53} predict a bell-shaped Poisson distribution for p_k . But for many real networks, p_k is highly skewed and decays much more slowly than a Poisson. For instance, the distribution decays as a power law $p_k \sim k^{-\gamma}$ for the Internet backbone¹¹, metabolic reaction networks⁹, the telephone call graph¹³ and the World-Wide Web¹⁰ (Fig. 6a). Remarkably, the exponent $\gamma \approx 2.1$ – 2.4 for all of these cases. Taken literally, this form of heavy-tailed distribution would imply an infinite variance. In reality, there are a few nodes with many links (Fig. 3d). For the World-Wide Web, think Yahoo; for metabolic networks, think ATP. Barabási, Albert and Jeong^{77,78} have dubbed these networks ‘scale-free’, by analogy with fractals, phase transitions and other situations where power laws arise and no single characteristic scale can be defined.

The scale-free property is common but not universal⁶². For coauthorship networks of scientists, p_k is fit better by a power law with an exponential cutoff⁴ (Fig. 6b); for the power grid of the western United States, p_k is an exponential distribution⁶² (Fig. 6c); and for a social network of Mormons in Utah⁷⁹, p_k is gaussian⁶² (Fig. 6d).

Nevertheless, the scale-free case has stimulated a great deal of theorizing. The earliest work is due to Simon^{80,81} in 1955, now independently rediscovered by Barabási, Albert and Jeong^{77,78}. They showed that a heavy-tailed degree distribution emerges automatically from a stochastic growth model in which new nodes are added continuously and attach themselves preferentially to existing nodes, with probability proportional to the degree of the target node. Richly connected nodes get richer, and the result is $p_k \sim k^{-3}$. More sophisticated models^{82–84} include the effects of adding or rewiring links, allowing nodes to age so that they can no longer accept new links, or varying the form of preferential attachment. These generalized models predict exponential and truncated power-law p_k in some parameter regimes, as well as scale-free distributions.

Could there be a functional advantage to scale-free architecture? Albert, Jeong and Barabási⁸⁵ suggested that scale-free networks are resistant to random failures because a few hubs dominate their topology (Fig. 3d). Any node that fails probably has small degree (like most nodes) and so is expendable. The flip side is that such networks are vulnerable to deliberate attacks on the hubs. These intuitive ideas have been confirmed numerically^{10,85} and analytically^{86,87} by examining how the average path length and size of the giant component depend on the number and degree of the nodes removed. Some possible implications for the resilience of the Internet^{79–81}, the design of therapeutic drugs⁹, and the evolution of metabolic networks^{9,59} have been discussed.

Generalized random graphs

As mentioned above, the simplest random graph predicts a Poisson degree distribution, and so cannot accommodate the other types of distribution found in real networks. Molloy and Reed^{88,89} introduced a more flexible class of random graphs in which any degree distribution is permitted. Given a sequence of non-negative integers $\{d_k\}$, where d_k denotes the number of nodes with degree k , consider the ensemble of all graphs with that prescribed degree sequence, and weight them all equally when computing statistical averages of interest. For this class of graphs, Molloy and Reed derived a simple condition for the birth of the giant component⁸⁸, and they also found an implicit formula for its size as a fraction of n , the total number of nodes⁸⁹. Specifically, let $n \gg 1$ and define

$$Q = \sum_{k=1}^{\infty} p_k k(k-2)$$

where $p_k = d_k/n$. If $Q < 0$, the graph consists of many small components. The average component size diverges as $Q \rightarrow 0$ from below, and a giant component exists for $Q > 0$. (In technical terms, these results hold ‘almost surely’; that is, with probability tending to 1 as $n \rightarrow \infty$.)

Aiello, Chung and Lu⁹⁰ applied these results to a random graph model for scale-free networks. For p_k of power-law form, the condition on Q implies that a giant component exists if and only if $\gamma < 3.47$, which holds for most scale-free networks measured so far. If $\gamma < 1$, there are so many high-degree hubs that the network forms one huge, connected piece. They also proved theorems about the number and size of small components outside the giant component, and compared these to a real graph of about 47 million telephone numbers and the calls between them in one day. They found that the data are best fit by an exponent $\gamma \approx 2.1$, which predicts correctly that the call graph is not connected but has a giant component.

The papers by Molloy and Reed^{88,89} and Aiello *et al.*⁹⁰ are mathematically rigorous. Newman, Strogatz and Watts⁹¹ recently developed a more heuristic approach based on generating functions. By handling the limit $n \rightarrow \infty$ in an intuitive way, their approach yields elementary derivations of the earlier results, along with new exact results for graphs with additional structure, such as directed or bipartite graphs.

The bipartite case is especially interesting for applications⁵⁸. By definition, in a bipartite graph there are two types of nodes, with links running only between different kinds (Fig. 7). For example, consider the network of the boards of directors of the Fortune 1,000 companies, the largest US corporations ranked according to revenues. This network is fascinating because the boards ‘interlock’ — some important people sit on several of them — and this overlap knits virtually all large US firms together into a giant web of corporate governance¹⁷.

Let p_j denote the probability that a director sits on exactly j boards, and let q_k denote the probability that a board consists of k directors. Figure 8 shows that p_j is approximately exponential, with most directors sitting on only one board, whereas q_k is strongly peaked around $k = 10$, indicating that most boards have about ten members. As a null hypothesis, assume that the Fortune 1,000 network is a random member of the ensemble of all bipartite graphs with the same p_j and q_k . Then generating functions yield predictions for various quantities of interest⁹¹. For example, suppose we want to calculate r_z , the probability that a random director works with a total of z other co-directors, summed over all the boards on which he or she serves. Let

$$f_0(x) = \sum_{j=0}^{\infty} p_j x^j$$

$$g_0(x) = \sum_{k=0}^{\infty} q_k x^k$$

be the generating functions associated with the empirical degree distributions p_j and q_k . If we now choose a random edge on the bipartite graph and follow it to the board at one of its ends, the distribution of the number of other edges leaving that board can be shown to be generated by $g_1(x) = g_0'(x)/\nu$, where $\nu = g_0'(1)$. Then for a randomly chosen director, the generating function for z is given by $G_0(x) = f_0(g_1(x))$. If we expand G_0 in a series as

$$G_0(x) = \sum_{z=0}^{\infty} r_z x^z,$$

the coefficients r_z are exactly the quantities we seek. They can be extracted by repeated differentiation: $r_z = (1/z!)(d^z G_0/dx^z)|_{x=0}$.

Figure 8c shows that the predicted r_z agrees almost perfectly with the actual distribution. Similarly, the clustering coefficient⁵⁶ predicted for the directors lies within 1% of the observed value (Table 1). Clearly the random model captures much of the structure of the real network.

However, for two other bipartite graphs — film actors and the movies they appeared in, and biomedical scientists and the papers they coauthored — the model⁹¹ underestimates the clustering coefficients by half (Table 1). The reason is that the random model quantifies only the generic portion of the clustering; it reflects the cliques that are formed automatically whenever a bipartite collaboration graph is projected onto the space of people, as in Fig. 7b. For the corporate board data, those cliques account for essentially all the clustering (simply because most directors sit on only one board, thus preventing clustering across boards). But for the scientists and actors, some further mechanisms must be at work. One possible explanation is that scientists tend to introduce pairs of their collaborators to each other, engendering new collaborations.

In this way the random model allows us to disentangle the generic features of bipartite graphs from those that could reflect sociological effects. Beyond their benchmarking role, generalized random graphs provide a promising new class of substrates on which dynamical processes can be simulated and even approached analytically. Using this approach, Watts⁹² has given an intriguing explanation of fads and

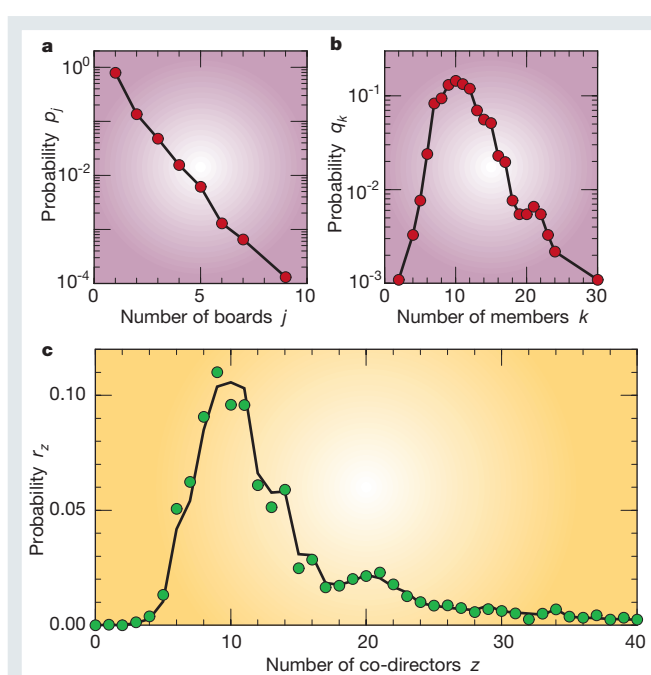


Figure 8 Structural properties of the Fortune 1,000 network of corporate directors for 1999. The data shown here comprise 7,673 directors and 914 boards. Most directors serve on the board of only one company, but about 20% sit on two or more boards, thereby creating interlocked directorates. This has potentially important consequences for business practices in the United States¹⁷. **a**, Distribution of the number of boards per director. The probability p_j that a director sits on exactly j boards decreases roughly exponentially with j . **b**, Distribution of the number of directors per board. The probability q_k that a board has k members is approximately lognormally distributed, with a typical board size around $k = 10$ members. **c**, Distribution of each director's total number of co-directors, summed over all the boards on which the director sits. The probability r_z of serving with z other co-directors is a complicated function of z , with a long tail and a secondary shoulder near $z = 20$, yet the theoretical prediction (solid line) agrees almost perfectly with the actual distribution (green circles). (Adapted from ref. 91.)

normal accidents as the natural consequence of cascade dynamics on sparse interaction networks.

Outlook

In the short run there are plenty of good problems about the nonlinear dynamics of systems coupled according to small-world, scale-free or generalized random connectivity. The speculations that these architectures are dynamically advantageous (for example, more synchronizable or error-tolerant) need to be sharpened, then confirmed or refuted mathematically for specific examples. Other ripe topics include the design of self-healing networks, and the relationships among optimization principles, network growth rules and network topology^{82–84,93–96}.

In the longer run, network thinking will become essential to all branches of science as we struggle to interpret the data pouring in from neurobiology, genomics, ecology, finance and the World-Wide Web. Will theory be able to keep up? Time to log back on to the Internet... □

1. Western Systems Coordinating Council (WSCC). Disturbance Report for the Power System Outage that Occurred on the Western Interconnection on August 10th, 1996 at 1548 PAST <<http://www.wscoc.com>> (October 1996).
2. Anonymous. Media: Six degrees from Hollywood. *Newsweek* 11 October 1999, 6 (1999).
3. Kirby, D. & Sahre, P. Six degrees of Monica. *New York Times* 21 February 1998, op. ed. page (1998).
4. Cohen, J. E., Briand, F. & Newman, C. M. *Community Food Webs: Data and Theory* (Springer, Berlin, 1990).
5. Williams, R. J. & Martinez, N. D. Simple rules yield complex food webs. *Nature* **404**, 180–183 (2000).
6. Kohn, K. W. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell* **10**, 2703–2734 (1999).

7. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
8. Bhalla, U. S. & Iyengar, R. Emergent properties of networks of biological signalling pathways. *Science* **283**, 381–387 (1999).
9. Jeong H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
10. Broder, A. *et al.* Graph structure in the web. *Comput. Netw.* **33**, 309–320 (2000).
11. Faloutsos, M., Faloutsos, P. & Faloutsos, C. On power-law relationships of the internet topology. *Comp. Comm. Rev.* **29**, 251–262 (1999).
12. Achacoso, T. B. & Yamamoto, W. S. *AY's Neuroanatomy of C. elegans for Computation* (CRC Press, Boca Raton, FL, 1992).
13. Abello, J., Buchsbaum, A. & Westbrook, J. A functional approach to external graph algorithms. *Lect. Notes Comput. Sci.* **1461**, 332–343 (1998).
14. Newman, M. E. J. The structure of scientific collaboration networks. *Proc. Natl Acad. Sci. USA* **98**, 404–409 (2001).
15. Seglen, P. O. The skewness of science. *J. Am. Soc. Inform. Sci.* **43**, 628–638 (1992).
16. Redner, S. How popular is your paper? An empirical study of the citation distribution. *Eur. J. Phys. B* **4**, 131–134 (1998).
17. Davis, G. F. The significance of board interlocks for corporate governance. *Corp. Govern.* **4**, 154–159 (1996).
18. Wilson, E. O. *Consilience* p.85 (Knopf, New York, 1998).
19. Weiss, C. O. & Vilaseca, R. *Dynamics of Lasers* (VCH, Weinheim, 1991).
20. Winful, H. G. & Wang, S. S. Stability of phase locking in coupled semiconductor laser arrays. *Appl. Phys. Lett.* **53**, 1894–1896 (1988).
21. Li, R. D. & Erneux, T. Preferential instability in arrays of coupled lasers. *Phys. Rev. A* **46**, 4252–4260 (1992).
22. Fabiny, L., Colet, P., Roy, R. & Lenstra, D. Coherence and phase dynamics of spatially coupled solid-state lasers. *Phys. Rev. A* **47**, 4287–4296 (1993).
23. Kourtchatov, S. Yu., Likhanskii, V. V., Naporotovich, A. P., Arecchi, F. T. & Lapucci, A. Theory of phase locking of globally coupled laser arrays. *Phys. Rev. A* **52**, 4089–4094 (1995).
24. Kozzyreff, G., Vladimirov, A. G. & Mandel, P. Global coupling with time delay in an array of semiconductor lasers. *Phys. Rev. Lett.* **85**, 3809–3812 (2000).
25. Winfree, A. T. *The Geometry of Biological Time* (Springer, New York, 1980).
26. Kuramoto, Y. *Chemical Oscillations, Waves, and Turbulence* (Springer, Berlin, 1984).
27. Wiesenfeld, K., Colet, P. & Strogatz, S. H. Frequency locking in Josephson arrays: connection with the Kuramoto model. *Phys. Rev. E* **57**, 1563–1569 (1998).
28. Turcotte, D. L. *Fractals and Chaos in Geology and Geophysics* 2nd edn (Cambridge Univ. Press, Cambridge, 1997).
29. May, R. M. *Stability and Complexity in Model Ecosystems* (Princeton Univ. Press, Princeton, 1973).
30. Levin, S. A., Grenfell, B. T., Hastings, A. & Perelson, A. S. Mathematical and computational challenges in population biology and ecosystem science. *Science* **275**, 334–343 (1997).
31. Arbib, M. (ed.) *The Handbook of Brain Theory and Neural Networks* (MIT Press, Cambridge, MA, 1995).
32. Pantaleone, J. Stability of incoherence in an isotropic gas of oscillating neutrinos. *Phys. Rev. D* **58**, 3002 (1998).
33. Stein, D. L. (ed.) *Lectures in the Sciences of Complexity* (Addison-Wesley, Reading, MA, 1989).
34. Pecora, L. M., Carroll, T. L., Johnson, G. A., Mar, D. J. & Heagy, J. F. Fundamentals of synchronization in chaotic systems: concepts and applications. *Chaos* **7**, 520–543 (1997).
35. VanWiggeren, G. D. & Roy, R. Communication with chaotic lasers. *Science* **279**, 1198–1200 (1998).
36. Collins, J. J. & Stewart, I. Coupled nonlinear oscillators and the symmetries of animal gaits. *J. Nonlin. Sci.* **3**, 349–392 (1993).
37. Pérez, C. J., Corral, A., Díaz-Guilera, A., Christensen, K. & Arenas, A. On self-organized criticality and synchronization in lattice models of coupled dynamical systems. *Int. J. Mod. Phys. B* **10**, 1111–1151 (1996).
38. Peskin, C. S. *Mathematical Aspects of Heart Physiology* 268–278 (Courant Institute of Mathematical Sciences, New York, 1975).
39. Mirollo, R. E. & Strogatz, S. H. Synchronization of pulse-coupled biological oscillators. *SIAM J. Appl. Math.* **50**, 1645–1662 (1990).
40. Abbott, L. F. & van Vreeswijk, C. Asynchronous states in neural networks of pulse-coupled oscillators. *Phys. Rev. E* **48**, 1483–1490 (1993).
41. Bressloff, P. C. Mean-field theory of globally coupled integrate-and-fire neural oscillators with dynamic synapses. *Phys. Rev. E* **60**, 2160–2170 (1999).
42. Golomb, D. & Hansel, D. The number of synaptic inputs and the synchrony of large, sparse neuronal networks. *Neural Comput.* **12**, 1095–1139 (2000).
43. Hopfield, J. J. Neurons, dynamics, and computation. *Phys. Today* **47**, 40–46 (1994).
44. Winfree, A. T. Biological rhythms and the behavior of populations of coupled oscillators. *J. Theor. Biol.* **16**, 15–42 (1967).
45. Strogatz, S. H. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D* **143**, 1–20 (2000).
46. Sakaguchi, H., Shinomoto, S. & Kuramoto, Y. Local and global self-entrainments in oscillator lattices. *Prog. Theor. Phys.* **77**, 1005–1010 (1987).
47. Daido, H. Lower critical dimension for populations of oscillators with randomly distributed frequencies: a renormalization-group analysis. *Phys. Rev. Lett.* **61**, 231–234 (1988).
48. Ermentrout, G. B. & Kopell, N. Frequency plateaus in a chain of weakly coupled oscillators. *SIAM J. Math. Anal.* **15**, 215–237 (1984).
49. Kopell, N. & Ermentrout, G. B. Symmetry and phaselocking in chains of weakly coupled oscillators. *Commun. Pure Appl. Math.* **39**, 623–660 (1986).
50. Sigvardt, K. A. & Williams, T. L. Models of central pattern generators as oscillators: the lamprey locomotor CPG. *Semin. Neurosci.* **4**, 37–46 (1992).
51. Kauffman, S. *At Home in the Universe* (Oxford, New York, 1995).
52. Erdős, P. & Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–61 (1960).
53. Bollobás, B. *Random Graphs* (Academic, London, 1985).
54. Kauffman, S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **22**, 437–467 (1969).
55. Kephart, J. O. & White, S. R. in *Proc. 1991 IEEE Comput. Soc. Symp. Res. Security Privacy* 343–359 (IEEE Computer Society Press, Los Alamitos, CA, 1991).
56. Watts, D. J. & Strogatz, S. H. Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442 (1998).
57. Watts, D. J. *Small Worlds* (Princeton Univ. Press, Princeton 1999).
58. Wasserman, S. & Faust, K. *Social Network Analysis: Methods and Applications* (Cambridge Univ. Press, New York, 1994).
59. Wagner, A. & Fell, D. The small world inside large metabolic networks. Preprint available at <http://www.santafe.edu/sfi/publications/Abstracts/00-07-041abs.html> (2000).
60. Adamic, L. The small world web. *Lect. Notes Comput. Sci.* **1696**, 443–452 (Springer, New York, 1999).
61. Kogut, B. & Walker, G. Small worlds and the durability of national networks: ownership and acquisitions in Germany. *Am. Sociol. Rev.* (in the press).
62. Amaral, L. A. N., Scala, A., Barthélémy, M. & Stanley, H. E. Classes of behavior of small-world networks. *Proc. Natl Acad. Sci. USA* **97**, 11149–11152 (2000).
63. Stephan, K. E. *et al.* Computational analysis of functional connectivity between areas of primate visual cortex. *Phil. Trans. R. Soc. Lond. B* **355**, 111–126 (2000).
64. Sporns, O., Tononi, G. & Edelman, G. M. Theoretical neuroanatomy: relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cereb. Cortex* **10**, 127–141 (2000).
65. Walsh, T. in *Proc. 16th Int. Joint Conf. Artif. Intell.* 1172–1177 <http://dream.dai.ed.ac.uk/group/tw/papers/wijcai99.ps>
66. Kleinberg, J. M. Navigation in a small world. *Nature* **406**, 845 (2000).
67. Milgram, S. The small world problem. *Psychol. Today* **2**, 60–67 (1967).
68. Wallinga, J., Edmunds, K. J. & Kretzschmar, M. Perspective: human contact patterns and the spread of airborne infectious diseases. *Trends Microbiol.* **7**, 372–377 (1999).
69. Ball, F., Mollison, J. & Scalia-Tomba, G. Epidemics with two levels of mixing. *Ann. Appl. Probab.* **7**, 46–89 (1997).
70. Keeling, M. J. The effects of local spatial structure on epidemiological invasions. *Proc. R. Soc. Lond. B* **266**, 859–867 (1999).
71. Boots, M. & Sasaki, A. ‘Small worlds’ and the evolution of virulence: infection occurs locally and at a distance. *Proc. R. Soc. Lond. B* **266**, 1933–1938 (1999).
72. Lago-Fernandez, L. F., Huerta, R., Corbacho, F. & Sigüenza, J. Fast response and temporal coherent oscillations in small-world networks. *Phys. Rev. Lett.* **84**, 2758–2761 (2000).
73. Barthélémy, M. & Amaral, L. A. N. Small-world networks: evidence for a crossover picture. *Phys. Rev. Lett.* **82**, 3180–3183 (1999).
74. Newman, M. E. J. Models of the small world: a review. *J. Stat. Phys.* **101**, 819–841 (2000).
75. Newman, M. E. J., Moore, C. & Watts, D. J. Mean-field solution of the small-world network model. *Phys. Rev. Lett.* **84**, 3201–3204 (2000).
76. Barbour, A. D. & Reinert, G. Small worlds. Preprint cond-mat/0006001 at <http://xxx.lanl.gov> (2000).
77. Barabási, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
78. Barabási, A.-L., Albert, R. & Jeong, H. Mean-field theory for scale-free random networks. *Physica A* **272**, 173–197 (1999).
79. Bernard, H. R., Killworth, P. D., Evans, M. J., McCarty, C. & Shelley, G. A. Studying social relations cross-culturally. *Ethnology* **27**, 155–179 (1988).
80. Simon, H. A. On a class of skew distribution functions. *Biometrika* **42**, 425–440 (1955).
81. Bornholdt, S. & Ebel, H. World-Wide Web scaling exponent from Simon’s 1955 model. Preprint cond-mat/0008465 at <http://xxx.lanl.gov> (2000).
82. Albert, R. & Barabási, A.-L. Topology of evolving networks: local events and universality. *Phys. Rev. Lett.* **85**, 5234–5237 (2000).
83. Dorogovtsev, S. N. & Mendes, J. F. F. Evolution of networks with aging of sites. *Phys. Rev. E* **62**, 1842–1845 (2000).
84. Krapivsky, P. L., Redner, S. & Leyvraz, F. Connectivity of growing random networks. *Phys. Rev. Lett.* **85**, 4629–4632 (2000).
85. Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
86. Cohen, R., Erez, K., ben-Avraham, D. & Havlin, S. Resilience of the Internet to random breakdowns. *Phys. Rev. Lett.* **85**, 4626–4628 (2000).
87. Callaway, D. S., Newman, M. E. J., Strogatz, S. H. & Watts, D. J. Network robustness and fragility: percolation on random graphs. *Phys. Rev. Lett.* **85**, 5468–5471 (2000).
88. Molloy, M. & Reed, B. A critical point for random graphs with given degree sequence. *Random Struct. Algorithms* **6**, 161–179 (1995).
89. Molloy, M. & Reed, B. The size of the giant component of a random graph with given degree sequence. *Combinatorics Probab. Comput.* **7**, 295–305 (1998).
90. Aiello, W., Chung, F. & Lu, L. A random graph model for power law graphs. *Exp. Math.* (in the press); preprint available at <http://math.ucsd.edu/~fan/power.pdf>.
91. Newman, M. E. J., Watts, D. J. & Strogatz, S. H. Random graphs with arbitrary degree distribution and their applications. Preprint cond-mat/0007235 at <http://xxx.lanl.gov> (2000).
92. Watts, D. J. A simple model of fads and cascading failures. Preprint available at <http://www.santafe.edu/sfi/publications/Abstracts/00-12-062abs.html> (2000).
93. Chorniak, C. Component placement optimization in the brain. *J. Neurosci.* **14**, 2418–2427 (1994).
94. Mitchison, G. Neuronal branching patterns and the economy of cortical wiring. *Proc. R. Soc. Lond. B* **245**, 151–158 (1991).
95. West, G. B., Brown, J. H. & Enquist, B. J. The fourth dimension of life: fractal geometry and the allometric scaling of organisms. *Science* **284**, 1677–1679 (1999).
96. Banavar, J. R., Colaiori, F., Flammini, A., Maritan, A. & Rinaldo, A. Topology of the fittest transportation network. *Phys. Rev. Lett.* **84**, 4745–4748 (2000).
97. Strogatz, S. H. *Nonlinear Dynamics and Chaos* (Perseus, New York, 1994).

Acknowledgements

Thanks to J. Ariaratnam, A.-L. Barabási, N. Martinez, M. E. J. Newman, D. Watts and A. Winfree for their comments on a draft of the manuscript, and to R. Albert, L. Amaral, M. Amin, W. Blake, A. Broder, D. Callaway, J. Collins, G. Davis, H. Ebel, K. Kohn, N. Martinez, R. Oliva, M. E. J. Newman, J. Thorp, D. Watts, J. Wiener, A. Winfree and H. Wang for providing data, figures and information. Research supported in part by the National Science Foundation, Department of Defense, and Electric Power Research Institute.

Synchronization and rhythmic processes in physiology

Leon Glass

Department of Physiology, Centre for Nonlinear Dynamics in Physiology and Medicine, McGill University, Montreal, Quebec, Canada H3G 1Y6

Complex bodily rhythms are ubiquitous in living organisms. These rhythms arise from stochastic, nonlinear biological mechanisms interacting with a fluctuating environment. Disease often leads to alterations from normal to pathological rhythm. Fundamental questions concerning the dynamics of these rhythmic processes abound. For example, what is the origin of physiological rhythms? How do the rhythms interact with each other and the external environment? Can we decode the fluctuations in physiological rhythms to better diagnose human disease? And can we develop better methods to control pathological rhythms? Mathematical and physical techniques combined with physiological and medical studies are addressing these questions and are transforming our understanding of the rhythms of life.

Physiological rhythms are central to life. We are all familiar with the beating of our hearts, the rhythmic motions of our limbs as we walk, our daily cycle of waking and sleeping, and the monthly menstrual cycle. Other rhythms, equally important but not as obvious, underlie the release of hormones regulating growth and metabolism, the digestion of food, and many other bodily processes. The rhythms interact with each other as well as the outside fluctuating, noisy environment under the control of innumerable feedback systems that provide an orderly function that enables life. Disruption of the rhythmic processes beyond normal bounds or emergence of abnormal rhythms is associated with disease. Figure 1 shows several examples of complex physiological rhythms.

The investigation of the origin and dynamics of these rhythmic processes — once the sole province of physicians and experimental physiologists — is coming under increasingly close examination by mathematicians and physicists. Mathematical analyses of physiological rhythms show that nonlinear equations (see Box 1) are necessary to describe physiological systems^{1–4}. In contrast to the linear equations of traditional mathematical physics (for example, Maxwell's equations, the heat equation, the wave equation or Schrödinger's equation), nonlinear equations rarely admit an analytical solution. Consequently, as Hodgkin and Huxley realized in their classic analysis of the properties of ionic channels in the membranes of squid nerve cells, numerical simulations are one essential feature of quantitative studies of physiological systems⁵. A complementary approach is to analyse qualitative aspects of simplified mathematical models of physiological systems. This involves a mathematical analysis of those features of physiological systems that will be preserved by classes of models that are sufficiently close to the real system. For example, periodic stimulation of a squid giant axon gives rise to a wide variety of regular and irregular rhythms that can be modelled by simple as well as complex mathematical models^{6–8}.

Although we know that deterministic equations can display chaotic dynamics, it is not straightforward to distinguish deterministic 'chaotic' dynamics from 'noisy' dynamics in real experimental data. The problems were underscored by Ruelle: "Real systems can in general be described as deterministic systems with some added

noise"⁹. Although in some carefully controlled situations it is possible to obtain good evidence that a system is obeying deterministic equations with a small amount of noise^{6–8,10}, more usually the origin and the amount of 'noise' is not easy to determine. In this review, I concentrate on three fundamental issues related to synchronization and rhythmic processes in physiology: origins of complex physiological rhythms; synchronization of physiological oscillations; and the function of noise and chaos in physiological processes with particular emphasis on stochastic resonance. Finally, I discuss the potential applications of these ideas to medicine.

Origins of complex physiological rhythms

Physiological rhythms are rarely strictly periodic but rather fluctuate irregularly over time (Fig. 1). The fluctuations arise from the combined influences of the fluctuating environment, the 'noise' that is inherent in biological systems, and deterministic, possibly chaotic, mechanisms. In most natural as opposed to laboratory settings, there is continual interaction between the environment and the internal control mechanisms, so that separation of dynamics due to intrinsic rather than extrinsic mechanism is not possible. Independent of the mechanism for the fluctuation, it is usually not clear whether the fluctuations are essential to the physiological function, or whether the physiological functions are carried out despite the fluctuation.

At a subcellular level, ionic channels in cell membranes open and close in response to voltage and chemical changes in a cell's environment. An ionic channel lets a specific ion pass through it provided there is a concentration gradient of that ion between the intracellular and extracellular medium and the channel is open. Because histograms of open and closed times of ionic channels are often well fit by an exponential or a sum of exponentials, theoretical models of channel activity often assume that the dynamics of channel opening and closing are governed by simple random processes such as the Poisson process^{11–13}. Figure 2a shows a schematic representation of five channels, each of which is open at time 0 and each of which closes randomly with a fixed probability of 0.1 ms^{-1} . Figure 2b shows the fraction of open channels as a function of time for a membrane with 5, 50 and 500 channels. As the numbers of channels in a membrane increases, the falloff of the fraction of open

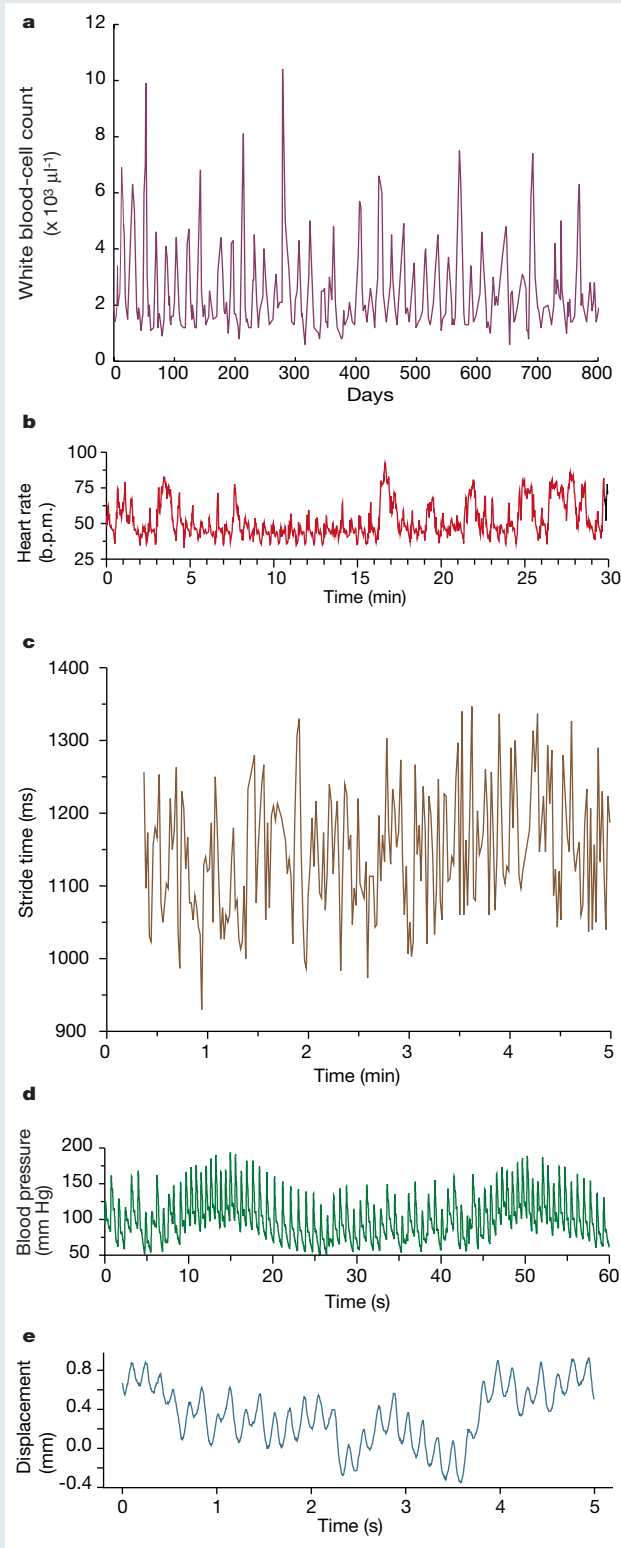


Figure 1 Representative physiological time series. **a**, White blood-cell count in a patient with cyclical neutropenia (tracing provided by D.C. Dale, C. Haurie and M. C. Mackey)³¹. **b**, Heart rate in a subject at high altitude (adapted with permission from the British Heart Journal)⁸⁵. **c**, Stride time in a patient with Huntington's disease (tracing provided by J. Hausdorff)^{86,87}. **d**, Blood pressure in a patient with sleep apnoea (tracing provided by A. Goldberger). **e**, Parkinsonian tremor measured from a finger (tracing provided by R. Edwards and A. Beuter⁷⁹). (The traces in panels **c** and **d** are adapted with permission from the Research Resource for Complex Physiologic Signals at <http://www.physionet.org>.)

Box 1

Properties of nonlinear equations

Nonlinear dynamics is a well developed discipline with many outstanding results⁸⁹. Three concepts are central — bifurcations, limit-cycle oscillations and chaos.

Bifurcations are changes in qualitative properties of dynamics. For example, as a parameter changes, steady states can become unstable and lead to stable oscillations, or a system with one stable steady state can be replaced by systems with multiple stable steady states. Physiological correlates are immediate. Drugs may lead to changes in control systems so that an abnormal, unhealthy rhythm is replaced by a more normal one. Mathematically, the drug induces a bifurcation in the dynamics, and as such, the actions of the drug can be analysed in a theoretical context. Often, the same type of bifurcation can be found in a host of different mathematical equations or experimental systems, and it is common to consider the 'universal' features of such bifurcations. Because many diseases are classified and identified by physicians based on characteristic qualitative features of dynamics, there is a natural match between the emphasis on qualitative dynamics in both mathematics and medicine.

Stable limit-cycle oscillations are a key feature of some nonlinear equations. Following a perturbation, a stable limit-cycle oscillation re-establishes itself with the same amplitude and frequency as before the perturbation. A perturbation to a linear oscillation may lead to a new amplitude of oscillation. For example, there is an intrinsic pacemaker that sets the rhythm in human hearts. If one or more electric shocks are delivered directly to the heart near the intrinsic pacemaker, the heart rhythm is modified transiently but re-establishes itself with the same frequency as before within a few seconds.

Chaos refers to aperiodic dynamics in deterministic equations in which there is a sensitivity to initial conditions. This means that even in the absence of stochastic processes, irregular rhythms can be generated. Although it is easy to consider mathematical systems in which all stochastic influences have been eliminated, in real physical and biological systems it is impossible to eliminate stochastic inputs. Thus, although chaotic dynamics is a clear mathematical concept, application of this concept to real biological systems is a difficult undertaking.

channels approaches the exponential distribution $e^{-0.1t}$. Although most models of channel opening and closing assume a stochastic mechanism such as the one above, deterministic chaotic models might also be consistent with the observed channel dynamics^{2,3}.

It is remarkable that the irregular openings and closings of ionic channels underlie all neural and muscular activities, including those that require great accuracy and precision such as reading this sentence, playing a violin concerto or carrying out a gymnastics routine. Because typical cells have of the order of at least 10^3 ionic channels of each type, deterministic equations can be used to model cells even though the underlying mechanism is probably stochastic^{11–13}. This notion is supported by studies of heart pacemaker cells in which channel dynamics have been modelled by random Markov processes. The resulting dynamics are similar to those generated by the deterministic models with an added 'noisy' fluctuation of period similar to what is observed experimentally^{14,15}. Additional regularization of dynamics can arise as a consequence of coupling of cells with irregular dynamics or cells that are heterogeneous^{15–17}.

Given the difficulty of analysing the origin of temporal fluctuations on a subcellular or cellular level, it is not surprising that the analysis of physiological rhythms in intact organisms provides added difficulties. In some cases, for example, for electrical activity of the

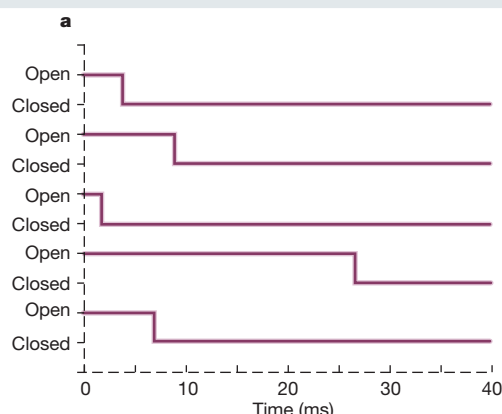
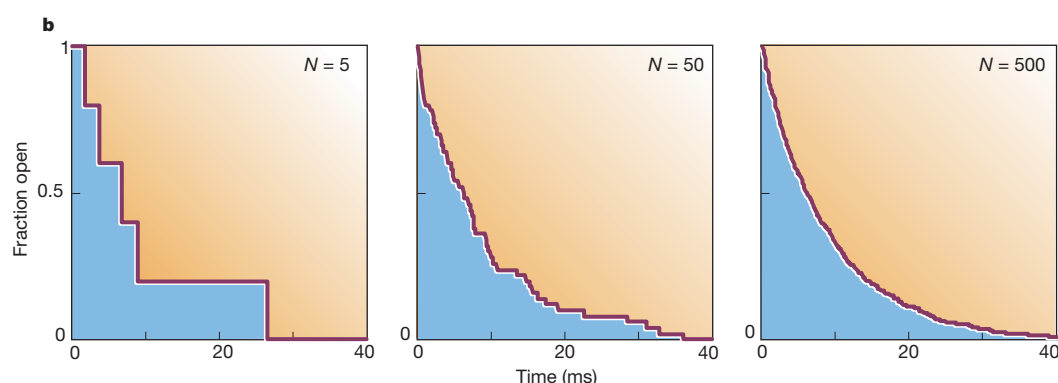


Figure 2 Schematic diagram showing dynamics in ionic channels that deactivate by a Poisson process (based on an analysis of dynamics observed in acetylcholine channels¹³). **a**, Channels that are open at $t = 0$, close randomly with a probability of 0.1 ms^{-1} . **b**, The fraction of open channels as a function of time for $N = 5$, $N = 50$ and $N = 500$ channels.



brain and the heart, the data can be collected easily using electrodes on the body surface, and computers provide a means for rapid analysis of large data sets. In other areas, such as endocrine function, determination of hormone fluctuations is difficult and expensive, requiring drawing blood at frequent intervals and performing expensive assays on the blood samples.

Consider the timing of the normal heartbeat as a paradigmatic example of physiological rhythm. Maintaining an adequate blood flow to the brain is essential to life. The heartbeat is generated by an autonomous pacemaker in the heart, but its frequency is mediated by neural activity controlled in turn by a large number of different feedback circuits all acting in parallel. As a consequence, the normal heart rate displays complex fluctuations in time in response to environmental factors such as breathing, exercise, changes in posture and emotion¹⁸. Diseases that impair heart function, for example damage to the heart caused by a heart attack or high blood pressure, lead to impaired pumping ability of the heart. In such cases, because the normal heart rate would not pump adequate blood to the brain, the heart generally beats more rapidly, and many of the normal fluctuations of heart rate are blunted. Lacking clearly defined and widely accepted operational definitions for chaotic dynamics, it is not surprising that agreement is lacking about whether or not normal heart dynamics are chaotic¹⁹ or not chaotic^{20,21}. Various tests of time variation have been applied to heart-rate variability to show that heart-rate fluctuation displays $1/f$ noise²², fractal dynamics with long-range correlations^{23,24}, and multifractal dynamics²⁵. Although similar dynamics in physical systems have been associated with self-organized criticality²⁶, in the biological context it is impossible to assert a mechanism based on the current observations. Indeed, it is possible that many of the properties observed reflect the response of individuals to a changing environment, and the environmental inputs themselves have interesting scaling properties²⁷.

There have been extensive quantitative analyses of data from many other physiological systems. In some instances the original motivation for the analyses was to determine whether the system dynamics was chaotic, although I believe the answer to this remains unclear. But

the data do reveal extraordinarily rich dynamics, and I mention several examples briefly. Electroencephalographic data reflect integrated activity from large numbers of cells in localized regions of the brain. A seizure shows characteristic changes in electroencephalographic records typified by larger-amplitude, regular and sustained oscillations reflecting broad synchronization of neural activity²⁸. Standard clinical interpretation of electroencephalographic data has been supplemented by a variety of quantitative analyses motivated by nonlinear dynamics^{29,30}. Analysis has also been carried out of a variety of normal and abnormal rhythms of blood-cell levels. Some blood disorders are characterized by pronounced fluctuations in levels of circulating white blood cells³¹. Blood-cell levels are controlled by feedback loops with long time delays and theoretical models have succeeded recently in analysing the effects of various manipulations. Endocrine function also provides a challenging area in view of the difficulty of obtaining temporal data, although sustained efforts have generated time series of various hormones including parathyroid hormone³², growth hormone³³ and prolactin³⁴.

These initial studies indicate extremely rich dynamics with differences between normal individuals and patients. The issue of whether or not the dynamics reflect chaos is much less interesting than elucidating the underlying mechanisms controlling the dynamics.

Synchronization of physiological oscillators

Although many cells in the body display intrinsic, spontaneous rhythmicity, physiological function derives from the interactions of these cells with each other and with external inputs to generate the rhythms essential for life. Thus, the heartbeat is generated by the sinoatrial node, a small region in the right atrium of the heart composed of thousands of pacemaker cells that interact with each other to set the normal cardiac rhythm^{14,15}. Nerve cells generating locomotion synchronize with definite phase relations depending on the species and the gait³⁵. And the intrinsic sleep-wake rhythm is usually synchronized to the light-dark cycle^{1,36}. In general, physiological oscillations can be synchronized to appropriate external or internal stimuli, so it is important to analyse the effects of

stimuli on intrinsic physiological rhythms. Even the simplest theoretical models (see Box 2 and Figs 3, 4) show the enormous complexity that can arise from periodic stimulation of nonlinear oscillations.

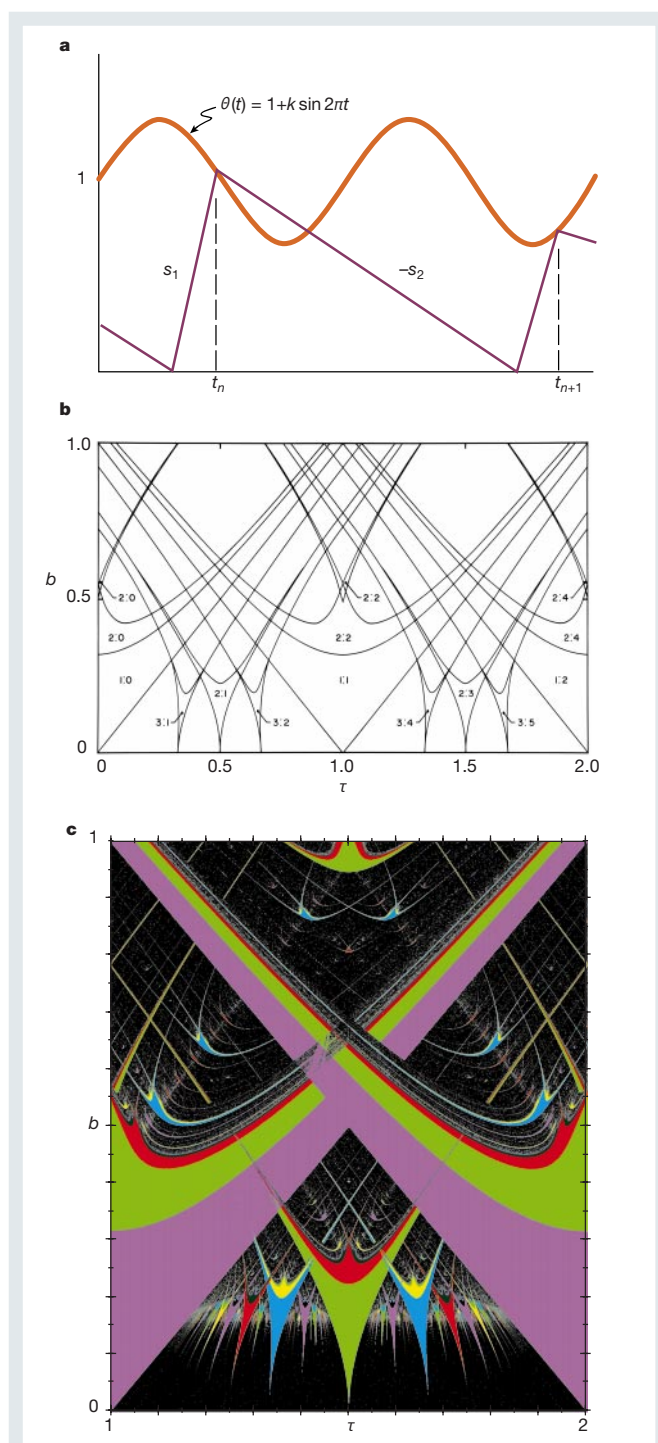


Figure 3 Integrate and fire model and locking zones. **a**, An activity increases linearly until it reaches a sinusoidal threshold and then decreases linearly. The successive times when the activity reaches threshold are determined iteratively. **b**, Schematic picture of the main Arnold tongues for the sine circle map (equation (3) in Box 2). $N:M$ locking reflects M cycles of the sine wave and N cycles of the activity. Chaotic dynamics are found only for $b > 1/2\pi$. **c**, Colour representation of a region of **b**. The colours code different periodic orbits (compare with **b**). The delicate geometry and self-similar structures are more evident in this representation. (Panels **a** and **b** modified with permission from ref. 88; colour version of the locking zones provided by J. Gallas.)

In vitro experiments in cardiac and neural tissue have clarified the effects of periodic stimulation in biological systems. Heart and nerve tissue are examples of excitable tissue⁴. This means that in response to a stimulus that is sufficiently large they will generate an event called an action potential. Following the action potential, for a time interval called the refractory period, a second stimulus does not elicit a second action potential. During periodic stimulation there are both periodic synchronized rhythms and aperiodic rhythms (see Box 2). Periodic stimulation of biological systems can give also give rise to aperiodic rhythms. For weak stimulation, it is common to find quasiperiodic rhythms, in which two rhythms with different frequencies march through each other with little interaction. Other aperiodic rhythms are chaotic. Identification of chaotic dynamics in periodically stimulated heart and nerve tissue is made more certain by the derivation of deterministic, theoretical models that are accurate both quantitatively and qualitatively and that correspond closely to both the regular and irregular dynamics observed experimentally^{6–8,10}. As the frequencies and amplitudes of the stimulation are varied, there is a characteristic organization of the phase-locking zones. However, if the dynamics are dissected on a fine scale, the detailed bifurcations in model systems differ from one another over some amplitude ranges. In experimental systems it is difficult to carry out locking studies in which stimulation parameters are varied finely because living preparations can be affected by stimulation and may change over the lengthy times needed to carry out the stimulation.

External inputs can often synchronize biological rhythms. Plants and animals display a circadian rhythm in which key processes show a 24-hour periodicity. This periodicity is usually set by the 24-hour light–dark cycle, but if this is removed by placing the organism in a constant environment, a cycle length different from 24 hours is observed. Thus the light–dark cycle entrains the intrinsic rhythm. If there is shift of the light–dark cycle, for example as might be generated by visiting a different time zone, then a time lag occurs until there is a new synchronization. Such phenomena have been modelled by both integrate and fire models and limit-cycle models^{36–39}.

Other circumstances in which physiological rhythms are stimulated by regular, periodic inputs occur in the context of medical devices. A mechanical ventilator assists breathing in experiments and in people who have respiratory failure. Such devices can be used in a variety of modes, but in the simplest, the physician sets the period and amplitude and the mix of gases for the ventilator, which then periodically inflates the patient's lungs. The resulting periodic lung inflation can interact with the person's intrinsic respiratory rhythm. In some instances, the respiratory rhythm will be entrained to the ventilator, but in other cases the patient will breath out when the ventilator is in the inflation phase^{40–42}.

These examples illustrate the effects of an external periodic input on intrinsic physiological rhythms. But the physiological rhythms also interact with one another. An example is the increase of the heart frequency during inspiration. Although the interactions between the respiratory and cardiac rhythms are not strong enough usually to lead to synchronization, such synchronization has been demonstrated in healthy high-performance swimmers⁴³.

It seems likely that many bodily activities require synchronization of cellular activity. For example, synchronization seems to be an essential component of many cortical tasks — coherent oscillations at 25–35 Hz are found in the sensorimotor cortex of monkeys⁴⁴ and 40–60-Hz oscillations are found in the visual cortex of cats⁴⁵.

Many different sorts of mathematical models have been proposed to account for synchronization in diverse systems. For example, synchronization has been observed in mathematical models of populations of cells that generate the heart beat in leeches¹⁶, the respiratory rhythm in rats⁴⁶, gamma and beta rhythms in the brain⁴⁷, and the circadian rhythm⁴⁸. However, because coupled oscillators show robust behaviour in which units tend to become synchronized⁴⁹, the observation of synchronization in models is not in itself a strong indicator of the suitability of the models.

Given the enormous numbers of connections between different cells and organs in the body, perhaps the biggest puzzle is not why bodily rhythms may in some circumstances become synchronized to

each other, but rather why there seems to be so much apparent independence between rhythms in different organs.

The function of noise and chaos in physiological systems

When looked at carefully, all physiological rhythms display fluctuations. Do the physiological systems function well despite these fluctuations, or are the fluctuations themselves intrinsic to the operation of the physiological systems? And if the fluctuations are essential to the functioning of the systems, is there some reason to believe that chaotic fluctuations would provide added function over stochastic fluctuations? There are no clear answers to these questions.

Chaotic dynamics might underlie normal function. In chaotic dynamics, there are typically many unstable periodic orbits, and there is also a sensitive dependence on initial conditions so that small, well-timed perturbations could lead to strong effects on the dynamics. Scientists have learned how to control chaotic dynamics⁵⁰ and nonlinear dynamics⁵¹ in model equations and in the laboratory. Electrical stimulation of complex dynamics in cardiac⁵² and neural⁵³ systems using chaos-control techniques has led to the regularization of complex rhythms, although the mechanisms of these effects are not certain⁵¹. Rabinovich and Abarbanel suggest that chaotic dynamics might be easier for the body to control than stochastic dynamics¹⁶. Another way the body might exploit chaos is by associating different unstable orbits with different percepts. Skarda and Freeman⁵⁴ proposed that the spatiotemporal fluctuations observed in the olfactory bulb in rabbits are chaotic. Each different odour is associated with and selects a different unstable spatiotemporal pattern of oscillation. However, in view of the difficulties associated with recording and interpreting spatiotemporal dynamics in living biological systems, and the huge gaps in understanding complex physiological functions involved in cognition and control, these claims remain intriguing hypotheses rather than established fact.

In normal circumstances, detection of signals is hampered by noise. For example, because the aesthetic and practical utility of sounds and images is reduced as noise level increases, designers of devices for recording and playback of sound and images make strenuous efforts to maximize the signal-to-noise ratio.

In other circumstances, however, the presence of noise and/or chaotic dynamics can improve detection of signals. Stochastic resonance refers to a situation in which the signal-to-noise ratio is maximum at an intermediate level of noise^{55–58}. For example, in tasks that are at the threshold of perception, the addition of noise can improve threshold detection. To illustrate this, consider a 'leaky' integrate and fire model⁵⁹ in the presence of noise as a model for stochastic resonance (Fig. 5). Assume that an activity (here a membrane potential of a nerve cell) x is governed by the differential equation

$$\frac{dx}{dt} = -cx + I + \xi \quad (1)$$

where c is a decay constant, I is constant input and ξ is added random noise. The activity rises to a threshold $f(t) = 1 + k \sin 2\pi t$, which represents an oscillating signal. If $c = 0$, then x increases linearly to the threshold (as shown in Fig. 3 and Box 2). When the activity reaches the threshold it elicits an action potential indicated by the arrows, followed by an immediate reset of the activity to zero. With $\xi = 0$ the maximum value of x is I/c . Consequently, if $(1 - k) > I/c$ and there is very low noise, then the activity would never reach threshold (Fig. 5a). But if there is moderate noise the activity can reach threshold, and this tends to occur at the troughs of the sine wave (Fig. 5b). Of course, if the noise is too great, then the activity is no longer as concentrated at the troughs (Fig. 5c). Thus, at intermediate values of the noise, the activity is well synchronized to the sinusoidal input, but rather than occurring on every cycle, there is a stochastic element resulting in a skipping pattern of activity in which excitation tends to occur at random multiples of a common divisor. Although the mechanisms are not well understood, similar rhythms, which may represent a 'stochastic phase locking', are observed frequently in physiological data^{56,57}.

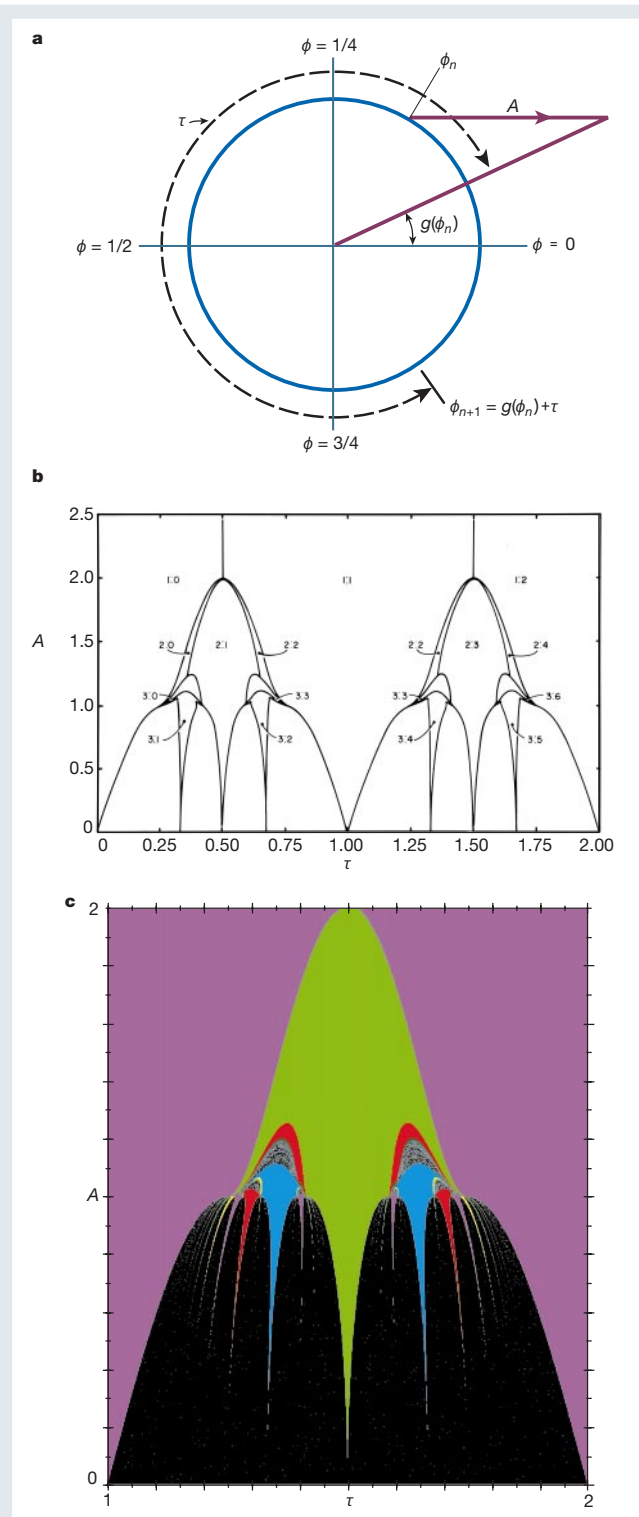


Figure 4 Poincaré oscillator model and locking zones. **a**, A stimulus of amplitude A delivered to the limit-cycle oscillator in equation (4) (Box 2) resets the phase. **b**, Schematic picture of the main Arnold tongues for the periodically forced Poincaré oscillator (equations (4) and (5) in Box 2). N : M locking reflects N stimuli for M cycles of the limit cycle oscillation. The Arnold tongue structure is present for $0 \leq A \leq 1$. **c**, Colour representation of a region of **b**. (Panels **a** and **b** modified with permission from ref. 88; colour version of the locking zones provided by J. Gallas.)

Box 2

Simple models for synchronization of nonlinear oscillators

Because biological oscillators are stable nonlinear oscillations, many of the general features of the interactions between the periodic input and the ongoing rhythm can be predicted without knowing precise details of the nature of the oscillator or the input. Simple toy models of synchronization show a host of phenomena shared by more complex models and real systems^{2,4,49,59,88,90–92}.

The integrate and fire model assumes that there is an activity that rises to a periodically oscillating threshold^{59,90} and then resets to a lower threshold. Depending on the setting, the activity might represent a nerve membrane potential, a substance that induces sleep or a neural activity associated with the timing of inspiration. There are also many ways in which one might model the activity. It could be modelled by a function that increases linearly in time, or by a function that saturates, or perhaps even by a biased random walk. One of the simplest embodiments of the integrate and fire model is to assume an oscillatory threshold $\theta(t) = 1 + k \sin 2\pi t$ and an activity that increases linearly to the threshold, followed by a linear decrease to zero (Fig. 3a). To illustrate some of the properties of this model, consider the situation in which there is a jump from 0 to the threshold, followed by a linear decay (slope of $-s_2$) to zero. Let t_n be the time when the oscillation is at the threshold value for the n th time. Then

$$t_{n+1} = t_n + 1/s_2 + (k/s_2) \sin 2\pi t_n \quad (2)$$

If $\tau = 1/s_2$, $b = k/s_2$ and $\phi_n = t_n \pmod{1}$, we obtain

$$\phi_{n+1} = \phi_n + \tau + b \sin 2\pi \phi_n \pmod{1} \quad (3)$$

This difference equation — sometimes called the sine circle map — displays an incredible array of complex rhythms⁹¹. One type of rhythm is synchronization of the activity to the sine function in an $N:M$ phase-locked rhythm so that for each M cycles of the sine wave the activity reaches the threshold N times. For low amplitudes ($0 \leq b \leq 1/2\pi$) of the sine-wave modulation, there is an orderly arrangement of the phase-locking zones, called Arnold tongues. For fixed values of b , as τ increases all rational ratios M/N are observed and there are quasiperiodic rhythms in which the two rhythms march through each other with slight interaction. For $b > 1/2\pi$, the simple geometry of Arnold tongues breaks down and there are chaotic dynamics as well as bistability in which two different stable rhythms exist simultaneously. Figure 3b,c gives a hint at the complexity of the organization of the locking zones.

A second prototypical oscillator was described originally by the French mathematician Poincaré

$$\frac{dr}{dt} = cr(1-r), \quad \frac{d\phi}{dt} = 2\pi \quad (4)$$

where the equation is written in a radial coordinate system (r, ϕ) . In this equation, there is a stable limit cycle at $r = 1$. More realistic models of biological oscillations also have limit cycles, but the variables in these models would reflect the mechanism of the underlying biological oscillation. For example, more realistic models of biological oscillators might be written in terms of ionic currents or neural activities. Because there are certain features of the qualitative response of biological oscillators to single and periodic stimulation that depend solely on the properties of nonlinear oscillations¹, the tractable analytic properties of the Poincaré oscillator make it an ideal system for theoretical analysis. We assume that the stimulus is a translation by a horizontal distance A from the current state point, followed by evolution according to equation (4) (Fig. 4a). If c is sufficiently large, then after the periodic perturbation there is a very rapid return to the limit cycle, and the dynamics can be described approximately by⁹²

$$\phi_{n+1} = \tau + g(\phi_n, A) \pmod{1}, \quad \text{where } g(\phi, A) = \frac{1}{2\pi} \arccos \frac{\cos 2\pi \phi + A}{(1 + A^2 + 2A \cos 2\pi \phi)^{1/2}} \quad (5)$$

Although for $0 \leq A \leq 1$ there is again a simple Arnold-tongue geometry that is the same topologically as that found in the integrate and fire model for low amplitudes, for higher amplitudes the detailed geometry is radically different (Fig. 4b, c).

Although noise is helpful in this artificial example, various proposals have been made suggesting that signal detection might be facilitated in physiological systems by similar mechanisms in which noise was either added externally or present intrinsically. Several reports indicate that added noise seems to enhance signal detection in *in vitro* preparations⁵⁸, in animal experiments^{60,61} and in human perception^{62–64}. Because noise is an intrinsic property of human physiological systems, a compelling question is whether or not noise may be acting to aid function in normal activities in a manner similar to the way it can act to enhance the detection of subthreshold activities.

Prospects for medical applications

In the physical sciences, advances in basic science have inevitably led to new technology. Although many technological innovations in medicine, such as X-ray imaging and nuclear magnetic resonance imaging, have derived from advances in the physical sciences, the recent mathematical analyses of temporal properties of physiological

rhythms have not yet led to medical advances, although several directions are under active consideration.

There is a wide spectrum of dynamical behaviour associated with both normal and pathological physiological functioning. Extremely regular dynamics are often associated with disease, including periodic (Cheyne–Stokes) breathing, certain abnormally rapid heart rhythms, cyclical blood diseases, epilepsy, neurological tics and tremors. However, regular periodicity can also reflect healthy dynamics — for example in the sleep–wake cycle and menstrual rhythms. Finally, irregular rhythms can also reflect disease. Cardiac arrhythmias such as atrial fibrillation and frequent ectopy, and neurological disorders such as post-anoxic myoclonus, are often highly irregular. The term ‘dynamical disease’ captures the notion that abnormal rhythms, which could be either more irregular or more regular than normal, arise owing to modifications in physiological control systems that lead to bifurcations in the dynamics⁶⁵. What is most important in distinguishing health from disease is that

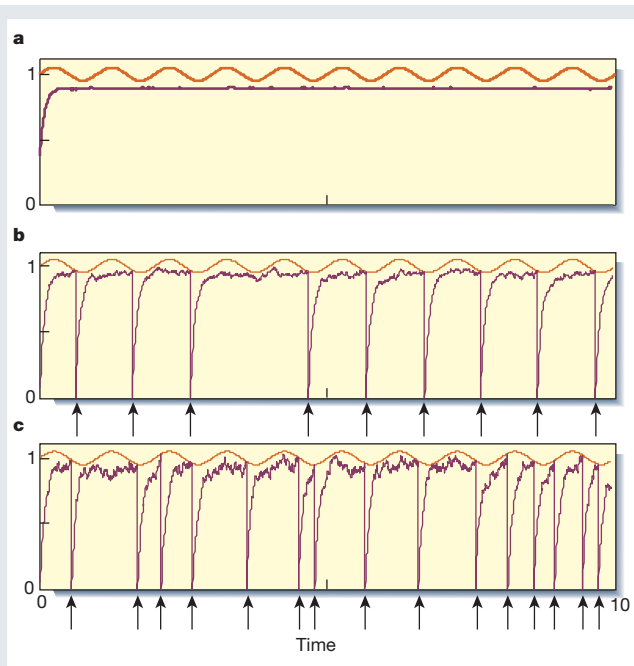


Figure 5 Schematic representation of a noisy, leaky integrate and fire model with three levels of noise. The activity rises to a sinusoidally modulated threshold following equation (1). **a**, The activity saturates and does not reach threshold. **b, c**, As the noise increases, the activity reaches threshold leading to an action potential (arrows). The ability to detect the frequency of the sine wave by the induced spiking activity will have a maximum efficiency at some intermediate level of noise. This illustrates the phenomenon of stochastic resonance^{55,57}.

there is a change in the dynamics from what is normal, rather than regularity or irregularity of the dynamics. Thus, generalizations such as “chaos may provide a healthy flexibility to the heart, brain, and other parts of the body”⁷⁶ must be considered warily.

Efforts are underway to develop better diagnostic and prognostic methods by analysis of dynamics of physiological rhythms. For example, in cardiology it would be extremely useful to have a measure that provides an independent predictor of significant cardiac events such as impending cardiac arrest. A variety of quantitative measures of cardiac activity derived from concepts introduced in nonlinear dynamics have been proposed as markers of increased risk of sudden death^{67–71}. However, because algorithms must be tested on large populations, it is difficult to design and implement clinical trials to test the utility of the various proposed measures. One recent initiative, the development of open databases that could serve as a repository of clinical data that are difficult and expensive to obtain, provides an emerging strategy that should prove indispensable for testing competing algorithms⁷². Likewise, the public availability of powerful data-processing methods should advance research⁷³.

Similar efforts at analysis of dynamics of time series are also underway in neurology. The earlier debates about whether the electroencephalogram displayed chaotic dynamics have been superseded by predictions of the onset of seizures using a variety of measures derived from nonlinear dynamics^{74–76}. Analysis of abnormalities in tremor and standing steadiness may provide a marker of heavy-metal toxicity^{77,78} or the early onset of Parkinson’s disease⁷⁹.

There are a number of ways in which knowledge about the synchronization of oscillators could be put to medical use. Because bodily functions show distinct periodicities, schedules for drug administration might be optimized. In cancer chemotherapy, treatments could be based on the circadian rhythm of cell division⁸⁰. Intriguing strategies to minimize the effects of jet lag have been developed, based on experimental studies of resetting the circadian

oscillator by modifying the exposure to light after travel^{1,81}, although I am unaware of any clinical studies that have assessed these proposals.

Medical devices that are used to regulate and artificially control cardiac and respiratory dynamics have been developed principally by engineers working with physicians. The empirical methods used to develop and test these devices have not included a detailed mathematical analysis of interactions of the physiological rhythm with the device. Indeed, devices usually have several adjustable parameters so that the physician can optimize the settings for operating the device based on direct testing in a patient. Recent work has investigated the application of algorithms motivated by nonlinear dynamics to control cardiac arrhythmias in humans (ref. 82 and D. J. Christini, K. M. Stein, S. M. Markowitz, S. Mittal, D. J. Slotwimer, M. A. Scheiner, S. Iwai and B. B. Lerman, unpublished results). Although there are not yet clinical applications, I anticipate that better understanding of the interactions between stimuli and physiological rhythms will lead to the development of better medical devices.

The recent identification of stochastic resonance in experimental systems has led to the suggestion that it might be useful to add noise artificially as a therapeutic measure. For example, in patients who have suffered strokes or peripheral nerve damage, detection tasks might be enhanced by addition of artificial noise to enhance tactile sensations⁶². Similarly, addition of sub-sensory mechanical noise applied to the soles of the feet of quietly standing healthy subjects reduced postural sway and seemed to stabilize the postural control (J. Niemi, A. Priplata, M. Salen, J. Harry and J. J. Collins, unpublished results). Another intriguing suggestion is that addition of noise might also enhance the efficiency of mechanical ventilators. The use of a variable inflation volume with occasional large inflations might act to prevent the collapse of alveoli and therefore maintain improved lung function in patients undergoing ventilation⁸³. Finally, low-level vibratory stimulation of a premature infant seemed to eliminate long apnoeic periods. Although the mechanism is unknown, it was hypothesized that a stable fixed point, corresponding to no breathing in the apnoeic infant, coexisted with the normal limit-cycle oscillation that corresponded to breathing. The vibration destabilized that fixed point thereby eliminating the apnoeic spells⁸⁴.

Conclusions

Physiological rhythms are generated by nonlinear dynamical systems. *In vitro* experimental systems often yield dynamics that can be successfully interpreted using both simplified and complicated mathematical models. These models make predictions about effects induced by parameter changes such as changing the frequency and amplitude of a periodic stimulus. However, rhythms in intact animals in an ambient environment have so far defied simple interpretation. Bodily rhythms such as the heart beat, respiration and cortical rhythms show complex dynamics, the function and origin of which are still poorly understood. Moreover, we do not understand if the complex dynamics are an essential feature, or if they are secondary to internal feedback and environmental fluctuations. Because of the complexity of biological systems and the huge jump in scale from a single ionic channel to the cell to the organ to the organism, for the foreseeable future all computer models will be gross approximations to the real system. In the physical sciences, scientific understanding has been expressed in elegant theoretical constructs and has led to revolutionary technological innovation. If the advances in understanding physiological rhythms will follow the same trajectory, then we are still just at the beginning. □

1. Winfree, A. T. *The Geometry of Biological Time* (Springer, New York, 1980; 2nd edn 2001).
2. Glass, L. & Mackey, M. C. *From Clocks to Chaos: The Rhythms of Life* (Princeton Univ. Press, Princeton, 1988).
3. Bassingthwaite, J. B., Liebowitch, L. S. & West, B. J. *Fractal Physiology* (Oxford Univ. Press, New York, 1994).
4. Keener, J. & Sneyd, J. *Mathematical Physiology* (Springer, New York, 1998).
5. Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol. (Lond.)* **117**, 500–544 (1952).
6. Aihara, K. & Matsumoto, G. in *Chaos in Biological Systems* (eds Degn, H., Holden, A. V. & Olson, L. F.) 121–131 (Plenum, New York, 1987).

7. Takahashi, N., Hanyu, Y., Musha, T., Kubo, R. & Matsumoto, G. Global bifurcation structure in periodically stimulated giant axons of squid. *Physica D* **43**, 318–334 (1990).
8. Kaplan, D. T. *et al.* Subthreshold dynamics in periodically stimulated squid giant axons. *Phys. Rev. Lett.* **76**, 4074–4077 (1996).
9. Ruelle, D. Where can one hope to profitably apply the ideas of chaos? *Phys. Today* **47**, 29–30 (July 1994).
10. Guevara, M. R., Glass, L. & Shrier, A. Phase-locking, period-doubling bifurcations and irregular dynamics in periodically stimulated cardiac cells. *Science* **214**, 1350–1353 (1981).
11. DeFelice, L. J. & Clay, J. R. in *Single-Channel Recording* Ch. 15 (eds Sakmann, B. & Neher, E.) 323–342 (Plenum, New York, 1983).
12. DeFelice, L. J. & Isaac, A. Chaotic states in a random world: relationships between the nonlinear differential equations of excitability and the stochastic properties of ion channels. *J. Stat. Phys.* **70**, 339–354 (1993).
13. Colquhoun, D. & Hawkes, A. G. in *Single-Channel Recording* Ch. 20 (eds Sakmann, B. & Neher, E.) 397–482 (Plenum, New York, 1995).
14. Wilders, R. & Jongsma, H. J. Beating irregularity of single pacemaker cells isolated from the rabbit sinoatrial node. *Biophys. J.* **60**, 2601–2613 (1993).
15. Guevara, M. R. & Lewis, T. A minimal single channel model for the regularity of beating of the sinoatrial node. *Chaos* **5**, 174–183 (1995).
16. Rabinovich, M. I. & Abarbanel, H. D. I. The role of chaos in neural systems. *Neuroscience* **87**, 5–14 (1998).
17. De Vries, G., Sherman, A. & Zhu, H. R. Diffusively coupled bursters: effects of cell heterogeneity. *B. Math. Biol.* **60**, 1167–1200 (1998).
18. Camm, A. J. *et al.* Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Circulation* **93**, 1043–1065 (1996).
19. Poon, C.-S. & Merrill, C. K. Decrease of cardiac chaos in congestive heart failure. *Nature* **389**, 492–495 (1998).
20. Kanter, J. K., Holstein-Rathlou, N.-H. & Agner, E. Lack of evidence for low-dimensional chaos in heart rate variability. *J. Cardiovasc. Electrophysiol.* **5**, 591–601 (1994).
21. Costa, M. *et al.* No evidence of chaos in the heart rate variability of normal and cardiac transplant human subjects. *J. Cardiovasc. Electrophysiol.* **10**, 1350–1357 (1999).
22. Kobayashi, M. & Musha, T. I/f fluctuation of heartbeat period. *IEEE Trans. Biomed. Eng.* **29**, 456–457 (1982).
23. Peng, C. K. *et al.* Long-range anticorrelations and non-Gaussian behavior of the heartbeat. *Phys. Rev. Lett.* **70**, 1343–1346 (1993).
24. Turcott, R. G. & Teich, M. Fractal character of the electrocardiogram: distinguishing heart failure and normal patients. *Ann. Biomed. Eng.* **24**, 269–293 (1996).
25. Ivanov, P. C. *et al.* Multifractality in human heartbeat dynamics. *Nature* **399**, 461–465 (1999).
26. Bak, P. *How Nature Works: The Science of Self-Organized Criticality* (Copernicus, New York, 1996).
27. Roach, D., Sheldon, A., Wilson, W. & Sheldon, R. Temporally localized contributions to measures of large-scale heart rate variability. *Am. J. Physiol.* **274** (Heart Circ. Physiol. H43), H1465–H1471 (1998).
28. Kiloh, L. *et al.* *Clinical Electroencephalography* (Butterworths, London, 1981).
29. Babloyantz, A. & Destexhe, A. Low-dimensional chaos in an instance of epilepsy. *Proc. Natl Acad. Sci. USA* **83**, 3513–3517 (1986).
30. Jansen, B. H. & Brandt, M. E. (eds) *Nonlinear Dynamical Analysis of the EEG* (World Scientific, Singapore, 1993).
31. Haurie, C., Dale, D. C. & Mackey, M. C. Cyclical neutropenia and other periodic hematological disorders: a review of mechanisms and mathematical models. *Blood* **92**, 2629–2640 (1998).
32. Prank, K. *et al.* Nonlinear dynamics in pulsatile secretion of parathyroid hormone in human subjects. *Chaos* **5**, 76–81 (1995).
33. Prank, K. *et al.* Self-organized segmentation of time series: separating growth hormone secretion in acromegaly from normal controls. *Biophys. J.* **70**, 2540–2547 (1996).
34. Veldman, R. G. *et al.* Increased episodic release and disorderliness of prolactin secretion in both micro- and macroprolactinomas. *Eur. J. Endocrinol.* **140**, 192–200 (1999).
35. Golubitsky, M., Stewart, I., Buono, P. L. & Collins, J. J. Symmetry in locomotor central pattern generators and animal gaits. *Nature* **401**, 693–695 (1999).
36. Strogatz, S. H. *The Mathematical Structure of the Human Sleep-Wake Cycle. Lecture Notes in Biomathematics* Vol. 69 (Springer, Berlin, 1986).
37. Daan, S., Beersma, D. G. M. & Borbely, A. A. Timing of human sleep: recovery process gated by a circadian pacemaker. *Am. J. Physiol.* **246**, R161–R178 (1984).
38. Jewett, M. E. & Kronauer, R. E. Refinement of a limit cycle oscillator model of the effects of light on the human circadian pacemaker. *J. Theor. Biol.* **192**, 455–465 (1998).
39. Leloup, J. C. & Goldbeter, A. A model for circadian rhythms in *Drosophila* incorporating the formation of a complex between the PER and TIM proteins. *J. Biol. Rhythms* **13**, 70–87 (1998).
40. Petriello, G. A. & Glass, L. A theory for phase-locking of respiration in cats to a mechanical ventilator. *Am. J. Physiol.* **246** (Regulat. Integrat. Comp. Physiol. 15), R311–R320 (1984).
41. Graves, C., Glass, L., Laporta, D., Meloche, R. & Grassino, A. Respiratory phase-locking during mechanical ventilation in anesthetized human subjects. *Am. J. Physiol.* **250** (Regulat. Integrat. Comp. Physiol. 19), R902–R909 (1986).
42. Simoin, P. M., Habel, A. M., Daubenspeck, J. A. & Leiter, J. C. Vagal feedback in the entrainment of respiration to mechanical ventilation in sleeping humans. *J. Appl. Physiol.* **89**, 760–769 (2000).
43. Schäfer, C., Rosenblum, M. G., Kurths, J. & Abel, H.-H. Heartbeat synchronized with ventilation. *Nature* **392**, 239–240 (1998).
44. Murthy, V. N. & Fetz, E. E. Coherent 25- to 35-Hz oscillations in the sensorimotor cortex of awake behaving monkeys. *Proc. Natl Acad. Sci. USA* **89**, 5670–5674 (1992).
45. Gray, C. M., König, P., Engel, A. K. & Singer, W. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **338**, 334–337 (1989).
46. Butera, R. J., Rinzel, J. & Smith, J. C. Models of respiratory rhythm generation in the pre-Bötzinger complex. II. Populations of coupled pacemaker neurons. *J. Neurophysiol.* **82**, 398–415 (1999).
47. Kopell, N., Ermentrout, G. B., Whittington, M. A. & Traub, R. D. Gamma rhythms and beta rhythms have different synchronization properties. *Proc. Natl Acad. Sci. USA* **97**, 1867–1872 (2000).
48. Achermann, P. & Kunz, H. Modeling circadian rhythm generation in the suprachiasmatic nucleus with locally coupled self-sustained oscillators: phase shifts and phase response curves. *J. Biol. Rhythms* **14**, 460–468 (1999).
49. Mirollo, R. E. & Strogatz, S. H. Synchronization properties of pulse-coupled biological oscillators. *SIAM J. Appl. Math.* **50**, 1645–1662 (1990).
50. Shinbrot, T. Progress in the control of chaos. *Adv. Phys.* **95**, 73–111 (1995).
51. Christini, D. J., Hall, K., Collins, J. J. & Glass, L. in *Handbook of Biological Physics* Vol. 4: *Neuroinformatics, Neural Modelling* (eds Moss, F. & Gielen, S.) 205–227 (Elsevier, Amsterdam, 2000).
52. Garfinkel, A., Spano, M. L., Ditto, W. L. & Weiss, J. N. Experimental control of cardiac chaos. *Science* **257**, 1230–1235 (1992).
53. Schiff, S. J. *et al.* Controlling chaos in the brain. *Nature* **370**, 615–620 (1994).
54. Skarda, C. A. & Freeman, W. J. How brains make chaos in order to make sense of the world. *Behav. Brain Sci.* **10**, 161–195 (1987).
55. Gammaitoni, L., Hänggi, P., Jung, P. & Marchesoni, F. Stochastic resonance. *Rev. Mod. Phys.* **70**, 223–288 (1998).
56. Longtin, A., Bulsara, A. & Moss, F. Time-interval sequences in bistable systems and the noise-induced transmission of information by sensory neurons. *Phys. Rev. Lett.* **67**, 656–659 (1991).
57. Longtin, A. Mechanisms of stochastic phase locking. *Chaos* **5**, 209–215 (1995).
58. Douglass, J. K., Wilkens, L., Pantazidou, E. & Moss, F. Stochastic resonance: noise-enhanced information transfer in crayfish mechanoreceptors. *Nature* **365**, 337 (1993).
59. Keener, J. P., Hoppensteadt, F. C. & Rinzel, J. Integrate-and-fire models of nerve membrane response to oscillatory input. *SIAM J. Appl. Math.* **41**, 503–517 (1981).
60. Collins, J. J., Imhoff, T. T. & Grigg, P. Noise-enhanced information transmission in rat SA1 cutaneous mechanoreceptors via aperiodic stochastic resonance. *J. Neurophysiol.* **76**, 642–645 (1996).
61. Greenwood, P. E., Ward, L. M., Russell, D. F., Neiman, A. & Moss, F. Stochastic resonance enhances the electrosensory information available to paddlefish for prey capture. *Phys. Rev. Lett.* **84**, 4773–4776 (2000).
62. Collins, J. J., Imhoff, T. T. & Grigg, P. Noise-mediated enhancements and decrements in human tactile sensation. *Phys. Rev. E* **56**, 923–926 (1997).
63. Simonotto, E. *et al.* Visual perception of stochastic resonance. *Phys. Rev. Lett.* **78**, 1186–1189 (1997).
64. Richardson, K. A., Imhoff, T. T., Grigg, P. & Collins, J. J. Using electrical noise to enhance the ability of humans to detect subthreshold mechanical stimuli. *Chaos* **8**, 599–603 (1998).
65. Mackey, M. C. & Glass, L. Oscillation and chaos in physiological control systems. *Science* **197**, 287–289 (1977).
66. Pool, R. Is it healthy to be chaotic? *Science* **243**, 604–607 (1989).
67. Skinner, J. E., Pratt, C. M. & Vybiral, T. A reduction in the correlation dimension of heartbeat intervals precedes imminent ventricular fibrillation in human subjects. *Am. Heart J.* **125**, 731–743 (1993).
68. Rosenbaum, D. S. *et al.* Electrical alternans and vulnerability to ventricular arrhythmias. *N. Engl. J. Med.* **330**, 235–241 (1994).
69. Huikuri, H. V. *et al.* Fractal correlation properties of R-R interval dynamics and mortality in patients with depressed left ventricular function after an acute myocardial infarction. *Circulation* **101**, 47–53 (2000).
70. Klingenstein, T. *et al.* Predictive value of T-wave alternans for arrhythmic events in patients with congestive heart failure. *Lancet* **356**, 651–652 (2000).
71. Lipsitz, L. A. Age-related changes in the “complexity” of cardiovascular dynamics: a potential marker of vulnerability to disease. *Chaos* **5**, 102–109 (1995).
72. Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, e215–e220 (2000). [Circulation Electronic Pages <<http://circ.ahajournals.org/cgi/content/abstract/101/23/e215>> (13 June 2000); see also <<http://www.physionet.org>>.]
73. Hegger, R., Kantz, H. & Schreiber, T. Practical implementation of nonlinear time series methods: the TISEAN package. *Chaos* **9**, 413–435 (1999); <<http://www.mpiikps-dresden.mpg.de/~tisean/>>.
74. Schiff, S. J. Forecasting brain storms. *Nature Med.* **4**, 1117–1118 (1998).
75. Lehnertz, K. & Elger, C. E. Can epileptic seizures be predicted? Evidence from nonlinear time series analysis of brain activity. *Phys. Rev. Lett.* **80**, 5019–5022 (1998).
76. Martinier, J. *et al.* Epileptic seizures can be anticipated by non-linear analysis. *Nature Med.* **4**, 1173–1176 (1998).
77. Beuter, A. & Edwards, R. Tremor in Cree subjects exposed to methylmercury: a preliminary study. *Neurotoxicol. Teratol.* **20**, 581–589 (1998).
78. Gerr, F., Letz, R. & Green, R. C. Relationships between quantitative measures and neurologist’s clinical rating of tremor and standing steadiness in two epidemiological studies. *NeuroToxicology* **21**, 753–760 (2000).
79. Edwards, R. & Beuter, A. Using time domain characteristics to discriminate physiologic and parkinsonian tremors. *J. Clin. Neurophysiol.* **17**, 87–100 (2000).
80. Mormont, M. C. & Lévi, F. Circadian-system alterations during cancer processes: a review. *Int. J. Cancer* **70**, 241–247 (1997).
81. Daan, S. & Lewy, A. J. Scheduled exposure to daylight: a potential strategy to reduce jet lag following transmeridian flight. *Psychopharmacol. Bull.* **20**, 566–588 (1984).
82. Ditto, W. L. *et al.* Control of human atrial fibrillation. *Int. J. Bif. Chaos* **10**, 593–601 (2000).
83. Suki, B. *et al.* Life-support systems benefits from noise. *Nature* **393**, 127–128 (1998).
84. Paydarfar, D. & Buerkle, D. M. Sporadic apnea: paradoxical transformation to eupnea by perturbations that inhibit inspiration. *Med. Hypotheses* **49**, 19–26 (1997).
85. Lipsitz, L. A. *et al.* Heart rate and respiratory dynamics on ascent to high altitude. *Br. Heart J.* **74**, 390–396 (1995).
86. Hausdorff, J. M. *et al.* Altered fractal dynamics of gait: reduced stride-interval correlations with aging and Huntington’s disease. *J. Appl. Physiol.* **82**, 262–269 (1997).
87. Hausdorff, J. M. *et al.* Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis. *J. Appl. Physiol.* **88**, 2045–2053 (2000).
88. Glass, L. Cardiac arrhythmias and circle maps—a classical problem. *Chaos* **1**, 13–19 (1991).
89. Wiggins, S. *Introduction to Applied Nonlinear Dynamical Systems and Chaos* (Springer, New York, 1990).
90. Glass, L. & Mackey, M. C. A simple model for phase locking of biological oscillators. *J. Math. Biol.* **7**, 339–352 (1979).
91. Perez, R. & Glass, L. Bistability, period doubling bifurcations and chaos in a periodically forced oscillator. *Phys. Lett.* **90A**, 441–443 (1982).
92. Guevara, M. R. & Glass, L. Phase-locking, period-doubling bifurcations and chaos in a mathematical model of a periodically driven biological oscillator: a theory for the entrainment of biological oscillators and the generation of cardiac dysrhythmias. *J. Math. Biol.* **14**, 1–23 (1982).

Acknowledgements

Thanks to M. R. Guevara, A. L. Goldberger, J. J. Collins, J. Milton and E. Cooper for helpful conversations; J. Lacuna, Y. Nagai and T. Inoue for assistance with the figures; and J. Gallas (Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil) for providing the colour representations of the locking zones in Figs 3c and 4c. My research has been supported by NSERC, MRC, MITACS, Canadian Heart and Stroke Foundation, FCAR and the Research Resource for Complex Physiologic Signals (NIH).

biotic resistance and virulence factors. The finding of a connection between bacterial conjugation and biofilm formation suggests that an important ecological consequence of the use of antibiotics and biocides in clinical medicine and agriculture may have been the selection of plasmid-bearing strains that are more likely to form a biofilm. Because biofilms are a common cause of persistent nosocomial infections that are difficult to eradicate owing to innate physiological properties¹⁴, this aspect may prove to be of relevant medical significance in addition to the conjugational spread of virulence factors themselves. □

Methods

Bacterial strains and plasmids

Bacterial strains are listed in Table 1 and were provided by the Collection of the Institut Pasteur (<http://www.pasteur.fr/applications/CIP/>) and the Unité des Agents Antibactériens (P. Courvalin and G. Gerbaud).

Biofilm

All experiments were performed in triplicate in 0.4% glucose M63B1 minimal medium at 37 °C. Continuous 60-ml microfermenters with four liquid and gas sampling ports (Pasteur Institute's Laboratory of Fermentation) were configured as continuous-flow culture bioreactors with a 40 ml h⁻¹ flow rate (*F*). 10⁸ bacterial inocula from overnight precultures grown in glucose minimal medium with required antibiotics were used to inoculate microfermenters, which were then cultivated for 3–48 h. The culture volume (*V*) was constant and the imposed dilution rate (*D*) was $D = F/V = 0.66 \text{ h}^{-1}$. Hence, the theoretical generation time (*T*) required for constant density culture in the microfermenter was $T = \ln 2/D = 1.05 \text{ h}$. The average generation time calculated in exponential batch culture for *E. coli* strains MG1655 and BM21 was 1.3 h. Therefore, the high input rate of fresh, diluting medium used in our experimental model was imposed to avoid any significant planktonic growth. Stirring was assured by aeration with sterile pressed air (0.3 bar). Submerged, removable Pyrex slides (total area of 22.4 cm²) served as growth substratum.

Microscopy and image analysis

Biofilm development was recorded with a Nikon Coolpix 950 digital camera. Epifluorescence, phase contrast and transmitted light microscopy were acquired with a Leitz Dialux 20EB microscope equipped with $\times 25$ to $\times 100$ objectives. Scanning confocal microscopy was performed at the confocal microscopy station of the Pasteur Institute.

Non-polar deletion of the *traA* gene

A non-polar mutation deleting the entire *traA* gene was created by allelic exchange¹⁵ using the primers TraAGB-5': 5'-AGGGAGGCGAGATAAGAGGAAGATATAACATTAAATACA CTCTAGTTTTATTTCATTTATCCGAAATTGAGGTAACTTATGAAAGCCACGTTGTG TCTCAA-3' and TraAGBnp-3': 5'-GCGGCTCTGGTTCAGTGTTCGCGGAAACG ATATTTCTTAAGTTTATCTCGTCTCCGACATCGTTTATTTCTTGTTAGAAAAA CTATCGAGCA-3', and *aphA* gene (kanamycin resistance) from Tn903 as template. The mutation was verified by PCR analysis.

Cloning of the *traA* gene

pTraA was constructed by PCR amplification of *traA* from strain TG1 using the primers TraAecorbs-5': 5'-AAAGAATTGCGAATTGAGGTAACTTATGAATGC-3' and TraAH3-3': 5'-CCCAAGCTTCGTTTATTTCTCTGTCAGAG-3'. I verified the nucleotide sequence of the construction.

Biofilm co-inoculation procedures

A preculture of a recipient strain BM21 (nal^r (nalidixic acid resistant)) was inoculated for 24 h and then co-inoculated with MG1655-S R1 (St^r (streptomycin resistant), Ap^r (ampicillin resistant) and Km^r (kanamycin resistant)) for another 24 h. Pyrex slides were removed and centrifuged in 15 ml of fresh M63B1 medium for 1 min. The number of colony-forming units was determined by plating serial dilutions of the resuspensions on medium supplemented with nalidixic acid (for recipient BM21 scoring), streptomycin, ampicillin, kanamycin (for donor MG1655-S R1 scoring), nalidixic acid, ampicillin and kanamycin (for BM21 R1 transconjugant scoring), and without antibiotic (for total cell scoring). Co-inoculation with BM21 strains carrying the plasmids described in Table 1 were generated using MG1655-S as recipient bacteria.

Received 8 March; accepted 6 June 2001.

- de la Cruz, I. & Davies, I. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.* **8**, 128–133 (2000).
- Lederberg, J. & Tatum, E. L. Gene recombination in *Escherichia coli*. *Nature* **158**, 558 (1946).
- Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
- Costerton, J. W., Lewandowski, Z., Caldwell, D. E., Korber, D. R. & Lappin-Scott, H. M. Microbial biofilms. *Annu. Rev. Microbiol.* **49**, 711–745 (1995).
- Hausner, M. & Wuertz, S. High rates of conjugation in bacterial biofilms as determined by quantitative in situ analysis. *Appl. Environ. Microbiol.* **65**, 3710–3713 (1999).

- Christensen, B. B. *et al.* Establishment of new genetic traits in a microbial biofilm community. *Appl. Environ. Microbiol.* **64**, 2247–2255 (1998).
- O'Toole, G. A. & Kolter, R. Flagellar and twitching motility are necessary for *Pseudomonas aeruginosa* biofilm development. *Mol. Microbiol.* **30**, 295–304 (1998).
- Pratt, L. A. & Kolter, R. Genetic analyses of bacterial biofilm formation. *Curr. Opin. Microbiol.* **2**, 598–603 (1999).
- Frost, L. S., Ippen-Ihler, K. & Skurray, R. A. Analysis of the sequence and gene products of the transfer region of the F sex factor. *Microbiol. Rev.* **58**, 162–210 (1994).
- Ried, G. & Henning, U. A unique amino acid substitution in the outer membrane protein OmpA causes conjugation deficiency in *Escherichia coli* K-12. *FEBS Lett.* **223**, 387–390 (1987).
- Lundquist, P. D. & Levin, B. R. Transitory derepression and the maintenance of conjugative plasmids. *Genetics* **113**, 483–497 (1986).
- Watnick, P. & Kolter, R. Biofilm, city of microbes. *J. Bacteriol.* **182**, 2675–2679 (2000).
- Bergstrom, C. T., Lipsitch, M. & Levin, B. R. Natural selection, infectious transfer and the existence conditions for bacterial plasmids. *Genetics* **155**, 1505–1519 (2000).
- Costerton, J. W., Stewart, P. S. & Greenberg, E. P. Bacterial biofilms: a common cause of persistent infections. *Science* **284**, 1318–1322 (1999).
- Chaveroche, M. K., Ghigo, J. M. & d'Enfert, C. A rapid method for efficient gene replacement in the filamentous fungus *aspergillus nidulans*. *Nucleic Acids Res.* **28**, E97 (2000).
- Courturier, M., Bex, F., Bergquist, P. L. & Maas, W. K. Identification and classification of bacterial plasmids. *Microbiol. Rev.* **52**, 375–395 (1988).
- Bukhari, A. I., Shapiro, J. A. & Adhya, S. L. (eds) *DNA Insertion Elements, Plasmids and Episomes* 601–656 (Cold Spring Harbor Laboratory, Cold Spring Harbor, 1977).
- Bradley, D. E., Taylor, D. E. & Cohen, D. R. Specification of surface mating systems among conjugative drug resistance plasmids in *Escherichia coli* K-12. *J. Bacteriol.* **143**, 1466–1460 (1980).
- Frost, L. S. in *Bacterial conjugation* (ed. Clewell, D. B.) 189–221 (Plenum, New York, 1993).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

I am grateful to C. Wandersman for support and constant interest during the course of this work. I also thank D. Mazel, G. Gerbaud and P. Courvalin for providing the plasmids and strains used in this study; A. Idja and P. Roux for technical assistance; R. Longin for providing facilities of the Pasteur Institute Laboratory of Fermentations; and D. Mazel, E. Stewart, D. A. Rowe-Magnus and P. Delepleire for helpful discussions and critical reading of the manuscript. This work was supported by grants from the Programme de Recherche Fondamentale en Microbiologie et Maladie Infectieuses, réseau Infections Nosocomiales (MENRT) and the Pasteur Institute.

Correspondence and requests for material should be addressed to the author (e-mail: jmghigo@pasteur.fr).

Force can overcome object geometry in the perception of shape through active touch

Gabriel Robles-De-La-Torre & Vincent Hayward

McGill University, Center for Intelligent Machines, Montréal, Canada H3A 2A7

Haptic (touch) perception normally entails an active exploration of object surfaces over time. This is called active touch^{1–3}. When exploring the shape of an object, we experience both geometrical⁴ and force cues. For example, when sliding a finger across a surface with a rigid bump on it, the finger moves over the bump while being opposed by a force whose direction and magnitude are related to the slope of the bump⁵. The steeper the bump, the stronger the resistance. Geometrical and force cues are correlated, but it has been commonly assumed that shape perception relies on object geometry alone. Here we show that regardless of surface geometry, subjects identified and located shape features on the basis of force cues or their correlates. Using paradoxical stimuli, for example combining the force cues of a bump with the geometry of a hole, we found that subjects perceived a bump. Conversely, when combining the force cues of a hole with the geometry of a bump, subjects typically perceived a hole.

In two experiments, human subjects explored surfaces by touch

using an apparatus that allowed us to separate force cues from surface geometry. Subjects explored the shape of surfaces through a manipulandum placed behind a curtain (Fig. 1). Subjects pressed down on the plate of the manipulandum with their index fingers while smoothly rolling it on the surfaces. There were three interchangeable physical surfaces. One surface was flat, one had a bump, and another had a hole; the latter two had a gaussian profile that was 0.3 cm high/deep and 4 cm wide. The plate was constrained to remain horizontal, and its vertical position was entirely determined by the geometry of the physical surfaces. A force-feedback haptic interface could produce a horizontal force as a function of the measured vertical component F_{sy} (the force applied vertically by the subjects, Fig. 2a) and of the plate's horizontal position (Fig. 1, Methods). The force of the interface interacted with the force returned by the physical surface. The interface's force was called a 'virtual surface' because it provided the same horizontal force component that an equivalent physical surface would return, regardless of the manipulandum's vertical position. For example, when subjects explored a physical bump or hole (Fig. 2a and b), they experienced a horizontal force F_{px} . When subjects explored a flat surface combined with a virtual bump or hole, they experienced the same horizontal force F_{px} (Fig. 2c, grey, dashed curve), but the manipulandum moved in a straight line (Fig. 2c, black line). In another example, the horizontal force component from a physical bump was cancelled out ('force-masked', Methods) by a spatially aligned virtual hole (Fig. 2d). Here, the vertical position of the manipulandum followed the geometry of the bump (Fig. 2d, black curve), but subjects did not experience horizontal force components (Fig. 2d, grey, dashed line). The manipulandum would stay in equilibrium on a slope regardless of how much the subjects pushed on it, just as if they were exploring a flat surface.

In experiment 1, we used the paradoxical stimuli just described to investigate whether subjects classified and located shape features by following geometrical or horizontal force cues. Subjects were tested using seven different conditions. In condition 1, only flat surfaces were presented. In condition 2, the flat surfaces were combined with virtual bumps, and in condition 3 with virtual holes (Fig. 2c). In condition 4 there were only ordinary physical bumps, and in condition 5 ordinary physical holes (Fig. 2b). In condition 6, the physical bumps were force-masked by virtual holes, and control virtual bumps were randomly placed elsewhere (Fig. 2e). Condition 7 mirrored condition 6 (Fig. 2e). The positions of physical and virtual surfaces were uncorrelated in conditions 6 and 7. The probability of classifying a stimulus as a hole or a bump was calculated for each subject under all conditions. The stimulus localization performance of the subjects was measured by the correlation between stimulus location and subjects' positioning of the manipulandum (Methods).

The flat surfaces were classified as flat (Fig. 3a, b, condition 1).

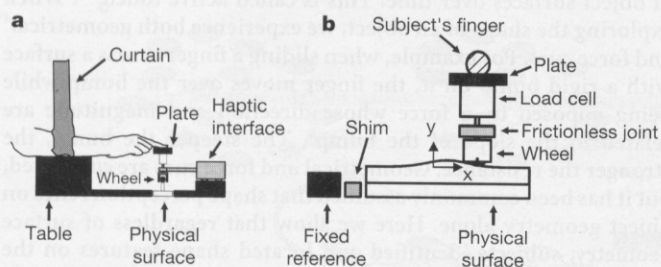


Figure 1 Side (a) and front (b) views of the apparatus. Subjects pressed down on the manipulandum's plate and rolled it sideways (x -axis in b) to explore the shape of an interchangeable physical surface. The entire workspace was hidden from view behind a black curtain. The manipulandum was connected to a haptic interface (Methods) that operated silently.

Virtual bumps, combined with the flat surface, were classified as bumps ($P < 0.001$; Fig. 3a, condition 2), and were accurately located (Fig. 3c). Virtual holes, combined with the flat surface, were also identified as holes ($P < 0.001$; Fig. 3b, condition 3) and were precisely located (Fig. 3d). When physical holes or bumps were presented alone, subjects also easily identified and located them (Fig. 3a–d, conditions 4 and 5). There was no significant difference in subjects' identification and localization performance when exploring physical or virtual surfaces. However, when force-masked physical holes or bumps were combined with virtual surfaces, subjects' features localization was affected drastically. Subjects' identification performance remained the same (Fig. 3a, b, conditions 6 and 7), but most subjects tracked the control virtual

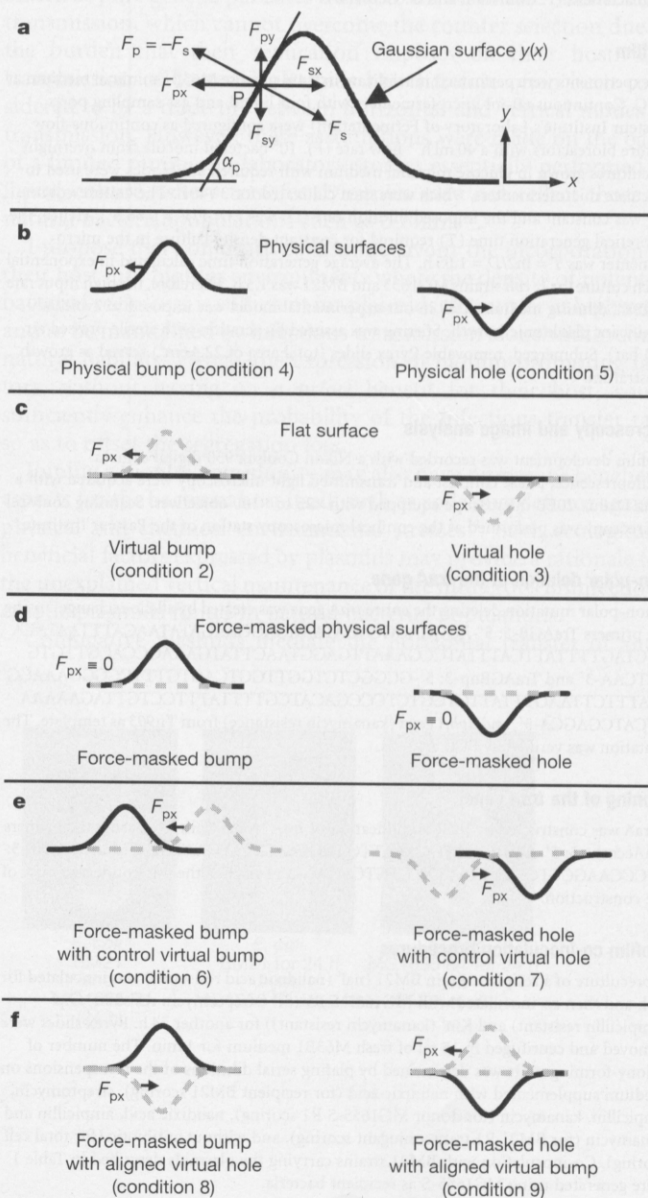


Figure 2 Experimental conditions 2–9. When the haptic interface was turned off, subjects experienced F_{px} , the horizontal force component that depends on surface geometry (a, b). Equivalent forces (c–f) were produced by the interface to generate stimuli with uncorrelated force (grey, dashed curves) and geometric cues (black curves). The force cues from virtual bump/hole stimuli (c) were equivalent to those of physical bumps/holes, but provided the geometrical cues of a flat surface. Force-masked stimuli (d–f) were used to create physical bumps/holes with spatially uncorrelated cues (e), physical holes with a bump's force cues, and physical bumps with a hole's force cues (f).

surfaces instead of the force-masked physical holes or bumps (Fig. 3c, d, conditions 6 and 7, thick circles, $P < 0.009$ for virtual bumps and $P < 0.001$ for virtual holes; see also Supplementary Information). The physical bumps or holes provided geometric but not horizontal force information to subjects, whereas the virtual surfaces provided force but not geometrical information. Given the subjects' instructions (see Methods), this suggests that the control virtual holes or bumps seemed deeper or bigger than the force-masked physical holes or bumps. Horizontal force was used instead of geometry to identify and locate shape features. This happened when geometrical information was absent (conditions 2 and 3) and when object geometry did not correlate with force (conditions 6 and 7). One subject displayed a significant localization performance when tracking force-masked bumps (Fig. 3c, condition 6). The position of the force-masked and virtual bumps presented to this subject had a spurious, non-significant correlation of 0.32 that explained the subject's significant tracking. Only this subject was exposed to spurious correlations. However, the subject clearly followed the virtual bumps and not the force-masked physical bumps (Supplementary Information). Some subjects located the force-masked shapes in some trials, and the control virtual shapes in other trials. These were subjects with significant localization correlations equal to or below 0.65 when tracking virtual bumps (Fig. 3c, condition 6; Supplementary Information), or virtual and force-masked holes (Fig. 3d, condition 7; Supplementary Information).

If subjects relied on force information to identify and locate shape features, then an easily identified and located physical bump should be perceptually transformed into a hole by combining it with the force cues of a hole. An equivalent perceptual reversal should happen for a physical hole. Experiment 2 explored this idea with a new group of subjects. Again, they were tested using condition 1 (flat surfaces, no virtual surface), conditions 2 and 3 (virtual surfaces, Fig. 2c), and conditions 4 and 5 (physical surface alone,

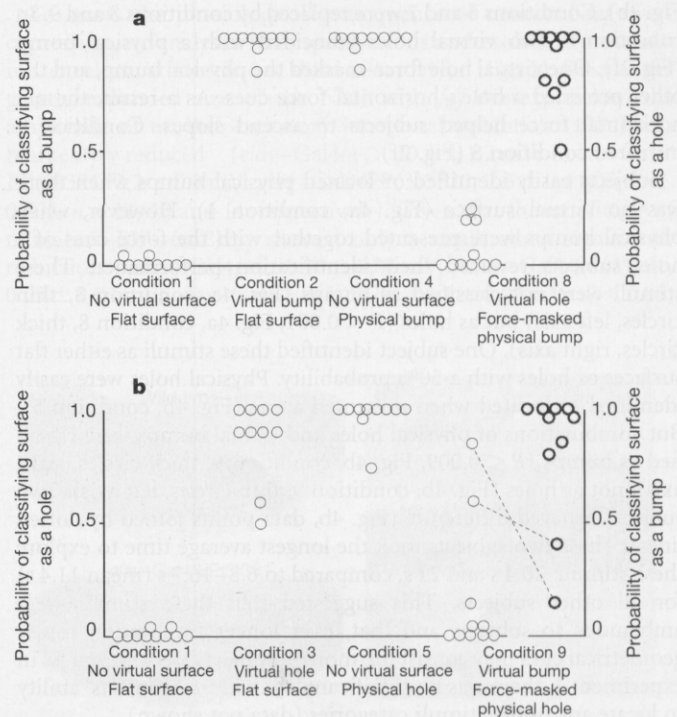


Figure 4 Force perceptually transformed geometrical shape. Each symbol represents data from a single subject. Subjects easily classified physical and virtual surfaces (a, b, conditions 1–5) as bumps or holes. However, when force-masked physical bumps coincided with virtual holes, stimuli were classified as holes (a, condition 8, thick circles, right axis). A mirrored perceptual change occurred when force-masked physical holes coincided with virtual bumps (b, condition 9, thick circles, right axis). Two subjects identified the physical holes but took the longest average times to make a judgement (text). Dotted lines join data points corresponding to the same subjects.

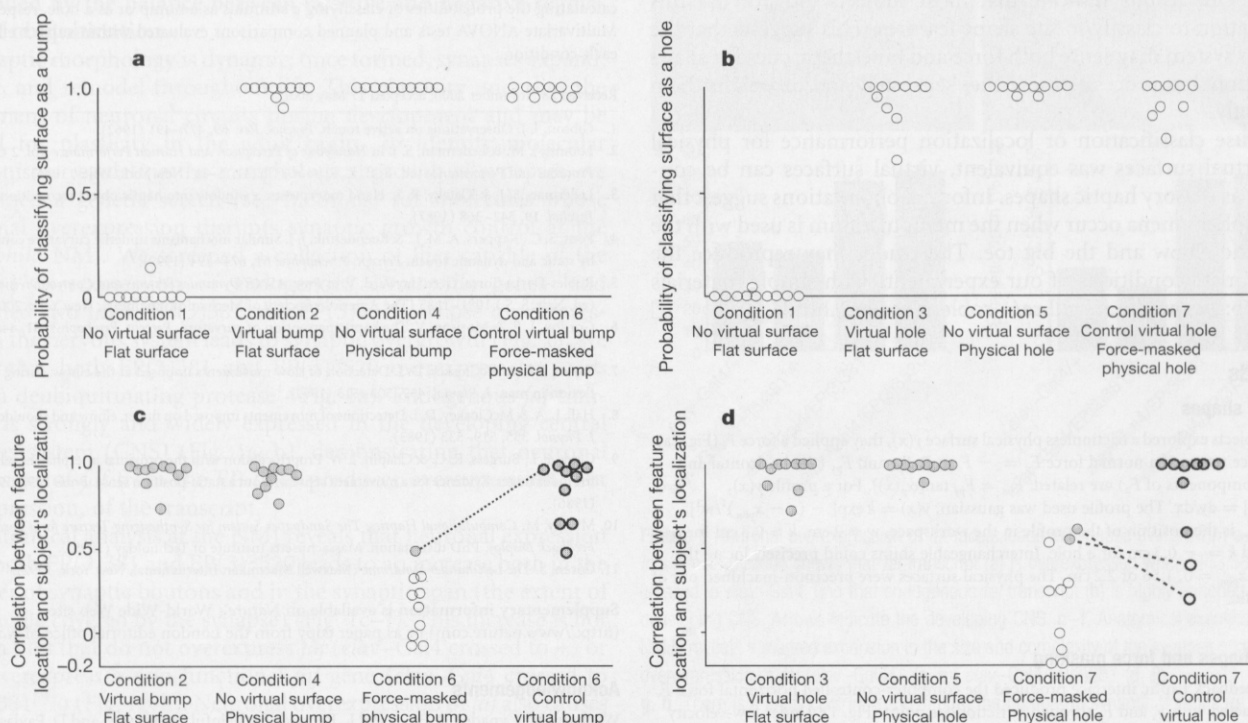


Figure 3 Subjects used force to identify or locate shape features. Each symbol represents data from one subject. Subjects easily classified stimuli as bumps or holes when a virtual surface or flat physical surface combination was presented, or when virtual and physical surfaces coexisted, but not when a flat surface was presented alone (a, b). Subjects tracked the position of physical and virtual surfaces (c, d, conditions 2–5). However,

when virtual surfaces and force-masked physical surfaces were presented side by side, most subjects tracked the control virtual surfaces (c, d, conditions 6 and 7, thick circles). Dotted lines link the data from the same subject. Grey circles indicate significant correlations ($P < 0.05$) between stimulus position and subjects' localization.

Fig. 2b). Conditions 6 and 7 were replaced by conditions 8 and 9. In condition 8, two virtual holes coincided with a physical bump (Fig. 2f). One virtual hole force-masked the physical bump, and the other provided a hole's horizontal force cues. As a result, the net horizontal force helped subjects to ascend slopes. Condition 9 mirrored condition 8 (Fig. 2f).

Subjects easily identified or located physical bumps when there was no virtual surface (Fig. 4a, condition 4). However, when physical bumps were presented together with the force cues of a hole, subjects reversed their identification performance. These stimuli were not classified as bumps (Fig. 4a, condition 8, thin circles, left axis) but as holes ($P < 0.001$, Fig. 4a, condition 8, thick circles, right axis). One subject identified these stimuli as either flat surfaces or holes with a 50% probability. Physical holes were easily identified or located when presented alone (Fig. 4b, condition 5). But combinations of physical holes and virtual bumps were classified as bumps ($P < 0.009$, Fig. 4b, condition 9, thick circles, right axis), not as holes (Fig. 4b, condition 9, thin circles, left axis). Two subjects behaved differently (Fig. 4b, data points joined by dotted lines). These two subjects took the longest average time to explore these stimuli: 20.4 s and 21 s, compared to 6.8–16.7 s (mean 11.4 s) for all other subjects. This suggested that these stimuli were ambiguous to subjects and that, over longer exploration times, geometrical cues may contribute more to subjects' perception. As in experiment 1, there was no significant difference in subjects' ability to locate any of the stimuli categories (data not shown).

Our results indicate that force or force-related information (for example, hand/finger velocity) can overcome geometrical information, such as slope, to determine shape perception through active touch. This happened for maximum slope differences well above reported thresholds: 5.71 degrees compared to 0.57 (ref. 6) and 3–4 degrees (ref. 4). Our subjects used a combination of finger, wrist and forearm movements to explore stimuli. Kinesthesia research^{7–9} suggests that subjects may have easily detected these movements, particularly those of the metacarpophalangeal joint of the index finger⁹. Our results indicate that most subjects did not use this information to classify/locate shape features. This suggests that the nervous system may sense both force and kinesthetic cues for shape perception, but processes these cues separately and/or weighs them differently.

Because classification or localization performance for physical and virtual surfaces was equivalent, virtual surfaces can be considered as illusory haptic shapes. Informal observations suggest that similar phenomena occur when the manipulandum is used with the wrist, the elbow and the big toe. The reader may reproduce the approximate conditions of our experiments with simple materials (see <http://www.cim.mcgill.ca/~roblesg/vsdemo.html>). □

Methods

Physical shapes

When subjects explored a frictionless physical surface $y(x)$, they applied a force F_x (Fig. 2a). The surface returned a normal force $F_y = -F_x$ at P . F_{yx} and F_{yy} (the horizontal and vertical components of F_y) are related: $F_{yx} = F_{yy} \tan[\alpha_y(x)]$. For a profile $y(x)$, $\tan[\alpha_y(x)] = dy/dx$. The profile used was gaussian, $y(x) = k \exp[-(x - x_{plac})^2/w^2]$, where x_{plac} is the position of the profile in the workspace, $w = 2$ cm, $k = 0.3$ cm for a bump and $k = -0.3$ cm for a hole. Interchangeable shims could precisely locate the profile at $x_{plac} = 0, 1.56$ or 2.2 cm. The physical surfaces were precision-machined out of hard plastic.

Virtual shapes and force masking

A force-feedback haptic interface produced the computer-controlled horizontal force F_x that interacted with F_x and F_y through a frictionless point (Fig. 1). Under low-velocity conditions, F_x , F_y and F_z balanced: $F_x + F_y + F_z = 0$. Along the x and y directions, $F_x + F_{yx} + F_z = 0$ and $F_y + F_{yy} = 0$. A load cell (Fig. 1) measured $F_{yy} = -F_y$ (Fig. 2a) and the haptic interface measured x , the horizontal position of the plate. The device was programmed to produce $F_x = F_{yx} \tan[\alpha_y(x - x_{plac})] + F_{yy} \tan[\alpha_y(x - x_{plac})]$. Because the force experienced by a subject in the horizontal direction was $F_x = -F_{yx} - F_y$, the first term of F_x defined a virtual surface that cancelled out ('force-masked') the horizontal component of the force returned by a physical surface located at x_{plac} . The second term of

F_x defined a virtual surface located at x_{plac} . The physical/virtual surface shapes were aligned when $x_{plac} = x_{plac}$. In conditions 2 and 3 (both experiments), the first term of F_x was not used.

Force-feedback haptic interface

A PenCAT/Pro (Immersion Canada Inc.) interface produced F_x and measured x . The interface had negligible friction and operated silently. Manipulandum's position/forces were sampled at 1 kHz. The force feedback was updated at the same rate. A manipulandum similar to ours was theoretically described by Minsky¹⁰ to illustrate a physical model of simulated textures.

Experimental procedure

Subjects gave informed consent, were paid to participate and naive to the purpose of the experiment, did not have calluses on their right index finger nor report any hand injury/disease. Ten right-handed subjects participated in experiment 1; four males and six females, ages 19–31. Ten new right-handed subjects participated in experiment 2; six males and four females, ages 18–31. Handedness was evaluated by using a standard questionnaire¹¹. In each trial, a stimulus was randomly selected from an experimental condition. The manipulandum's position was randomly set either to the right or left end of the workspace. A subject's index finger was placed on the centre of the manipulandum's plate (Fig. 1). Subjects were instructed to locate the highest/lowest point of the highest/deepest perceived bump/hole. There was no time limit for exploration, but proceeding quickly was encouraged. After locating the feature, subjects did not change the manipulandum's position. Subjects ended the trial by pressing a button on a computer keyboard to identify the located feature. Buttons were labelled 'bump', 'hole' and 'flat'. No feedback was given. Subjects withdrew their fingers from the manipulandum after pressing the button. Only physical surfaces were presented during 25 practice trials. During practice, the shims were not used and the surface's position was randomly varied across the workspace. Each subject proceeded to complete 140 experimental trials. Each experimental condition was tested 20 times. Subjects had periodic breaks after every 50 trials, but could rest at any time. A typical experiment lasted about one hour and fifty minutes.

Data analysis

For each experimental condition, subjects' localization performance was measured by the Pearson correlation coefficient between physical/virtual surface location and subjects' final manipulandum position. If a stimulus provided the force cues of a bump (virtual or physical) but was classified as a hole or as a flat surface, the trial was not used to compute the correlation. Conversely, a trial was not used for correlation calculations if the stimulus provided the force cues of a hole (virtual or physical), but was classified as a bump or as a flat surface. Subjects' classification performance was measured by calculating the probabilities of classifying a stimulus as a bump or as a hole, respectively. Multivariate ANOVA tests and planned comparisons evaluated within-subject effects of each condition.

Received 28 November 2000; accepted 17 May 2001.

- Gibson, J. I. Observations on active touch. *Psychol. Rev.* **69**, 477–491 (1962).
- Loomis, J. M. & Lederman, S. J. in *Handbook of Perception and Human Performance* Vol. 2 *Cognitive Processes and Performance* (ed. Boff, K. R. et al.) Ch. 33, 1–41 (New York, Wiley, 1986).
- Lederman, S. J. & Klatzky, R. L. Hand movements: a window into haptic object recognition. *Cogn. Psychol.* **19**, 342–368 (1987).
- Pont, S. C., Kappers, A. M. L. & Koenderink, J. J. Similar mechanisms underlie curvature comparison by static and dynamic touch. *Percept. Psychophys.* **61**, 874–894 (1999).
- Robles-De-La-Torre, G. & Hayward, V. in *Proc. ASME Dynamics Systems and Control Division* Vol. 2 (ed. Nair, S. S.) 1081–1085 (The American Society of Mechanical Engineers, New York, 2000).
- Gordon, I. E. & Morison, V. The haptic perception of curvature. *Percept. Psychophys.* **31**, 446–450 (1982).
- Taylor, J. L. & McCloskey, D. I. Detection of slow movements imposed at the elbow during active flexion in man. *J. Physiol.* **457**, 503–513 (1992).
- Hall, L. A. & McCloskey, D. I. Detections of movements imposed on finger, elbow and shoulder joints. *J. Physiol.* **335**, 519–533 (1983).
- Clark, F. J., Burgess, R. C. & Chapin, J. W. Proprioception with the proximal interphalangeal joint of the index finger. Evidence for a movement sense without a static-position sense. *Brain* **109**, 1195–1208 (1986).
- Minsky, M. *Computational Haptics: The Sandpaper System for Synthesizing Texture for a Force-Feedback Display*. PhD dissertation, Massachusetts Institute of Technology (1995).
- Coren, S. *The Left-Hander Syndrome* (Maxwell Macmillan International, New York, 1992).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank M. Cynader, B. Frost and L. Requaft for helpful comments, and D. Pavlasek and J. Boka for help in designing/building the manipulandum and surfaces. The research was funded by Canada's Network of Centers of Excellence programme, Institute for Robotics and Intelligent Systems, and the Natural Sciences and Engineering Research Council of Canada.

Correspondence and requests for materials should be addressed to G.R. (e-mail: roblesg@cim.mcgill.ca).

Scars of the Wigner Function

Fabricio Toscano,^{1,2} Marcus A. M. de Aguiar,³ and Alfredo M. Ozorio de Almeida¹

¹*Centro Brasileiro de Pesquisas Físicas, Rua Xavier Sigaud 150, 22290-180, RJ, Rio de Janeiro, Brazil*

²*Instituto de Física, Universidade do Estado do Rio de Janeiro, R. São Francisco Xavier 524, 20559-900, RJ, Rio de Janeiro, Brazil*

³*Instituto de Física, "Gleb Wataghin," Universidade Estadual de Campinas, 83-970, Campinas, SP, Brazil*
(Received 10 December 1999; revised manuscript received 10 August 2000)

We propose a picture of Wigner function scars as a sequence of concentric rings along a two-dimensional surface inside a periodic orbit. This is verified for a two-dimensional plane that contains a classical hyperbolic orbit of a Hamiltonian system with 2 degrees of freedom. The stationary wave functions are the familiar mixture of scarred and random waves, but the spectral average of the Wigner functions in part of the plane is nearly that of a harmonic oscillator and individual states are also remarkably regular. These results are interpreted in terms of the semiclassical picture of chords and centers.

DOI: 10.1103/PhysRevLett.86.59

PACS numbers: 05.45.Mt, 03.65.Sq

Sixteen years have passed since Heller [1] detected scars of periodic orbits in individual eigenfunctions of chaotic systems. Explanations in terms of wave packets [1] or the semiclassical Green function [2,3] do predict an enhancement of intensity near the projection of a Bohr-quantized periodic orbit. However, such theories apply to a collective superposition of states near this quantization condition. In spite of some quite striking visual evidence, the issue of scarring for individual states is confused by the fact that different branches of the same periodic orbit may overlap so that their contributions interfere in the position space and this must be superposed on the expected random wave background for the chaotic state. This has led to the need to make quantitative assessments of scarring strength [4]. We here show that the phase space picture can provide sharper qualitative evidence of the influence of a periodic orbit on its Wigner functions.

The old conjecture of Berry and Voros [5] that the Wigner functions for eigenstates of chaotic Hamiltonians are concentrated on the corresponding classical energy shells is widely disseminated [6], as it is compatible with Schnirelman's theorem [7]. Nonetheless, no such restriction arises in the more recent semiclassical theory for the mixture of these states over a narrow energy window, i.e., the spectral Wigner function [8,9]. Indeed, the computations presented in this Letter show that individual Wigner functions, as well as their energy average can oscillate with large amplitudes deep inside the energy shell.

At points $\mathbf{x} = (\mathbf{q}, \mathbf{p})$ inside the energy shell, the semiclassical spectral Wigner function is [8,9]

$$W(\mathbf{x}, E, \varepsilon) = \sum_j A_j e^{-\varepsilon t_j / \hbar} \cos \left[\frac{S_j(\mathbf{x}, E)}{\hbar} + \gamma_j \right]. \quad (1)$$

The sum is over all the trajectory segments on the E shell with end points, $\mathbf{x}_{j\pm}$, centered on \mathbf{x} , as sketched in Fig. 1. The action is merely the symplectic area, $S_j = \oint \mathbf{p} \cdot d\mathbf{q}$, for the circuit taken along the orbit from \mathbf{x}_{j-} to \mathbf{x}_{j+} and closed by the chord $-\xi_j(\mathbf{x})$. The time of traversal for the stretch along the trajectory is t_j and ε is the width of the

energy window over which we average individual Wigner functions $W_n(\mathbf{x})$, i.e.,

$$W(\mathbf{x}, E, \varepsilon) = (2\pi\hbar)^{D/2} \sum_n \frac{\varepsilon/\pi}{(E - E_n)^2 + \varepsilon^2} W_n(\mathbf{x}), \quad (2)$$

where D is the dimension of the phase space (\mathbf{q}, \mathbf{p}) . We shall not be concerned with the Maslov phase γ_j or with the amplitude $A_j(\mathbf{x}, E)$, except to note that these are purely classical quantities that vary smoothly with \mathbf{x} inside the shell, as compared with the high frequency oscillations of the cosine factor.

The energy shell itself is a caustic, because the chord $\xi \rightarrow 0$ as \mathbf{x} approaches the shell. In this limit, the contributing trajectories either shrink to a point or they are very close to a periodic orbit. The modification of (1) leads to the Berry theory of scars [10], further refined in [8], for evaluation points \mathbf{x} close to the energy shell.

The point here is that large open segments of periodic orbits can be constructed by adding multiple windings to a primitive, small segment of a periodic orbit. If conditions for phase coherence, to be discussed, are satisfied, we can then obtain a scar deep inside the energy shell. This scar is located at the two-dimensional *central surface* constructed by the centers of all the chords with end points on the periodic orbit. For the case where the phase space

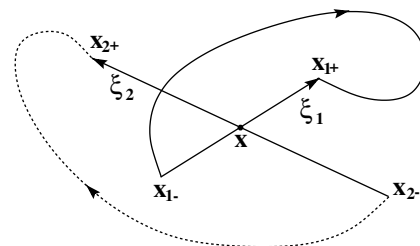


FIG. 1. In general, there are many orbit segments with end points \mathbf{x}_{j-} and \mathbf{x}_{j+} centered on a given point \mathbf{x} . The circuit, closed by the chord $-\xi_j = \mathbf{x}_{j-} - \mathbf{x}_{j+}$, defines the phase of the semiclassical contribution to (1).

dimension $D = 2$, the central surface is the phase space region enclosed by the convex hull of the orbit that here is the energy shell. For $D \geq 4$, the two-dimensional central surfaces are still bounded by one-dimensional periodic orbits that lie within the higher dimensional energy shell. In the simplest case where the orbit is plane and convex, the central surface coincides with the part of the plane within the orbit just as in the case where $D = 2$. If the orbit is not plane, then the central surface will be more complicated and it may exhibit singularities and self-intersections. In general, it will not coincide everywhere with the invariant surface formed by changing continuously the energy of this periodic orbit.

We emphasize that these scars of the Wigner function that correspond to a very recognizable oscillatory pattern have special localization properties in phase space of dimension $D \geq 4$. Previous explorations of phase space representations of eigenstates have been too blunt to discern these fine features. The work of Feingold *et al.* [11] involves a projection from a three-dimensional section, which is fine for the Husimi function, but washes away the details of an oscillatory Wigner function. On the other hand, Agam and Fishman [12] restrict their investigation to the neighborhood of the orbit and, hence, to the energy shell. This is just the edge of the central surface.

In this Letter we tested our prediction for the simplest case of an unstable periodic orbit lying in a two-dimensional plane. Accidentally this plane is a symmetry plane; however, we have checked that the enhanced amplitude of the spectral Wigner function accompanies the distortion of the orbit due to a nonlinear canonical transformation. Since the Wigner function is not invariant in this case, we thus obtain an essentially new system, in which the new central surface is not a symmetry surface. Thus, our scarring effect cannot be attributed to a special property of the symmetry plane. This case is essentially different from recent studies of higher dimensional systems where the scars arise over marginally stable invariant planes that are indeed special [13].

The chord structure for segments of a periodic orbit is sketched in Fig. 2. This is similar to a system with $D = 2$, for which Berry [14] showed that there is only one chord for most points \mathbf{x} . However, we must now distinguish between the chords $\xi_{\text{in}} = -\xi_{\text{out}}$, and the sum in (1) includes each different winding of the periodic orbit for primitive orbit segments, “in” and “out,” associated with the two respective chords. Evidently all these “in” and “out” contributions build up if the energy is close to one of the values for which this periodic orbit is Bohr quantized. We show in [15] that the condition for these two contributions to be in phase is the same as that for a maximal contribution to the Gutzwiller trace formula [6]. Therefore, near the quantization energy, the sum of all the contributions in (1), owing to chords in the periodic orbit, can be approximated by an expression analogous to Berry’s simplest semiclassical approximation [14,16] for the Wigner function in $D = 2$

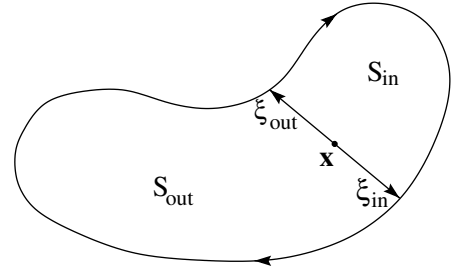


FIG. 2. If the end points lie on a periodic orbit, we can add an infinite number of windings to the two shortest primitive segments, “in” and “out,” divided by the chords $\xi_{\text{in}} = -\xi_{\text{out}}$. All the points \mathbf{x} , centers of chords in the periodic orbit, form a two-dimensional *central surface*. If the orbit is plane (as in the graph) the central surface is the part of the plane enclosed by the convex hull of the orbit.

systems, but now evaluated over points \mathbf{x} in the two-dimensional central surface. Hence, in this case, we have an enhanced contribution to the spectral Wigner function with the phase $S_{\text{in}}(\mathbf{x})/\hbar$ and the Maslov phase $\gamma = 0$. The successive contours $S_{\text{in}}(\mathbf{x}) = \text{const}$ thus determine rings of constant phase along the central surface. Therefore, we predict rings of positive and negative amplitude superimposed on the background of contributions from uncorrelated trajectory segments that also contribute to the spectral Wigner function. The edge of this system of rings along the orbit itself is the object of the Berry theory [10], though it does not deal with multiple windings.

We have tested this prediction for the Nelson Hamiltonian ($D = 4$) [17,18],

$$H(\mathbf{q}, \mathbf{p}) = (p_1^2 + p_2^2)/2 + 0.05q_1^2 + (q_2 - q_1^2/2)^2. \quad (3)$$

Evidently the plane $q_1 = p_1 = 0$ is classically invariant, so we obtain an isolated periodic orbit on this plane for each energy (the vertical family). This is then, the simplest case where the central surface is flat—the central plane. The restriction of the Hamiltonian to this plane defines a harmonic oscillator, so there is a single chord with end points in the periodic orbit (except for sign) for each point (p_2, q_2) inside the orbit, except at $q_2 = p_2 = 0$.

Using the method of [18] we computed the wave functions $\langle \mathbf{q} | n \rangle$ for the eigenstates of (3) within the range of energies $0.821 \leq E \leq 0.836$ taking $\hbar = 0.05$. The Wigner function for each state was then calculated directly by the double integral over $\mathbf{Q} = (Q_1, Q_2)$:

$$W_n(\mathbf{q}, \mathbf{p}) = \int \frac{d\mathbf{Q}}{(2\pi\hbar)^2} \left\langle \mathbf{q} + \frac{\mathbf{Q}}{2} | n \right\rangle \left\langle n | \mathbf{q} - \frac{\mathbf{Q}}{2} \right\rangle \times e^{-i(\mathbf{p} \cdot \mathbf{Q}/\hbar)}. \quad (4)$$

In Fig. 3 we display the average over Wigner functions for the energy window chosen, surrounding the 11th Bohr energy for the vertical orbit that has Maslov index 3. This is more convenient than smoothing with the Lorentzian window in (2) and should produce essentially the same results.

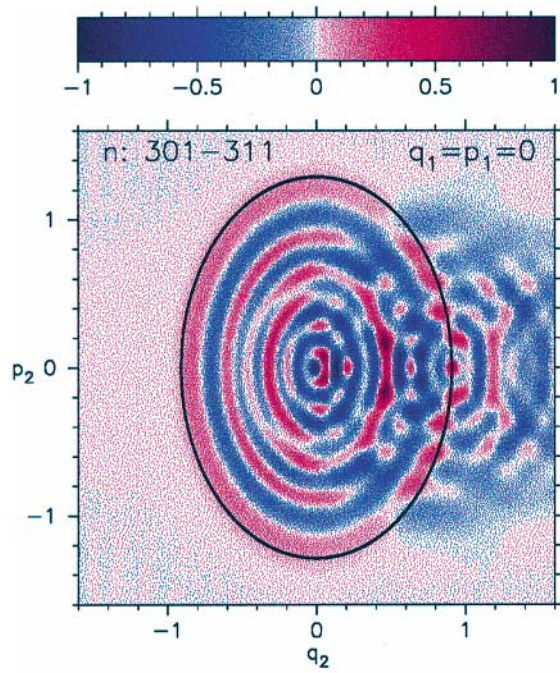


FIG. 3 (color). Averaged Wigner function on the (q_2, p_2) plane for an energy window containing 11 eigenstates centered on the 11th Bohr level. The thick dark line is the superposition of each energy shell in the plane (periodic orbit) for energies in the range considered; the middle one corresponding to the Bohr-quantization energy.

As an example, Fig. 4 shows some of the individual Wigner functions in this window for \mathbf{x} in the (q_2, p_2) plane. Immediately, we notice that all these Wigner functions are remarkably regular, roughly in the region $q_2 < 0$. Indeed, Fig. 3 coincides in this region with the Wigner function of the restricted Hamiltonian, also with respect to the precise phase of the oscillations. The individual Wigner functions follow the same phase contours, but there is a varying phase shift. If we extrapolate the semiclassical theory to energy smoothings of the order of the energy spacing, the conclusion is that there are no orbit segments, other than the periodic orbit itself, contributing to the Wigner functions on this surface up to the Heisenberg time. This is more dramatic for the state $n = 305$ whose Wigner function on all the central surface is almost the same as that of the restricted Hamiltonian (see Fig. 4). It is important to note that only this state seems to have a visual scar of the vertical orbit in position space, although superimposed to a random wave background.

For $q_2 > 0$, we have found other orbit segments that do not lie on the central plane, but which do contribute to the spectral Wigner function on it. Indeed, this is the case for segments of both the symmetric periodic orbits shown in Fig. 5 projected onto position space. Needless to say, the multiple windings of these periodic orbits could produce total contributions of the same order as the plane orbit, if they are nearly Bohr quantized. This supplies a qualitative explanation for the existence of interference patterns to the

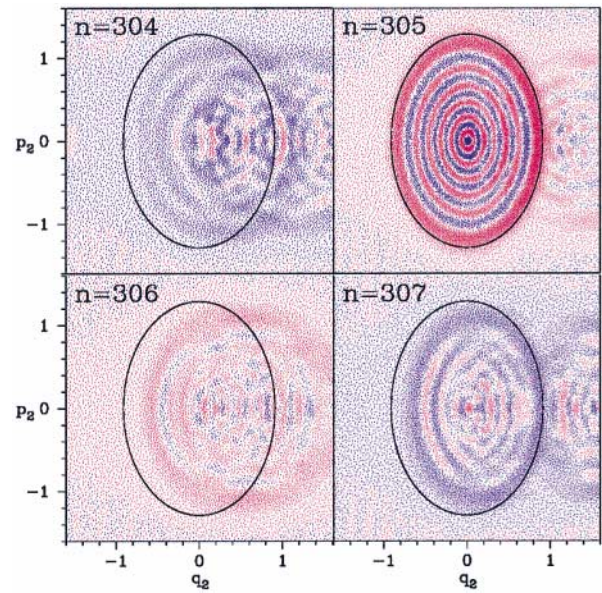


FIG. 4 (color). Some of the individual Wigner functions on the (q_2, p_2) plane that were averaged in Fig. 3. The plots are globally normalized to point out the enhanced amplitude of the scar for the $n = 305$ state. The eigenenergy of this state is at a distance $\approx 1.5\Delta$ from the Bohr energy (Δ the mean level spacing in the chosen average window).

left of the Wigner functions in Figs. 3 and 4. Notice that the second of the orbits in Fig. 5 reaches into a region with q_2 greater than is attained by the plane periodic orbit. So we account for Wigner functions reaching outside the energy shell (which coincides with the periodic orbit on the central plane). Note that these are only particular examples of the many symmetric periodic orbits whose central surface intersects the central plane along lines with $q_2 > 0$ [17,18]. Therefore, it will be hard to make quantitative predictions in this region in contrast to that where $q_2 < 0$, for which no such orbits of short period were found. The enhanced scar pattern of the Wigner function of the state $n = 305$ over all the central plane evidently washes out other possible contributions.

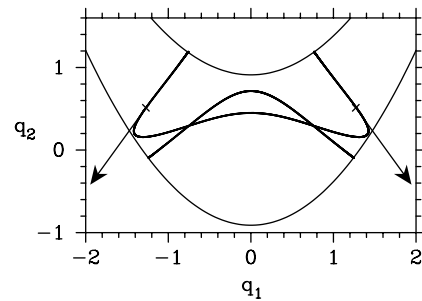


FIG. 5. Two of the many symmetric periodic orbits that have chords centered on the (q_2, p_2) plane. The graph is in position space, so that the momenta for the symmetric end points are displayed as vectors. All centers have $q_2 \geq 0$. The thin line is an equipotential of (3) for the energy of the 11th Bohr level.

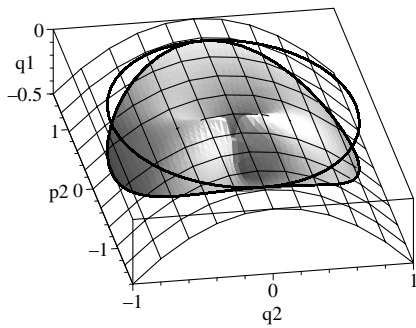


FIG. 6. The application of a particular nonlinear canonical transformation to the Hamiltonian (3) deforms the invariant plane $q_1 = p_1 = 0$ into the squared surface in the graph that also has $p_1 = 0$. The curve on this surface is the distortion of the original periodic orbit (ellipse displayed in the invariant plane). The grey surface is the new central surface for the deformed periodic orbit.

We show that the symmetry of the central plane is not responsible for the scarring observed by applying a nonlinear canonical transformation, described in [15], that deforms it into a new invariant surface (Fig. 6). Now the new central surface of the distorted periodic orbit no longer coincides everywhere with the invariant surface (Fig. 6), losing in this way any symmetry. We calculated the Wigner functions for this new system, over the central surface of the periodic orbit, for eigenstates within the same range of energies surrounding the 11th Bohr level. Both the average and the individual Wigner functions, projected over the (q_2, p_2) plane, display the same features as those shown in Figs. 3 and 4, respectively. Reference [15] also considers the thickness of scar surfaces.

Our main point is that individual trajectory segments do play a role in the spectral Wigner function. Indeed, phase coherent contributions from a periodic orbit can add up to scars in the form of concentric rings deep inside the energy shell. Even though the Berry-Voros hypothesis [5] ties in more intuitively with the concept of quantum ergodicity, it should be noted that the chord picture of Wigner functions does not contradict Shnirelman's theorem and subsequent exact results [7], because the oscillations make negligible contributions to integrals with smooth functions, i.e., projections that give true probability densities. However, the oscillations are unavoidable in the study of interference effects.

It would seem that the structure of interfering chords, from which the spectral Wigner function is built up, could generate only a messy picture where scars would be even harder to recognize than in wave functions for which quantitative methods are needed. We have now shown that the converse can be true in the simplest case, i.e., that the

relatively low dimension of the central surface can allow it to miss, over an appreciable region, the contribution of chords from other orbits. In this region we obtain a regular ring structure, not only for the energy average, but even for the pure Wigner functions shown in Fig. 4. Though we cannot yet predict why the relative weight of this particular periodic orbit should vary so markedly among the states close to the Bohr-quantized energy, we are now able to recognize in Fig. 4 a remarkable example of an individual scar.

This work was financed by CNPq-CLAF, FAPERJ, and Pronex-MCT.

-
- [1] E. J. Heller, Phys. Rev. Lett. **53**, 1515 (1984).
 - [2] E. B. Bogomolny, Physica (Amsterdam) **31D**, 169 (1988).
 - [3] A. M. Ozorio de Almeida, *Hamiltonian Systems: Chaos and Quantization* (Cambridge University Press, Cambridge, 1988).
 - [4] L. Kaplan and E. J. Heller, Ann. Phys. (N.Y.) **264**, 171 (1998); L. Kaplan, Nonlinearity **12**, R1 (1999).
 - [5] M. V. Berry, J. Phys. A **10**, 2083 (1977); A. Voros, in *Stochastic Behavior in Classical and Quantum Hamiltonian Systems*, edited by G. Casati and J. Ford, Lectures Notes in Physics Vol. 93 (Springer, Berlin, 1979), p. 326.
 - [6] M. C. Gutzwiller, *Chaos in Classical and Quantum Mechanics* (Springer-Verlag, Berlin, 1990).
 - [7] A. I. Schnirelman, Usp. Mat. Nauk **29**, 181 (1974); Y. Colin de Verdiere, Commun. Math. Phys. **102**, 497 (1985); S. Zelditch, Duke Math. J. **55**, 919 (1987).
 - [8] A. M. Ozorio de Almeida, Phys. Rep. **295**, 265 (1998).
 - [9] M. V. Berry, in *Chaos and Quantum Physics*, Proceedings of the Les Houches Summer School, Session LII, edited by M. J. Giannoni, A. Voros, and J. Zinn-Justin (North-Holland, Amsterdam, 1991), p. 251.
 - [10] M. V. Berry, Proc. R. Soc. London A **423**, 219 (1989).
 - [11] M. Feingold *et al.*, Phys. Lett. A **146**, 199 (1990).
 - [12] O. Agam and S. Fishman, J. Phys. A **26**, 2113 (1993); Phys. Rev. Lett. **73**, 806 (1994).
 - [13] T. Prosen, Phys. Lett. A **233**, 332 (1997); T. Papenbrock, T. H. Seligman, and H. A. Weidenmüller, Phys. Rev. Lett. **80**, 3057 (1998); T. Papenbrock and T. Prosen, Phys. Rev. Lett. **84**, 262 (2000).
 - [14] M. V. Berry, Philos. Trans. R. Soc. London **287**, 237 (1977).
 - [15] Work in preparation.
 - [16] F. Toscano and A. M. Ozorio de Almeida, J. Phys. A **32**, 6321 (1999).
 - [17] M. Baranger and K. T. R. Davies, Ann. Phys. (N.Y.) **177**, 330 (1987).
 - [18] D. Provost and M. Baranger, Phys. Rev. Lett. **71**, 662 (1993).

Classical and Quantum Hamiltonian Ratchets

Holger Schanz,¹ Marc-Felix Otto,¹ Roland Ketzmerick,¹ and Thomas Dittrich²

¹Max-Planck-Institut für Strömungsforschung und Institut für Nichtlineare Dynamik der Universität Göttingen, Bunsenstraße 10, 37073 Göttingen, Germany

²Departamento de Física, Universidad Nacional, Santafé de Bogotá, Colombia

(Received 21 November 2000; published 26 July 2001)

We explain the mechanism leading to directed chaotic transport in Hamiltonian systems with spatial and temporal periodicity. We show that a mixed phase space comprising both regular and chaotic motion is required and we derive a classical sum rule which allows one to predict the chaotic transport velocity from properties of regular phase-space components. Transport in quantum Hamiltonian ratchets arises by the same mechanism as long as uncertainty allows one to resolve the classical phase-space structure. We derive a quantum sum rule analogous to the classical one, based on the relation between quantum transport and band structure.

DOI: 10.1103/PhysRevLett.87.070601

PACS numbers: 05.60.-k, 05.45.Mt

Stimulated by the biological task of explaining the functioning of molecular motors, the study of ratchets [1] has widened to a general exploration of “self-organized” transport, i.e., transport without external bias, in nonlinear systems [2]. Along with this process, there has been a tendency to reduce the models under investigation from realistic biophysical machinery to the minimalist systems customary in nonlinear dynamics. External noise, for example, which originally served to account for the fluctuating environment of molecular motors, has been replaced by deterministic chaos. This required one to include inertia terms in the equations of motion, thus leaving the regime of overdamped dynamics and leading to deterministic inertia ratchets with dissipation [3,4]. It is then a consequent but radical step to abandon friction altogether. Indeed, transport in Hamiltonian ratchets was observed numerically if all symmetries were broken that generate to each trajectory a countermoving partner [5,6].

As a parallel development, the desire to realize ratchets in artificial, nanostructured electronic systems requires one to consider quantum effects [6,7]. Quantum Hamiltonian ratchets, however, have been studied only in the framework of one-band systems where no transport occurs [6].

In this paper we explain how a Hamiltonian ratchet works. We rely on methods which—although well established in studies of deterministic dynamics—have never before been applied to ratchets. We derive a classical and an analogous quantum sum rule for transport allowing the following conclusions: (i) Directed transport is a property associated with individual invariant sets of the dynamics. A necessary condition for nonzero transport is a mixed phase space with coexisting regular and chaotic regions. (ii) Transport in chaotic regions can be described quantitatively by using topological and further properties of adjacent regular regions only. (iii) Quantum transport persists for all times and approaches the classical transport when \hbar is small compared to the major invariant sets of the classical phase space.

We consider a Hamiltonian of the form $H(x, p, t) = T(p) + V(x, t)$, where $T(p)$ is the kinetic energy. The force $-V'$ is periodic in space and time, $V'(x + 1, t) = V'(x, t + 1) = V'(x, t)$, and has zero mean $\int_0^1 dt \int_0^1 dx V'(x, t) = 0$. Usually directed transport is demonstrated by following selected trajectories over very long times [5,6] or an ensemble of trajectories which generates spatial distributions as shown in Figs. 1a and 1c. While this is easily implemented numerically, it gives no clue about the origin of the transport (but see Ref. [9]). Instead, we shall exploit the periodicity of the dynamics with respect to space and time and analyze transport in terms of the invariant sets of phase space, *reduced to*

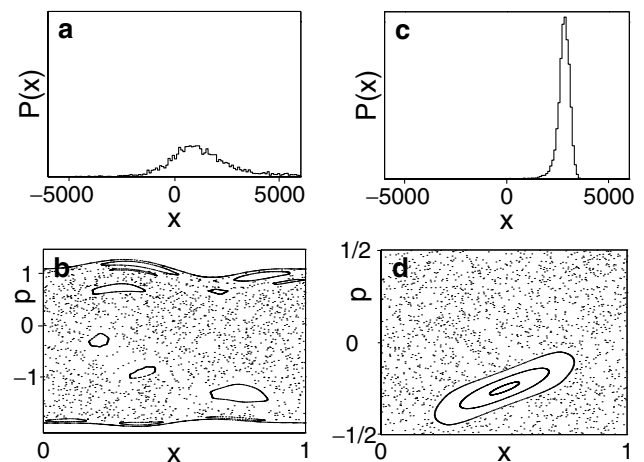


FIG. 1. (a) Spatial distribution $P(x)$ of a continuously driven system [8] after 20 000 time periods showing the directed transport in a Hamiltonian ratchet. Initially, 10^4 trajectories were started at random on the line $p = 0$, $x \in [0, 1]$ in the chaotic sea. (b) Poincaré section p vs x of a unit cell at integer times showing the main chaotic sea, the upper and lower limiting KAM-tori, and the major embedded regular islands. (c), (d) As (a), (b), but for the kicked Hamiltonian (4) showing a much more pronounced directed transport.

the spatiotemporal unit cell $x, t \in [0, 1)$. For any finite invariant set M we define ballistic transport as phase-space volume times average velocity expressed as

$$\tau_M = \int_0^1 dt \int_0^1 dx \int_{-\infty}^{+\infty} dp \chi_M(x, p, t) \frac{\partial H}{\partial p}, \quad (1)$$

where χ_M is the characteristic function of M . Transport is additive for the union of two or more disjoint invariant sets, i.e., for $M = \bigcup_i M_i$, with $M_i \cap M_j = \emptyset$ for all $i \neq j$, we have

$$\tau_M = \sum_i \tau_{M_i}. \quad (2)$$

This sum rule for Hamiltonian transport has far-reaching consequences to be discussed in the following.

For a generic Hamiltonian system, phase space is mixed and comprises an infinite number of minimal invariant sets of different types. For the sake of definiteness we will restrict the following discussion to the most interesting case of a chaotic region containing embedded regular islands (Fig. 1b). In any of these invariant sets the time-averaged velocity v_i is the same for almost all initial conditions (assuming ergodicity for chaotic components). Hence, for a chaotic region, $\tau_{\text{ch}} = A_{\text{ch}} v_{\text{ch}}$ with A_{ch} denoting its area in a stroboscopic Poincaré section. For an embedded island we have $\tau_i = A_i v_i$ where A_i includes the areas of the narrow chaotic layers inside the island and of the infinite hierarchy of island chains surrounding it because all these invariant sets share the same mean velocity. This velocity v_i is identical to the rational winding number $w_i = x_i/t_i$ of the stable fixed point at the center of the island. In extended phase space, this corresponds to a shift of the island by x_i spatial after t_i time periods. Typically, the chaotic set is bounded from above and below by two noncontractible KAM-tori $p_{a,b}(x, t)$ enclosing the spatial unit cell. Treating the phase-space region in between as the global invariant set M appearing on the left-hand side of Eq. (2), its transport τ_M is obtained from Eq. (1) as $\langle T \rangle_a - \langle T \rangle_b$ with $\langle T \rangle_{a,b} = \int_0^1 dt \int_0^1 dx T(p_{a,b}(x, t))$ denoting averages of the kinetic energy over the tori. Using the sum rule (2) we can now express transport of the chaotic region in terms of its adjoining regular components (KAM-tori and islands) as

$$A_{\text{ch}} v_{\text{ch}} = \langle T \rangle_a - \langle T \rangle_b - \sum_i A_i w_i. \quad (3)$$

This is our main result on classical transport in Hamiltonian ratchets. Not only does Eq. (3) provide an efficient method to determine the chaotic drift velocity, it also expresses the simple principle generating directed ballistic motion: Decomposing phase space into different invariant sets, these will in general have average velocities different from each other and also different from zero but related by the sum rule (2). Therefore a necessary condition for directed chaotic transport in Hamiltonian ratchets is a *mixed phase space*.

Lévy flights [10] are a characteristic feature of chaotic motion in a generic mixed phase space and indeed they were observed in Hamiltonian ratchets [5,6]. They reflect the slow exchange between subsets of a chaotic region, separated by leaky barriers [10]. As these subsets are not invariant, their contributions to Eq. (2) are contained in the contribution of the chaotic invariant set. Lévy flights lead to power-law tails in spatial distributions. For example, the asymmetric shapes of the peaks visible in Figs. 1a and 1c can be attributed to these tails. Notably, in Fig. 1c one clearly sees a mean transport to the right although the same data show no indication of a power-law tail in this direction. We stress that the sum rule allows one to predict the mean velocity of chaotic trajectories without any reference to such details of the chaotic dynamics. This suggests that Lévy flights are not a necessary element of the mechanism of chaotic transport in Hamiltonian ratchets.

In Ref. [5] it was shown that a necessary condition for directed transport is the breaking of all symmetries which to each trajectory generate a countermoving partner. For a chaotic set invariant under such a symmetry, this is in agreement with Eq. (3) because then the right-hand side vanishes identically. However, chaotic sets can also occur as symmetry-related pairs transporting in opposite directions. Moreover, if phase space cannot be decomposed into invariant subsets, e.g., for an ergodic system, there cannot be transport even with all symmetries broken.

Up to now we have considered only transport of invariant sets of the unit cell. For an arbitrary initial distribution transport is determined by projection onto these invariant sets [11]. Therefore, the location of an initial distribution within an invariant set is irrelevant. This applies also to the location within the temporal unit cell, i.e., to the question of phase dependence discussed in [12]. In particular, in case that the plane $p = 0$, $0 \leq x, t \leq 1$ is completely within the chaotic invariant set, any initial condition restricted to this plane will result in the same average transport. We now understand how a Hamiltonian ratchet makes particles initially at rest ($p = 0$) move with a predetermined mean velocity as, e.g., in Fig. 1a.

We have checked Eq. (3) numerically for a continuously driven system [8]. We determined the areas A_i and winding numbers w_i for the regular islands shown in the Poincaré section of Fig. 1b as well as $\langle T_a \rangle$ and $\langle T_b \rangle$ for the limiting KAM-tori, yielding $v_{\text{ch}} = 0.092 \pm 0.011$. The error estimate includes the uncertainty in the location of the bounding KAM-tori and the contribution from neglected small islands. The result is in agreement with the value $v_{\text{ch}} = 0.082 \pm 0.002$ determined with much more computational effort from the spatial distribution of 10^4 trajectories, started with $p = 0$ (Fig. 1a).

As a minimal model for directed chaotic transport in Hamiltonian ratchets, we propose a kicked Hamiltonian

$$H(x, p, t) = T(p) + V(x) \sum_n \delta(t - n). \quad (4)$$

It reduces the dynamics to a map for position and momentum $x_{n+1} = x_n + T'(p_n)$, $p_{n+1} = p_n - V'(x_{n+1})$, just after the kick. As an example we take a symmetric potential $V(x) = (x \bmod 1 - 1/2)^2/2$ and an asymmetric kinetic energy $T(p) = |p| + 3 \sin(2\pi p)/(4\pi^2)$. We consider the dynamics on a cylinder with transport along the x axis and $p \in [-1/2, +1/2]$ being a periodic variable. Figure 1d shows the Poincaré section for one unit cell. There are only two major invariant sets—a chaotic sea and a regular island centered around a periodic orbit with winding number $w_{\text{reg}} = -1$. According to Eq. (1), transport of the full phase space vanishes identically because of the periodic momentum variable. Applying the sum rule (2) the contributions to transport from the two invariant sets cancel exactly,

$$A_{\text{ch}} v_{\text{ch}} + A_{\text{reg}} w_{\text{reg}} = 0. \quad (5)$$

We find the transport velocity of the chaotic component as $v_{\text{ch}} = f_{\text{reg}}/(1 - f_{\text{reg}})$, where $f_{\text{reg}} = A_{\text{reg}}/(A_{\text{reg}} + A_{\text{ch}})$ denotes the relative area of the regular island. From Fig. 1d, $f_{\text{reg}} = 0.117 \pm 0.001$, thus $v_{\text{ch}} = 0.133 \pm 0.001$ in agreement with $v_{\text{ch}} = 0.1344 \pm 0.0003$ from the spatial distribution of Fig. 1c.

In order to extend our concept of directed transport in Hamiltonian systems to quantum ratchets, we first demonstrate by a numerical example that quantum Hamiltonian ratchets can work. Figure 2 shows that the average velocity of a wave packet initialized in the chaotic sea varies between 0 for large and the classical value v_{ch} for small values of \hbar . We explain this behavior in the following.

In analogy with our approach to classical transport, we consider the invariants of the quantum dynamics, the stationary states of the time-evolution operator over one period, i.e., $\hat{U} = e^{-i\hat{V}/\hbar} e^{-i\hat{T}/\hbar}$ for the kicked Hamiltonian Eq. (4). They satisfy $\hat{U}|\phi_{\alpha,k}\rangle = \exp[-2\pi i \epsilon_{\alpha}(k)] \times |\phi_{\alpha,k}\rangle$, with the quasienergy $\epsilon_{\alpha}(k) \in [0, 1]$ [13]. Similarly, spatial periodicity implies $\phi_{\alpha,k}(x+1, t) = \exp(2\pi i k) \phi_{\alpha,k}(x, t)$ with quasimomentum $k \in [0, 1]$ where h is chosen rational for systems periodic in p . Quantum transport is related to the expectation values in

the stationary states $\bar{v}_{\alpha,k} \equiv \langle\langle \phi_{\alpha,k} | \hat{v} | \phi_{\alpha,k} \rangle\rangle$ of the velocity operator $\hat{v} = \hat{T}'(\hat{p})$, where $\langle\langle \cdot \rangle\rangle = \int_0^1 dx \int_0^1 dt (\cdot)$. Using a generalization of the Hellmann-Feynman theorem to time-periodic systems [13], we express velocities by band slopes as

$$\bar{v}_{\alpha,k} = d\epsilon_{\alpha}(k)/dk. \quad (6)$$

This allows one to discuss quantum transport in terms of spectral properties. Examples for quasienergy band spectra are shown in Fig. 3 together with the corresponding velocity distributions. The semiclassical regime is characterized by the existence of two different types of bands and corresponding eigenstates [14,15]: Bands pertaining to regular states appear as straight lines in the spectrum, while the chaotic bands show oscillations and wide avoided crossings among themselves. Associating the terms chaotic and regular with the bands is supported by the Husimi representations of the corresponding eigenfunctions (insets in Fig. 3a). The new aspect introduced into this picture by directed chaotic transport is the overall slope of the chaotic bands.

Only on a coarse quasienergy scale, the two sets of bands appear to cross. On a sufficiently fine scale, all crossings are avoided. Consequently the actual bands change their character between regular and chaotic at each of the narrow crossings and have *no* overall slope. Switching from the latter (“adiabatic”) to the former (“diabatic”) viewpoint is a well-controlled procedure [14]. Formally, the behavior of the bands can be described in terms of their winding number (average slope) with respect to the periodic (ϵ, k) space: In the adiabatic as well as in the diabatic case, all quasienergy bands must close after an integer number of periods in the ϵ and k directions, so that their winding number \bar{w} must be rational. Clearly, in the adiabatic case, all winding numbers are zero. Going from the adiabatic to the diabatic case amounts to a mere reconnection of bands at the crossings, preserving the sum of winding numbers. Thus it must be zero also in the diabatic representation,

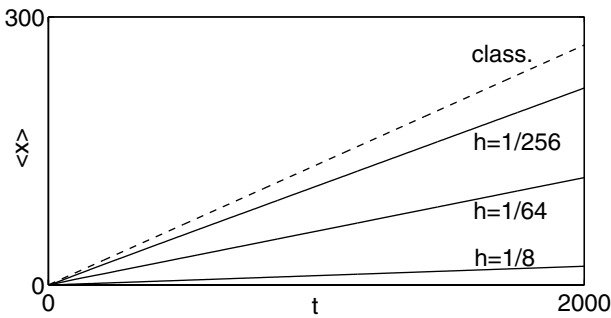


FIG. 2. Mean position vs time for a wave packet in system (4) initialized as the momentum eigenstate with $p = 0$ for various values of \hbar (full lines). For decreasing \hbar the classical prediction $v_{\text{ch}} t$ (dashed line) is approached.

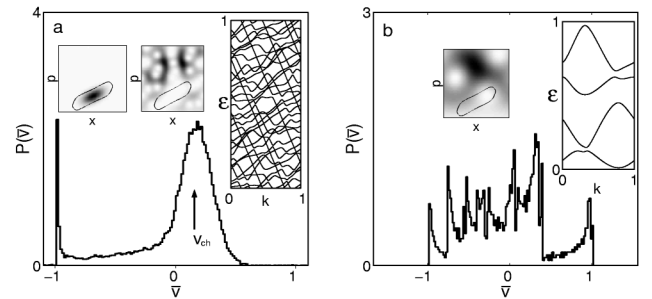


FIG. 3. Distribution of quantum velocities \bar{v} obtained according to Eq. (6) for (a) $\hbar = 1/128$ and (b) $\hbar = 1/4$. The right insets show band spectra for $\hbar = 1/32$ and $\hbar = 1/4$, respectively. The smaller insets are Husimi representations of characteristic wave functions together with the border of the classical regular island. In (a) the regular and the chaotic wave functions can be associated with the two peaks of $P(\bar{v})$ centered around $\bar{v}_{\text{reg}} = -1$ and \bar{v}_{ch} .

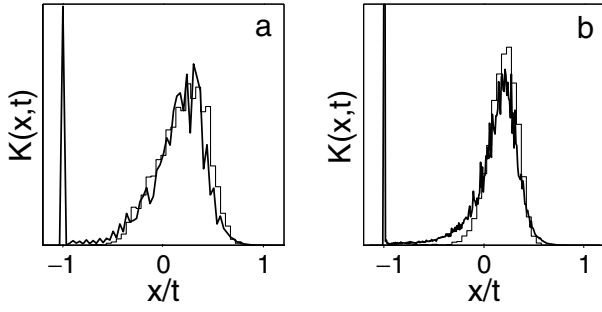


FIG. 4. Form factor $K(x, t)$ for $h = 1/128$ (thick line) at (a) $t = 30$ and (b) $t = 111$, the Heisenberg time of the chaotic component. The two distinct peaks centered around v_{reg} and v_{ch} are the fingerprints of directed transport. The time-dependent width of the chaotic peak follows the classical distribution $P^{(\text{ch})}(x, t)$ of chaotic trajectories (thin line).

$$\sum_{\alpha} \bar{w}_{\alpha}^{(\text{ch})} + \sum_{\alpha} \bar{w}_{\alpha}^{(\text{reg})} = 0. \quad (7)$$

This is the quantum-mechanical analog of the classical sum rule (5). Because of the localization of the regular states on tori inside the regular island, the winding number of the regular bands in (ϵ, k) space is in the semiclassical limit identical to the winding number in (x, t) space of the central periodic orbit, i.e., $\bar{w}_{\alpha}^{(\text{reg})} = w_{\text{reg}}$. Moreover, in this limit, the fractions of regular and chaotic bands correspond to the relative phase-space volumes f_{reg} and $1 - f_{\text{reg}}$, respectively. We therefore obtain from Eq. (7) the mean slope of the chaotic bands, $\bar{w}^{(\text{ch})} = f_{\text{reg}}/(1 - f_{\text{reg}}) = v_{\text{ch}}$, as the classical drift velocity. This is confirmed in Fig. 3a.

The asymptotic quantum transport velocity for a given initial wave packet $|\psi\rangle$ is an average of band slopes weighted with the overlaps $|\langle\psi|\phi_{\alpha,k}\rangle|^2$. We can now explain our observations in Fig. 2: For $\hbar \ll A_{\text{reg}}$, an initial wave packet prepared in the chaotic region of a single unit cell of the extended system is a superposition of chaotic eigenfunctions from the entire band spectrum. Consequently, its drift velocity is given by the mean slope of the chaotic bands and thus by the classical value v_{ch} . In contrast, for $\hbar \gg A_{\text{reg}}$, there are no states restricted to the regular or the chaotic set and hence quantum transport does not correspond to classical transport in this regime.

Our analysis based on the winding numbers can be applied to predict the mean quantum transport in the semiclassical regime from the classical value. The band spectra, however, encode more detailed information about quantum transport. It can be extracted by a double Fourier transform $\epsilon \rightarrow t$, $k \rightarrow x$ (discrete position in units of the spatial period) and subsequent squaring of the spectral density, translating two-point correlations in the bands into the entire time evolution of the spatial distribution on the scale of the spatial period. A formal definition of the result-

ing generalized form factor $K(x, t)$ and further details are found in [16]. A semiclassical theory for the form factor in Hamiltonian ratchets, which will be published elsewhere, requires one to account for the simultaneous presence of regular and chaotic regions in phase space. It relates the form factor $K(x, t)$ to the respective contributions of these invariant sets to the classical spatiotemporal distribution $P(x, t)$ (Fig. 4).

We benefited from discussions with S. Flach, P. Hänggi, M. Holthaus, and O. Yevtushenko. M.F.O. acknowledges financial support from the Volkswagen foundation. T.D. is grateful for the hospitality enjoyed during stays at the MPI für Physik komplexer Systeme, Dresden, and the MPI für Strömungsforschung, Göttingen, generously financed by the MPG.

-
- [1] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics* (Addison-Wesley, Reading, MA, 1966), Vol. 1, Chap. 46.
 - [2] F. Jülicher, A. Ajdari, and J. Prost, *Rev. Mod. Phys.* **69**, 1269 (1997).
 - [3] R. Bartussek, P. Hänggi, and J. G. Kissner, *Europhys. Lett.* **28**, 459 (1994); P. Jung, J. G. Kissner, and P. Hänggi, *Phys. Rev. Lett.* **76**, 3436 (1996).
 - [4] J. L. Mateos, *Phys. Rev. Lett.* **84**, 258 (2000).
 - [5] S. Flach, O. Yevtushenko, and Y. Zolotaryuk, *Phys. Rev. Lett.* **84**, 2358 (2000).
 - [6] I. Goychuk and P. Hänggi, in *Lecture Notes on Physics*, edited by J. Freund and T. Pöschel (Springer, Berlin, 2000), Vol. 557, pp. 7–20.
 - [7] P. Reimann, M. Grifoni, and P. Hänggi, *Phys. Rev. Lett.* **79**, 10 (1997); P. Reimann and P. Hänggi, *Chaos* **8**, 629 (1998).
 - [8] $H = p^2/2 + V(x, t)$ with $\partial_x V(x, t) = (2\pi/\omega^2) \times [\cos(2\pi x) + 0.6 \cos(4\pi x + 0.4) - 2.3 \sin(2\pi t) - 1.38 \sin(4\pi t + 0.7)]$ and $\omega = 2.4$ corresponding to the parameter set (3) of Fig. 1 in Ref. [5] where we have scaled the spatial and temporal period to 1.
 - [9] S. Denisov and S. Flach, nlin.CD/0104006.
 - [10] T. Geisel, in *Lévy Flights and Related Phenomena in Physics*, edited by G.M. Zaslavsky (Springer, Berlin, 1995).
 - [11] T. Dittrich, R. Ketzmerick, M. F. Otto, and H. Schanz, *Ann. Phys. (Leipzig)* **9**, 755 (2000).
 - [12] O. Yevtushenko, S. Flach, and K. Richter, *Phys. Rev. E* **61**, 7215 (2000).
 - [13] H. Sambe, *Phys. Rev. A* **7**, 2203 (1973).
 - [14] A. R. Kolovsky, S. Miyazaki, and R. Graham, *Phys. Rev. E* **49**, 70 (1994); S. Miyazaki and A. R. Kolovsky, *Phys. Rev. E* **50**, 910 (1994).
 - [15] In this work we neglect the existence of hierarchical states, but see R. Ketzmerick, L. Hufnagel, F. Steinbach, and M. Weiss, *Phys. Rev. Lett.* **85**, 1214 (2000).
 - [16] T. Dittrich, B. Mehlig, H. Schanz, and U. Smilansky, *Phys. Rev. E* **57**, 359 (1998).

Peculiar Scaling of Self-Avoiding Walk Contacts

Marco Baiesi,^{1,*} Enzo Orlandini,^{1,†} and Attilio L. Stella^{1,2,‡}

¹*INFN-Dipartimento di Fisica, Università di Padova, I-35131 Padova, Italy*

²*Sezione INFN, Università di Padova, I-35131 Padova, Italy*

(Received 13 March 2001; published 27 July 2001)

The nearest neighbor contacts between the two halves of an N -site lattice self-avoiding walk offer an unusual example of scaling random geometry: for $N \rightarrow \infty$ they are strictly finite in number but their radius of gyration R_c is power law distributed $\propto R_c^{-\tau}$, where $\tau > 1$ is a novel exponent characterizing universal behavior. A continuum of diverging length scales is associated with the R_c distribution. A possibly superuniversal $\tau = 2$ is also expected for the contacts of a self-avoiding or random walk with a confining wall.

DOI: 10.1103/PhysRevLett.87.070602

PACS numbers: 05.70.Jk, 36.20.Ey, 64.60.Ak, 64.60.Kw

The self-avoiding walk (SAW) is a classical problem in statistical mechanics, playing a central role in our understanding of polymer statistics and intimately related to magnetic critical phenomena and percolation [1]. In its most simple version the SAW amounts to the statistical characterization of equally probable single chain conformations, with no overlaps or intersections. These conformations are made by $N - 1$ successive nearest neighbor (nn) steps (N sites) on a lattice in d dimensions. For $N \rightarrow \infty$, quantities like the radius of gyration with respect to the center of mass of SAW configurations, R_g , have an average $\langle R_g \rangle \sim N^{\nu_{\text{SAW}}}$, where ν_{SAW} is the SAW metric exponent. When $d > 1$ self-avoidance does not prevent the conformations from involving close approaches of different parts of the chain: here we generally count as contacts the pairs of nn lattice sites which are visited, not consecutively, by the SAW. The totality of such contacts grows on average proportional to N [2] and possesses the same fractal dimension ($\equiv 1/\nu_{\text{SAW}}$) as the whole SAW.

In the present Letter we discuss so far unexplored features of a particular subset of SAW contacts. Such features represent an unusual example of how scale invariance can manifest itself in a finite, nonextensive portion of an infinite fractal set. The scaling exponent τ of the probability distribution of the gyration radius of such a subset represents a novel characterization of SAW universal behavior. The fact that $\tau > 1$ implies the existence of a whole continuum of diverging characteristic lengths in the SAW problem, in addition to the length $\langle R_g \rangle$.

The subset on which we focus here is that of the contacts between the two halves of a SAW (Fig. 1). Counting such contacts alone allows one to get rid of less interesting effects, which are extensive in N . Being related to problems like network formation, transport, or intramolecular reactions, these contacts can be of particular interest for applications in which the two half chains are made of different monomers, as for diblock copolymers [3]. When the chain represents a homopolymer, the subset we consider is particularly significant in relation to the effect of nearest neighbor interactions on the SAW [4]. This is the case of models of the polymer Θ collapse [1,5,6], where

an attractive nn energy $\epsilon < 0$ is associated to each contact. In this case a Boltzmann factor $e^{-\epsilon/T}$ weighs each contact occurring in a configuration at temperature T . If for such a model one counts contacts only between the two halves of the SAW, the average number of them is of the order N^0 in the high T regime, increases as N^ϕ , with $0 < \phi < 1$, at the Θ point, and scales as N at low T . ϕ turns out to coincide with the crossover exponent ϕ_Θ at $T = T_\Theta$, where $\langle R_g \rangle \sim N^{\nu_\Theta}$, with $\nu_\Theta \neq \nu_{\text{SAW}}$ [7]. Similar behaviors occur if the attractive interactions are associated exclusively with this subset of contacts, and the model describes a diblock copolymer zipping transition [7,8]. In different T regimes of the Θ and zipping transitions the contacts between the two halves of the SAW behave in a way analogous to the monomers adhering to a wall in polymer adsorption [7]. Most recently it has also been found that in $d = 2$ the fractal dimensions of the contacts between the two half chains are the same at the homopolymer Θ point and at the diblock copolymer zipping transition [8]. Thus, intriguing universality aspects can be expected to underlie the geometry of SAW contacts.

The contacts between the two halves of a polymer have already been studied by renormalization group methods [9] in the high T regime controlled by excluded volume. The focus there was on the scaling of their average number, $\langle N_c \rangle$, and precisely on the scaling correction exponent describing its approach to a finite limit for $N \rightarrow \infty$. This finite limit shows that only a vanishing fraction of the

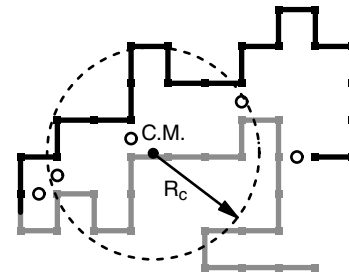


FIG. 1. SAW in two dimensions: the contacts between its two halves (light and dark, respectively) are indicated by open circles. c.m. is the center of mass of these contacts and R_c is their radius of gyration.

total number of contacts pertains to approaches of remote portions of the chain. In fact, $\langle N_c \rangle$ gives only limited information on the contact statistics. The full probability distribution function (PDF) for N_c , $P(N_c, N)$, could in principle have higher moments diverging with $N \rightarrow \infty$, and we devote a first effort to check this possibility, which is normally not considered in polymer statistics. We perform several Monte Carlo simulations to investigate this PDF and its moments in various dimensions and for $80 \leq N \leq 2000$. Since the configurations with many contacts are very rare, we employ a pruned-enriched Rosenbluth method [6] tuned to increase the sampling of configurations with a high number of contacts. This enables us to obtain good statistics also in the tail of the PDF, which for all d turns out to be a negative exponential without substantial dependence on N , as displayed in Fig. 2 for SAW on a cubic lattice.

These results show that this set of contacts is strictly finite in the $N \rightarrow \infty$ limit. In spite of this, one can still ask whether these contacts have interesting geometrical scaling properties, not just reducing to those of a spatially bounded random set. If we indicate by $P_{\text{rad}}(R_c, N)$ the cumulative PDF of R_c (Fig. 1) over all SAW configurations, strict boundedness would mean that the moments $\langle R_c^q \rangle \sim \int dR_c R_c^q P_{\text{rad}}(R_c, N)$ remain finite, for $N \rightarrow \infty$, $\forall q$. We extrapolate the moments in the form $\langle R_c^q \rangle \sim N^{\sigma_q}$. The data are generated by sampling SAW of fixed length with a Monte Carlo algorithm based on pivot moves [10], which have been proved to be very efficient for SAW's [11]. Since the contacts between the two halves of the SAW are mainly located close to the junction point, local moves [12] are also attempted in this region. We consider hypercubic lattices and the fcc lattice in $d = 3$. Contrary to what happens for $P(N_c, N)$, for all d we find that σ_q is positive and grows linearly with q , at sufficiently high q . The data are consistent with a behavior

$$\sigma_q = \begin{cases} 0 & \text{if } q \leq \tau - 1, \\ \nu[q - (\tau - 1)] & \text{if } q > \tau - 1, \end{cases} \quad (1)$$

where $\sigma_q = 0$ may represent logarithmic divergences for $0 < q \leq \tau - 1$. In $d = 2$, for example (see Fig. 3),

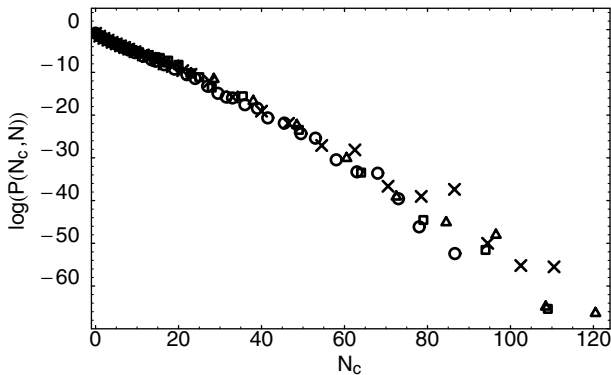


FIG. 2. Histograms of $\ln[P(N_c, N)]$ for SAW on a cubic lattice, with $N = 200$ (\circ), 400 (\times), 800 (\triangle), and 1600 (\square).

we find $\tau = 1.93(2)$ and $\nu = 0.76(1)$, consistent with $\nu = \nu_{\text{SAW}} = 3/4$ [13]. The form (1) is compatible with a PDF having a scaling form $P_{\text{rad}}(R_c, N) \approx R_c^{-\tau} f(R_c/N^\nu)$. That $\nu = \nu_{\text{SAW}}$ is quite plausible since $N^{\nu_{\text{SAW}}}$ is the only expected characteristic length in the problem. However, as discussed below, P_{rad} itself introduces a multiplicity of new length scales for the SAW. The fact that the moments do not approach zero for $q < \tau - 1$ (i.e., σ_q does not become negative) should be due to the circumstance that $f(x)$ is not converging to zero for $x \rightarrow 0$. In other terms, the scaling function g , if we write $P_{\text{rad}} \approx N^{-\tau\nu} g(R_c/N^\nu)$, is singular for its argument approaching zero, $g(x) \sim x^{-\tau}$. We verify these assumptions on the structure of P_{rad} by scaling collapse plots for various d . For example, Fig. 4 shows the data collapse of $\ln[f(x)]$ for $d = 3$. A similar collapse plot for $\ln[g(x)]$ is reported in the inset.

It is indeed the singular character of $g(x)$ for $x \rightarrow 0$, which allows the exponent τ to take a nontrivial value > 1 , while maintaining the zeroth, normalization moment of the PDF equal to 1. The lower length cutoff l (lattice spacing in this case) is crucial to obtain a finite integral, in $dx = d(R_c/N^\nu)$, of the PDF in the continuum limit, because the main contribution comes from small values of x , close to $x_- \equiv l/N^\nu$. This contribution has an N dependence which compensates the diverging factor $N^{-\nu(\tau-1)}$ extracted in front of the integral. $\tau > 1$ means that we cannot associate with the R_c PDF a unique characteristic length. Indeed, putting $\xi_q \equiv \langle R_c^q \rangle^{1/q}$ we find $\xi_q \sim N^{\nu_q}$ with $\nu_q \equiv \sigma_q/q$, for $q \in [\tau - 1, +\infty)$. This means that the self-similarity of contacts has an intrinsic multiscaling character.

The τ found here is a novel exponent for the SAW [14]. It is a measure of the spread of the region within which one-half of the chain feels the presence of the other one in the SAW configurations. A higher τ (see Table I) indicates more localized contacts. Like the global geometry of the SAW, the spread of contacts is determined by the interplay between entropic and excluded volume effects. The nonmonotonic behavior of τ is remarkable, which takes minimum values in $d = 3$ and $d = 4$, indicating these dimensionalities as the optimal ones for a broad

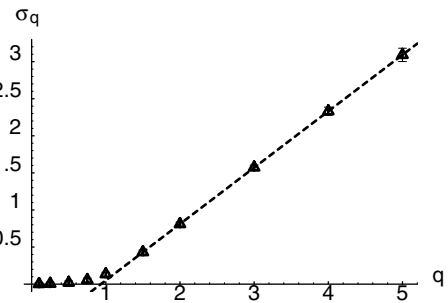


FIG. 3. σ_q vs q and extrapolation of $\tau - 1$ for $d = 2$. We expect $\sigma_q = 0$ for $q \leq \tau - 1$. Numerically a logarithmic divergence cannot be easily distinguished from a power law one with $\sigma_q \approx 0$.

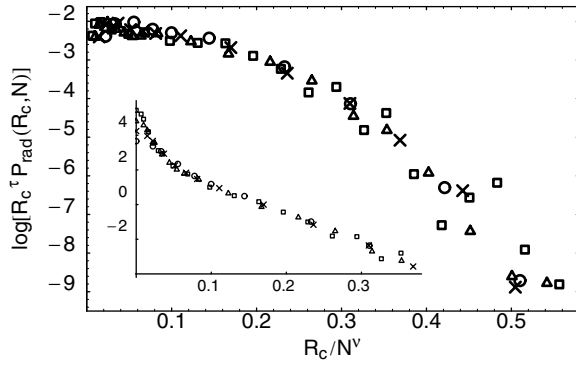


FIG. 4. Collapses of $\ln[f(x)]$ (see text) and $\ln[g(x)]$ (inset) for SAW on cubic lattice, with $N = 200$ (\circ), 400 (\times), 800 (\triangle), and 1600 (\square).

interpenetration of the two half chains. At high d there is soon much room for the two branches of the SAW to develop without giving rise to contacts, and this tends to localize them more. τ is rather high in $d = 2$: we interpret this as a consequence of the peculiar topology of the two-dimensional lattice, which makes it more difficult than, e.g., in $d = 3$ for the two half chains to approach each other at large length scales. The dependence of τ on d , also above the upper critical dimension $d_u = 4$, is a further indication of the peculiar novel character of this exponent.

In $d = 2$ and $d = 3$ we also investigate the behavior of the contacts between the two half chains in the presence of nn attractive interactions (Θ point model). While for $T > T_\Theta$ the behaviors of their PDF's appear the same as at $T = \infty$, at the Θ point the moments of P diverge as those of P_{rad} , while τ becomes equal to 1 and $\sigma_q/q = \nu_\Theta$ for the latter PDF. This suggests that the disappearance of the singular scaling function in P_{rad} could be a good criterion for locating the transition point. This is illustrated in Fig. 5, referring to the Θ point in $d = 3$, simulated as in Ref. [6] with chains up to $N = 2000$. We collect data from two runs at $-\epsilon/T = 0.25, 0.27$, and we use the multiple histogram method [15] to calculate the moments in the surrounding interval of temperatures. The result from the values of q examined is $-\epsilon/T_\Theta = 0.274(4)$, consistent with accurate estimates by other methods: for example, $0.275(8)$ in [5] and $0.2690(3)$ in [6]. Similar results are valid for models of the diblock copolymer zipping transition [16].

TABLE I. Extrapolated values of τ and ν for various lattices, using SAW with length from N_{\min} to N_{\max} .

Lattice	τ	ν	ν_{SAW}	N_{\min}	N_{\max}
$2d$	1.93(2)	0.76(1)	$3/4$ [13]	1000	10 000
$3d$	1.51(2)	0.595(5)	0.5877(6) [11]	1000	10 000
fcc	1.52(2)	0.594(5)	0.5877(6)	3000	6000
$4d$	1.51(3)	0.52(2)	$1/2$	1500	8000
$5d$	2.0(2)	0.49(2)	$1/2$	1000	8000
$6d$	2.9(1)	0.49(1)	$1/2$	1000	5000

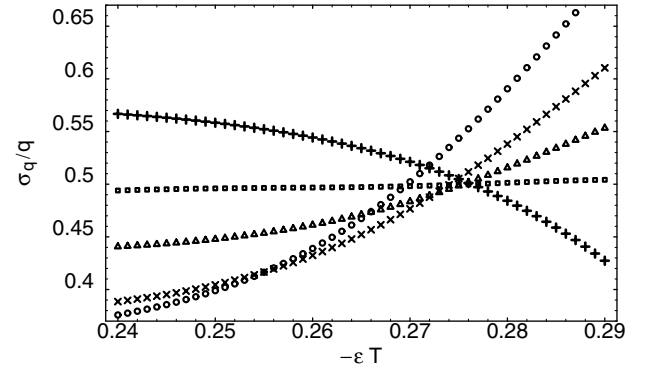


FIG. 5. Extrapolation of σ_q/q close to T_Θ in $d = 3$, for $q = 0.5$ (\circ), $q = 1$ (\times), $q = 2$ (\triangle), and $q = 4$ (\square). The crossings are consistent with the expectation to find $\tau = 1$ and $\sigma_q/q = \nu_\Theta = 1/2$ right at $T = T_\Theta$. The curve with the opposite trend (+) refers to the effective ν exponent of the SAW radius of gyration.

For SAW with attractive interactions the average number of contacts between the two half chains behaves in the various temperature ranges as the mean number of SAW-wall contacts in a polymer adsorption transition. Thus, it makes sense to check whether also in the adsorption process the high T ordinary regime is characterized by peculiar scaling features of the kind described above. To this purpose we study the PDF's of the number and the radius of SAW-wall contacts at $T = \infty$, where the wall exerts only a geometrical confinement effect on the chain. Here we consider as radius the mean distance R_c^* of a contact from the point where the SAW is grafted to the wall [17]. Also in this case all the moments of the PDF of N_c appear to remain finite for $N \rightarrow \infty$. The PDF's of R_c show singular scaling functions with an exponent $\tau \simeq 2$, in $d = 2$ and $d = 3$ (Table II).

We study also the case of a random walk (RW) confined by a wall, for which we are able to compute exactly τ . Consider first a RW on a square lattice tilted at 45° with respect to the coordinate axes, in such a way that the nn of a site have coordinates $\{(1, 1), (-1, 1), (-1, -1), (1, -1)\}$. In this way a nn step moves the RW with nonzero and independent displacements along both coordinate directions. This simplifies the calculations, but we expect the final results to be valid for every lattice model, because independence is asymptotically recovered for long RW. The generalization to d dimensions gives 2^d nn vectors of the form $(1, 1, \dots, 1), (-1, 1, \dots, 1), \dots, (-1, -1, \dots, -1)$ and, for example, in $d = 3$ one obtains the bcc lattice. Let the walk start from the origin, near a $d - 1$ -dimensional hard wall perpendicular to the x coordinate: the problem in the x direction is equivalent to a one-dimensional RW which steps

TABLE II. As in Table I, but for SAW-wall contacts.

Lattice	τ	ν	N_{\min}	N_{\max}
$2d$	1.99(3)	0.744(5)	1000	6000
$3d$	1.968(34)	0.58(1)	1000	8000

to nn sites with equal probability. The probability to be again at $x = 0$ after n steps is given by $P_n(0) = 2^{-n} \binom{n}{n/2}$. The effect of the impenetrable wall is represented by forbidden sites located at $\bar{x} = -1$. So one has to subtract the probability to travel through this point. The method of images [18] in this simple case says that this is equal to the probability to go from $x = 0$ to its mirror image with respect to \bar{x} , $x = -2$. So $\bar{P}_n(0) = P_n(0) - P_n(-2) = P_n(0)2(n+2)^{-1}$ is the probability to be on the surface. The q th moment of the average distance of a contact from the origin is thus

$$\langle R_c^*(N)^q \rangle \sim \frac{\sum_{n=2,4,6,\dots}^N \bar{P}_n(0) (n^{1/2})^q}{\sum_{n=2,4,6,\dots}^N \bar{P}_n(0)}, \quad (2)$$

where $n^{1/2}$ is the root mean square displacement after n steps. Using $n! \approx \sqrt{2\pi} n^{n+1/2} e^{-n}$ for the binomials, one recovers that the denominator is finite for $N \rightarrow \infty$, while the numerator is equivalent to a sum of the type $\sum_{n=2}^N n^{(q-3)/2}$, diverging as $N^{(q-1)/2}$ if $q > 1$. This means $\tau = 2$ for RW near a wall, in any d . The exact result for RW suggests that, in any dimension, for SAW confined by a wall, the scaling of the contacts is the same as for RW. Thus, excluded volume effects seem to play no role in determining τ for the contacts of the SAW with a confining wall.

In summary, the contacts discussed above represent a very peculiar example of scaling random set: in fractal physics we are familiar with sets in which the number of elements is growing to infinity together with their average radius of gyration, and criticality implies a nontrivial scaling for the PDF's of both quantities. A typical example is percolation clusters [19], whose size PDF has a singular scaling function with nontrivial τ , as we instead find here for P_{rad} . In the case considered here the scale invariance of the set is not accompanied by its number of elements being broadly, power law distributed. The criticality is in fact triggered by the length of the whole chain, N , becoming infinite, while N_c remains finite. A τ exponent, possibly superuniversal, can also be defined for the contacts between a SAW and a $d-1$ -dimensional confining wall in the ordinary regime. The case of a RW in the presence of a wall can also be treated, yielding the exact classical value $\tau = 2$, which could apply also to SAW, independent

of d . Exact determinations of these exponents for SAW in low ($d = 2$) or high dimensionality and the possible descriptions of the new scalings within the renormalization group framework remain a challenge for the future.

Partial support from the European Network No. ERBFMRXCT980183 and from MURST through COFIN-99 is acknowledged. We thank U. Bastolla and E. Carlon for help and useful discussions.

*E-mail address: baiesi@pd.infn.it

†E-mail address: orlandini@pd.infn.it

‡E-mail address: stella@pd.infn.it

- [1] C. Vanderzande, *Lattice Models of Polymers* (Cambridge University Press, Cambridge, 1998).
- [2] This can be easily proved on the basis of the Kesten pattern theorem: H. Kesten, *J. Math. Phys. (N.Y.)* **4**, 960 (1963).
- [3] S. F. Bates and G. H. Fredrickson, *Phys. Today* **52**, No. 2, 32 (1999).
- [4] J. F. Douglas and T. Ishinabe, *Phys. Rev. E* **51**, 1791 (1995).
- [5] M. C. Tesi and E. J. Janse van Rensburg, E. Orlandini, and S. G. Whittington, *J. Stat. Phys.* **29**, 2451 (1996).
- [6] P. Grassberger, *Phys. Rev. E* **56**, 3682 (1997).
- [7] E. Orlandini, F. Seno, and A. L. Stella, *Phys. Rev. Lett.* **84**, 294 (2000).
- [8] M. Baiesi, E. Carlon, E. Orlandini, and A. L. Stella, *Phys. Rev. E* **63**, 041801 (2001).
- [9] S. Müller and L. Schäfer, *Eur. Phys. J. B* **2**, 351 (1998).
- [10] N. Madras and A. Sokal, *J. Stat. Phys.* **50**, 109 (1988).
- [11] B. Li, N. Madras, and A. D. Sokal, *J. Stat. Phys.* **80**, 661 (1995).
- [12] P. H. Verdier and W. H. Stockmayer, *J. Chem. Phys.* **36**, 227 (1961).
- [13] B. Nienhuis, *Phys. Rev. Lett.* **49**, 1063 (1982).
- [14] One can also consider asymmetric chain partitions. If the ratio of block lengths is kept fixed for $N \rightarrow \infty$, τ does not appear to change with respect to the symmetric case.
- [15] A. M. Ferrenberg and R. H. Swendsen, *Phys. Rev. Lett.* **61**, 2635 (1988); **63**, 1195 (1989).
- [16] M. Baiesi, E. Orlandini, and A. L. Stella (to be published).
- [17] This gives results equivalent to those for R_c .
- [18] M. N. Barber and B. N. Ninham, *Random and Restricted Walks* (Gordon and Breach, New York, 1970).
- [19] D. Stauffer, *Introduction to Percolation theory* (Taylor and Francis, London and Philadelphia, 1985).

Transition to Coherence in Populations of Coupled Chaotic Oscillators: A Linear Response Approach

Dmitri Topaj, Won-Ho Kye, and Arkady Pikovsky

Department of Physics, University of Potsdam, Postfach 601553, D-14415 Potsdam, Germany

(Received 21 March 2001; published 26 July 2001)

We consider the collective dynamics in an ensemble of globally coupled chaotic maps. The transition to the coherent state with a macroscopic mean field is analyzed in the framework of the linear response theory. The linear response function for the chaotic system is obtained using the perturbation approach to the Frobenius-Perron operator. The transition point is defined from this function by virtue of the self-excitation condition for the feedback loop. Analytical results for the coupled Bernoulli maps are confirmed by the numerics.

DOI: 10.1103/PhysRevLett.87.074101

PACS numbers: 05.45.Xt, 05.70.Fh

Ensembles of globally coupled nonlinear oscillators attracted much attention recently. Such models arise in the study of Josephson junctions [1], multimode lasers [2], and charge density waves [3]. In the living world one uses similar models to describe chirps of grasshoppers [4], neurons [5,6], and yeast cells [7]. A particular interest attracted ensembles of chaotic oscillators [8–12]. Recently, an experimental investigation of 64 globally coupled chaotic electrochemical oscillators have been performed [13]. These studies have revealed that already coupling of identical chaotic oscillators demonstrates nontrivial synchronization patterns.

The first step in the emerging of order and coherence in an ensemble of globally coupled elements is the appearance of a macroscopic mean field. Indeed, the global coupling can be represented as the coupling through the mean field, and the absence of the latter means the absence of interaction and full independence of the elements. In the thermodynamic limit, where the number of elements tends to infinity, one can interpret the appearance of the mean field as a nonequilibrium phase transition. Such a transition is well described for ensembles of noise-driven oscillators [14,15], by virtue of the bifurcation analysis of the nonlinear Fokker-Planck equation. In the case of chaotic deterministic systems one has to consider the nonlinear Frobenius-Perron or Liouville equations (in the cases of discrete and continuous dynamics, correspondingly), which is by far a more difficult task.

In this Letter, we develop an approach to the description of the transition in the ensembles of chaotic elements, based on the response theory for chaos. The main idea is to “break” the feedback loop due to coupling in the ensemble, and to consider the effect of a small periodic force on a chaotic oscillator. With the help of the linear response theory (cf. [16,17]) we can find the linear response of the distribution function of chaos to this force. Then, for each type of coupling the response function can be calculated. After that we can again “close” the feedback loop, reducing the problem of the onset of the mean field to the analysis of stability of a linear discrete dynamical system with a

given response characteristic. Note that a similar approach for noise-driven periodic oscillators has been developed in [18].

Our basic model is the system of N globally coupled chaotic maps

$$\begin{aligned} x_i(t+1) &= f(x_i(t)) + \varepsilon g(x_i(t))a(t), \\ a(t) &= \frac{1}{N} \sum_{i=1}^N q(x_i(t)). \end{aligned} \quad (1)$$

Here ε is the coupling constant. Note that the coupling performs via the mean field a . We write the coupling in a general form, using arbitrary functions $g(x)$ and $q(x)$. The only natural condition is that in the thermodynamic limit the mean field vanishes for $\varepsilon = 0$, i.e., $\langle q(x) \rangle_0 = 0$, where $\langle \rangle_0$ denotes the average over the stationary distribution of the map $x \mapsto f(x)$. This condition ensures that the disordered state with vanishing mean field a exists for all couplings ε (as will be shown below, the instability of this state leads to the transition to ordered state with nonzero mean field a).

In previous investigations [19–22] it has been demonstrated numerically that in system (1) a transition from vanishing to a macroscopic mean field a can occur at some critical coupling strength ε_c . Generally, this transition can be interpreted as a bifurcation in the self-consistent nonlinear Frobenius-Perron equation for the probability density [20,21]. However, except for a special case of noisy homographic maps [22], no analytical approach to the description of the transition has been developed. Below, we develop such an approach, basing it on the linear response theory of chaos.

To apply the linear response theory we consider a supplementary problem of the effect of small periodic force (with frequency ω) on a single map:

$$x(t+1) = f(x(t)) + \alpha g(x(t))e^{i\omega t}. \quad (2)$$

The forcing affects the dynamics, and our goal is to find the variations of the probability distribution density in the first order in $\alpha \ll 1$. Denoting the right-hand side of (2)

as $F_t(x)$, we write the Frobenius-Perron operator for the density $\rho_t(x)$

$$\rho_{t+1}(x) = \int \delta[x - F_t(y)] \rho_t(y) dy. \quad (3)$$

Considering the map as defined on the interval $[0, 2\pi)$ (this can always be achieved with a normalization), we can introduce the Fourier transform of the density $\rho_t(x) = \sum \psi_t(k) e^{ikx}$ and to obtain from (3) the corresponding Frobenius-Perron operator in the Fourier space:

$$\begin{aligned} \psi_{t+1}(k) &= \sum_l R_t(k, l) \psi_t(l), \\ R_t(k, l) &= \frac{1}{2\pi} \int_0^{2\pi} \exp[i l x - i k F_t(x)] dx. \end{aligned} \quad (4)$$

Taking into account that $F_t(x) = f(x) + \alpha g(x) e^{i\omega t}$ we can write up to the first order in α

$$R(k, l) = R^0(k, l) + \alpha R^1(k, l) e^{i\omega t}, \quad (5)$$

where

$$\begin{aligned} R^0(k, l) &= \frac{1}{2\pi} \int_0^{2\pi} \exp[i l x - i k f(x)] dx, \\ R^1(k, l) &= \frac{-ik}{2\pi} \int_0^{2\pi} g(x) \exp[i l x - i k f(x)] dx. \end{aligned}$$

Substituting this in (4) and writing $\psi_t(k) = \psi^0(k) + \alpha e^{i\omega t} \psi^1(k)$ we obtain the equation for the complex amplitude of the perturbation ψ^1

$$e^{i\omega} \psi^1 = \sum_l [R^0(k, l) \psi^1(l) + R^1(k, l) \psi^0(l)]. \quad (6)$$

One can formally write a solution of this linear system, but from this formal expression it is difficult to judge whether the solution is nonsingular. Indeed, as is well known in the theory of chaos (see, e.g., [16,23]), the response to small perturbations can be singular. This happens, e.g., in structurally unstable chaotic systems. In such systems, small changes of a parameter lead to a topologically nonequivalent dynamics, which can, e.g., be seen in the symbolic description or in the representation via unstable periodic orbits. Note that to this class belong even many systems where chaos persists in the whole parameter range (e.g., the Lorenz attractor and the tent map), let alone such non-hyperbolic examples where small perturbation can lead to a periodic window (like in the logistic map). Response of the structurally unstable system is expected to be singular [23], which, in particular, can be seen from the fractal dependence of some statistical characteristics on a parameter [24]. Therefore, in order to be in the realm of validity of the linear response theory, we have to consider a structurally stable map $x \mapsto f(x)$.

Below, we study the simplest such map—the Bernoulli map $f(x) = 2x \bmod 2\pi$. In this case $R^0(k, l) = \delta(2k - l)$ and $\psi^0(k) = \delta(k)(2\pi)^{-1}$; thus Eq. (6) reduces to

$$\psi^1(k) e^{i\omega} = \psi^1(2k) - \frac{ik}{2\pi} G(2k), \quad (7)$$

where $G(l)$ is the Fourier harmonics of $g(x)$,

$$G(l) = \frac{1}{2\pi} \int_0^{2\pi} e^{-ilx} g(x) dx. \quad (8)$$

The solution of (7) gives the final expression for the response of the probability density:

$$\psi^1(k) = \frac{-ik}{4\pi} \sum_{m=1}^{\infty} 2^m e^{-im\omega} G(2^m k). \quad (9)$$

The next step is the closing of the feedback loop in model (1). To this end we have to find which mean field is generated in the first order in α in system (2). In the thermodynamic limit $N \rightarrow \infty$ we can calculate the mean field as the average over the probability density:

$$\lim_{N \rightarrow \infty} \sum_{i=1}^N \frac{1}{N} q(x_i(t)) = \langle q(x(t)) \rangle = \int_0^{2\pi} \rho_t(x) q(x) dx.$$

Substituting here the expression for the density

$$\rho_t(x) = \sum_k [\psi^0(k) + \alpha e^{i\omega t} \psi^1(k)] e^{ikx},$$

we obtain

$$\langle q \rangle = \alpha K(\omega) e^{i\omega t}, \quad K(\omega) = \sum_k Q(-k) \psi^1(k), \quad (10)$$

where $Q(k)$ are the Fourier harmonics of the function $q(x)$ defined similar to (8). Here we have taken into account that the unperturbed invariant density does not contribute to the mean field. In the particular case of the Bernoulli map we obtain from (9) and (10)

$$K(\omega) = \sum_{k=-\infty}^{\infty} \sum_{m=1}^{\infty} \frac{-ik}{4\pi} 2^m e^{-im\omega} G(2^m k) Q(-k). \quad (11)$$

The function $K(\omega)$ is the linear response function of system (2) for the observable $q(x)$. In terms of Eq. (1) it describes the linear dynamics of the mean field a in the spectral representation. The condition of the generation of a in the original ensemble (1), i.e., the condition of instability of the state $a = 0$, can be formulated as the condition for nondecay of perturbations within the feedback loop (the self-consistency condition). The total amplification is the product of the factors ε and $K(\omega)$. Thus the instability threshold is defined by the relation

$$\varepsilon K(\omega) = 1. \quad (12)$$

This relation determines the frequency of oscillations at the instability threshold [from the imaginary part of (12)] and the critical value of coupling ε_c [from the real part of (12)].

As a particular example we consider the following coupling in the Bernoulli map:

$$g(x) = \sin 2x + \sin 4x, \quad q(x) = \cos x. \quad (13)$$

Substituting this in (11) yields

$$K(\omega) = -\frac{1}{4\pi} (2e^{-i2\omega} + e^{-i\omega}), \quad (14)$$

and from (12) we obtain the following critical values for the coupling:

$$\begin{aligned} \varepsilon_{c1} &= 2\pi, & \omega_1 &= \arccos(-\tfrac{1}{4}); \\ \varepsilon_{c2} &= -\tfrac{4\pi}{3}, & \omega_2 &= 0. \end{aligned} \quad (15)$$

These results of the theory are confirmed in numerical simulations, presented in Figs. 1 and 2.

Essential features of the transitions can be deduced already from the linear analysis above. The transition to coherence at ε_{c1} has nonzero frequency, so it should be

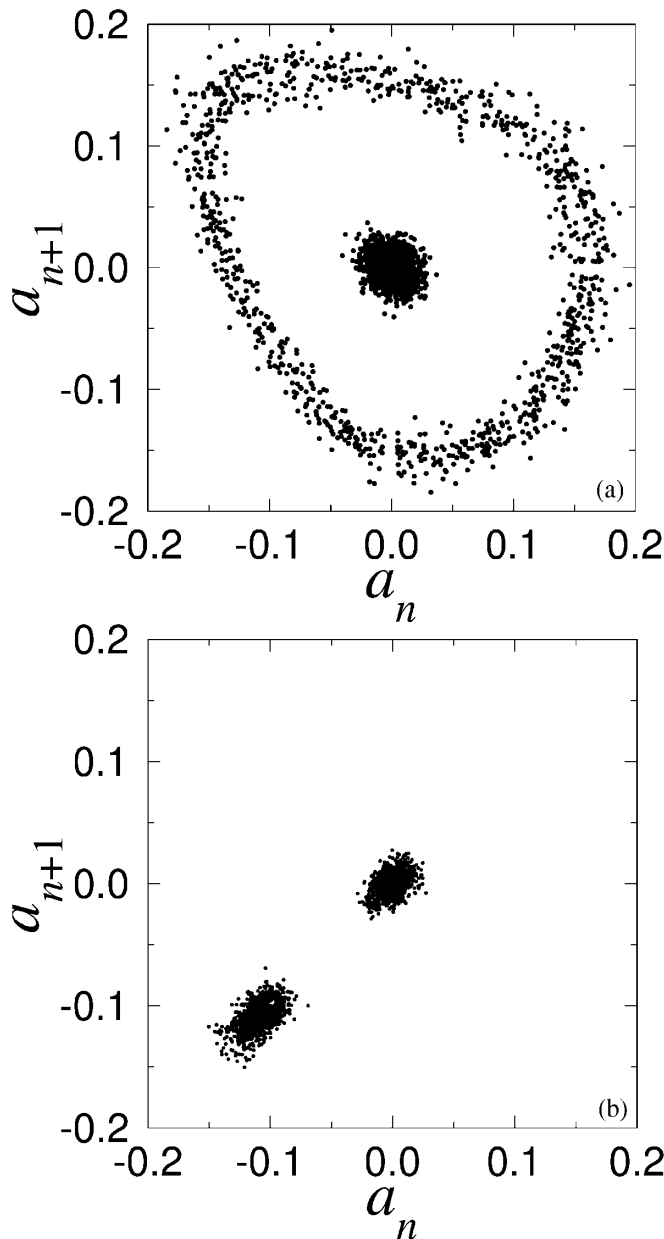


FIG. 1. Dynamics of the mean field in the ensemble of $N = 10^4$ coupled maps represented as a recurrence plot a_{n+1} vs a_n . (a) The transition at ε_{c1} . In the disordered state at $\varepsilon = 5$ the mean field vanishes up to finite-size fluctuations (a cloud around $a_n = a_{n+1} = 0$); in the coherent state ($\varepsilon = 7$) nearly periodic oscillations are observed. (b) The transition at ε_{c2} is one from a zero equilibrium point ($\varepsilon = -3$) to a nearly constant mean field at $\varepsilon = -5$.

a Neimark (discrete time Hopf) bifurcation. Numerics (Figs. 1a and 2) shows that it is supercritical. The transition at ε_{c2} occurs with the unit multiplier; thus one can expect here a transcritical transition to a nonzero fixed point (Fig. 1b) (a pitchfork bifurcation is excluded due to the absence of symmetry $a \mapsto -a$). Note that the scaling laws expected for these transitions ($\langle a^2 \rangle \sim |\varepsilon - \varepsilon_c|$ for the Neimark one and $\langle a^2 \rangle \sim |\varepsilon - \varepsilon_c|^2$ for the transcritical one) are distorted by finite-size effects.

The linear response approach above can be obviously generalized to the case when the mean field has its own dynamics. Such a situation appears, e.g., in a series array of Josephson junctions coupled by means of an external load [25]. The junctions are coupled via the common current, which obeys an additional equation (for the oscillatory circuit load considered in [25] this is the equation of a driven damped linear oscillator). In the linear response theory this additional dynamics of the mean field can be easily incorporated just by multiplying the response function of the chaotic map $K(\omega)$ with the response function of the mean field dynamics $L(\omega)$, thus yielding the stability condition

$$\varepsilon L(\omega) K(\omega) = 1$$

instead of (12). As a simple example let us consider the ensemble (1) with the inertial dynamics the mean field

$$a(t) = \gamma a(t-1) + \frac{1}{N} \sum_{i=1}^N q(x_i(t)).$$

In this case $L(\omega) = (1 - \gamma e^{i\omega})^{-1}$ with an obvious modification of the transition values (15).

Another straightforward generalization of the theory above is including an additive noise in the dynamics. In this case Eq. (1) is rewritten as

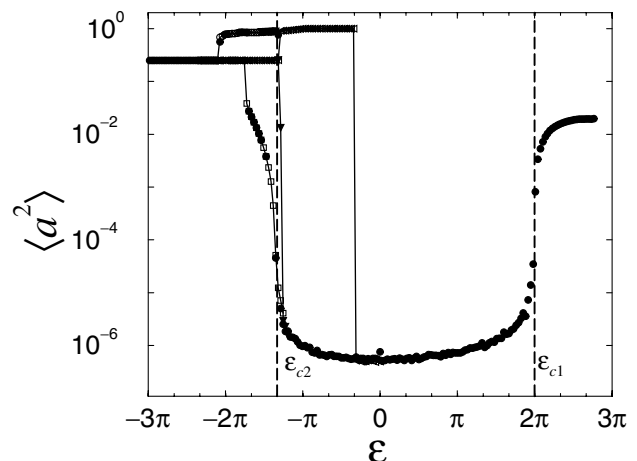


FIG. 2. The dependence of $\langle a^2 \rangle$ in the ensemble of $N = 10^6$ maps on the coupling ε reveals transitions at the critical values predicted by the linear theory (15) (dashed lines). Two large-amplitude branches at $\varepsilon < \varepsilon_{c2}$ correspond to one-cluster solution (largest branch, this solution is a fixed point $x_i = 0$ for small $|\varepsilon|$ and exhibits period doublings for larger $|\varepsilon|$) and to a two-cluster period-two solution (middle branch). Jumps to these branches make the transition at ε_{c2} hysteretic.

$$x_i(t+1) = f(x_i(t)) + \xi_i(t) + \varepsilon g(x_i(t))a(t), \quad (16)$$

$$a(t) = \frac{1}{N} \sum_{i=1}^N q(x_i(t)),$$

where $\xi_i(t)$ are independent equally distributed random variables. The Frobenius-Perron equation (3) is generalized to this case by including the convolution of ρ with the probability density of ξ (see, e.g., [26]). In the Fourier space one simply multiplies the operator R (4),(5) with the characteristic function of noise:

$$\tilde{R}(k, l) = w(k)R(k, l),$$

where \tilde{R}, R are the operators with and without noise, respectively, and $w(k)$ is the Fourier transform of the probability density W_ξ of the random noise:

$$w(k) = \int_{-\infty}^{\infty} W_\xi(x) e^{-ikx} dx.$$

With this modification, all the linear response theory holds. Moreover, the presence of the fast decaying factor $w(k)$ regularizes the Frobenius-Perron equation, so that a non-singular solution for the response function can be expected even when the deterministic dynamics is structurally unstable, or even when in the deterministic dynamics periodic windows are present. In the particular case of Bernoulli maps, the final expression for the response function K (11) is modified to

$$K(\omega) = \sum_{k=-\infty}^{\infty} \frac{-ikQ(-k)}{4\pi} \sum_{m=1}^{\infty} 2^m e^{-im\omega} G(2^m k) \\ \times \prod_{l=0}^{m-1} w(2^l k).$$

Formally, the generalization of the theory above to the case of continuous-time oscillators is also simple. One just writes the Liouville equation instead of the Frobenius-Perron one, or the Fokker-Planck equation if the noise is present. However, analytic solution of the perturbation problem appears to be hardly feasible. Nevertheless, one can still proceed numerically, determining the linear response function from the simulations [17]. Because the fluctuations-dissipation theorem does not hold for generic chaotic systems, one has to rely on direct simulations. Namely, one has to numerically integrate the periodically driven system [in the discrete case to iterate the map (2)] and to calculate the spectral component of the output at the driving frequency. The amplitude and the phase of this component yield the linear response function $K(\omega)$. Then the condition similar to (12) gives the threshold of instability and the frequency of the appearing oscillations of the mean field. Examples of this analysis will be presented elsewhere.

In conclusion, we have developed a theory of the transition to coherent collective behavior in ensembles of globally coupled chaotic maps. The main idea is to define the linear response function according to the mode of coupling. Noteworthy, this function can be obtained from the analysis of a single system. The critical coupling follows then

from the stability condition for the feedback loop of the mean field (12). We have also outlined several straightforward generalizations of the method, e.g., to the noisy systems. Less obvious is a generalization to the case of nonidentical interacting systems; it is the subject of current work.

We thank the Deutsche Forschungsgemeinschaft (SFB 555) for financial support.

-
- [1] P. Hadley, M. R. Beasley, and K. Wiesenfeld, *Phys. Rev. B* **38**, 8712 (1988).
 - [2] K. Wiesenfeld, C. Bracikowski, G. James, and R. Roy, *Phys. Rev. Lett.* **65**, 1749 (1990).
 - [3] S. H. Strogatz, C. M. Marcus, R. M. Westervelt, and R. E. Mirollo, *Physica (Amsterdam)* **36D**, 23 (1989).
 - [4] E. Sismundo, *Science* **249**, 55 (1990).
 - [5] E. A. Stern, D. Jaeger, and C. J. Wilson, *Nature (London)* **394**, 475 (1998).
 - [6] N. F. Rulkov, *Phys. Rev. Lett.* **86**, 183 (2001).
 - [7] S. Dano, P. G. Sorensen, and F. Hynne, *Nature (London)* **402**, 320 (1999).
 - [8] H. Bohr, K. S. Jensen, T. Petersen, B. Rathjen, E. Mosekilde, and N.-H. Holstein-Rathlou, *Parallel Comput.* **12**, 113 (1989).
 - [9] A. Pikovsky, M. Rosenblum, and J. Kurths, *Europhys. Lett.* **34**, 165 (1996).
 - [10] D. H. Zanette and A. S. Mikhailov, *Phys. Rev. E* **57**, 276 (1998).
 - [11] D. H. Zanette and A. S. Mikhailov, *Phys. Rev. E* **58**, 872 (1998).
 - [12] H. Sakaguchi, *Phys. Rev. E* **61**, 7212 (2000).
 - [13] W. Wang, I. Z. Kiss, and J. L. Hudson, *Chaos* **10**, 248 (2000).
 - [14] R. C. Desai and R. Zwanzig, *J. Stat. Phys.* **19**, 1 (1978).
 - [15] L. L. Bonilla, J. C. Neu, and R. Spigler, *J. Stat. Phys.* **67**, 313 (1992).
 - [16] S. Grossmann, *Z. Phys. B* **57**, 77 (1984).
 - [17] C. Reick, in *Computational Physics Nonlinear Dynamical Phenomena in Physical, Chemical and Biological Systems. Proceedings of the 3rd IMACS International Conference on Computational Physics, Lyngby, Denmark, 1994* (IMACS, Rutgers University, Piscataway, NJ, 1994), pp. 209–214.
 - [18] S. H. Strogatz and R. E. Mirollo, *J. Stat. Phys.* **63**, 613 (1991).
 - [19] K. Kaneko, *Physica (Amsterdam)* **55D**, 368 (1992).
 - [20] A. S. Pikovsky and J. Kurths, *Phys. Rev. Lett.* **72**, 1644 (1994).
 - [21] A. S. Pikovsky and J. Kurths, *Physica (Amsterdam)* **76D**, 411 (1994).
 - [22] M. Griniasty and V. Hakim, *Phys. Rev. E* **49**, 2661 (1994).
 - [23] S. V. Ershov, *Phys. Lett. A* **177**, 180 (1993).
 - [24] R. Klages and J. R. Dorfman, *Phys. Rev. Lett.* **74**, 387 (1995).
 - [25] K. Wiesenfeld and J. W. Swift, *Phys. Rev. E* **51**, 1020 (1995).
 - [26] A. Lasota and M. C. Mackay, *Probabilistic Properties of Deterministic Systems* (Cambridge University Press, Cambridge, 1985).

Statistical mechanics of complex networks

Réka Albert* and Albert-László Barabási

Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556

(Published 30 January 2002)

Complex networks describe a wide range of systems in nature and society. Frequently cited examples include the cell, a network of chemicals linked by chemical reactions, and the Internet, a network of routers and computers connected by physical links. While traditionally these systems have been modeled as random graphs, it is increasingly recognized that the topology and evolution of real networks are governed by robust organizing principles. This article reviews the recent advances in the field of complex networks, focusing on the statistical mechanics of network topology and dynamics. After reviewing the empirical data that motivated the recent interest in networks, the authors discuss the main models and analytical tools, covering random graphs, small-world and scale-free networks, the emerging theory of evolving networks, and the interplay between topology and the network's robustness against failures and attacks.

CONTENTS

I. Introduction	48	C. Random graphs with power-law degree distribution	66
II. The Topology of Real Networks: Empirical Results	49	D. Bipartite graphs and the clustering coefficient	66
A. World Wide Web	49	VI. Small-World Networks	67
B. Internet	50	A. The Watts-Strogatz model	67
C. Movie actor collaboration network	52	B. Properties of small-world networks	68
D. Science collaboration graph	52	1. Average path length	68
E. The web of human sexual contacts	52	2. Clustering coefficient	69
F. Cellular networks	52	3. Degree distribution	70
G. Ecological networks	53	4. Spectral properties	70
H. Phone call network	53	VII. Scale-Free Networks	71
I. Citation networks	53	A. The Barabási-Albert model	71
J. Networks in linguistics	53	B. Theoretical approaches	71
K. Power and neural networks	54	C. Limiting cases of the Barabási-Albert model	73
L. Protein folding	54	D. Properties of the Barabási-Albert model	74
III. Random-Graph Theory	54	1. Average path length	74
A. The Erdős-Rényi model	54	2. Node degree correlations	75
B. Subgraphs	55	3. Clustering coefficient	75
C. Graph evolution	56	4. Spectral properties	75
D. Degree distribution	57	VIII. The Theory of Evolving Networks	76
E. Connectedness and diameter	58	A. Preferential attachment $\Pi(k)$	76
F. Clustering coefficient	58	1. Measuring $\Pi(k)$ for real networks	76
G. Graph spectra	59	2. Nonlinear preferential attachment	77
IV. Percolation Theory	59	3. Initial attractiveness	77
A. Quantities of interest in percolation theory	60	B. Growth	78
B. General results	60	1. Empirical results	78
1. The subcritical phase ($p < p_c$)	60	2. Analytical results	78
2. The supercritical phase ($p > p_c$)	61	C. Local events	79
C. Exact solutions: Percolation on a Cayley tree	61	1. Internal edges and rewiring	79
D. Scaling in the critical region	62	2. Internal edges and edge removal	79
E. Cluster structure	62	D. Growth constraints	80
F. Infinite-dimensional percolation	62	1. Aging and cost	80
G. Parallels between random-graph theory and percolation	63	2. Gradual aging	81
V. Generalized Random Graphs	63	E. Competition in evolving networks	81
A. Thresholds in a scale-free random graph	64	1. Fitness model	81
B. Generating function formalism	64	2. Edge inheritance	82
1. Component sizes and phase transitions	65	F. Alternative mechanisms for preferential attachment	82
2. Average path length	65	1. Copying mechanism	82
		2. Edge redirection	82
		3. Walking on a network	83
		4. Attaching to edges	83
		G. Connection to other problems in statistical mechanics	83
		1. The Simon model	83
		2. Bose-Einstein condensation	85

*Present address: School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

IX. Error and Attack Tolerance	86
A. Numerical results	86
1. Random network, random node removal	87
2. Scale-free network, random node removal	87
3. Preferential node removal	87
B. Error tolerance: analytical results	88
C. Attack tolerance: Analytical results	89
D. The robustness of real networks	90
1. Communication networks	90
2. Cellular networks	91
3. Ecological networks	91
X. Outlook	91
A. Dynamical processes on networks	91
B. Directed networks	92
C. Weighted networks, optimization, allometric scaling	92
D. Internet and World Wide Web	93
E. General questions	93
F. Conclusions	94
Acknowledgments	94
References	94

I. INTRODUCTION

Complex weblike structures describe a wide variety of systems of high technological and intellectual importance. For example, the cell is best described as a complex network of chemicals connected by chemical reactions; the Internet is a complex network of routers and computers linked by various physical or wireless links; fads and ideas spread on the social network, whose nodes are human beings and whose edges represent various social relationships; the World Wide Web is an enormous virtual network of Web pages connected by hyperlinks. These systems represent just a few of the many examples that have recently prompted the scientific community to investigate the mechanisms that determine the topology of complex networks. The desire to understand such interwoven systems has encountered significant challenges as well. Physics, a major beneficiary of reductionism, has developed an arsenal of successful tools for predicting the behavior of a system as a whole from the properties of its constituents. We now understand how magnetism emerges from the collective behavior of millions of spins, or how quantum particles lead to such spectacular phenomena as Bose-Einstein condensation or superfluidity. The success of these modeling efforts is based on the simplicity of the interactions between the elements: there is no ambiguity as to what interacts with what, and the interaction strength is uniquely determined by the physical distance. We are at a loss, however, to describe systems for which physical distance is irrelevant or for which there is ambiguity as to whether two components interact. While for many complex systems with nontrivial network topology such ambiguity is naturally present, in the past few years we have increasingly recognized that the tools of statistical mechanics offer an ideal framework for describing these interwoven systems as well. These developments have introduced new and challenging problems for statistical physics and unexpected links to major topics in

condensed-matter physics, ranging from percolation to Bose-Einstein condensation.

Traditionally the study of complex networks has been the territory of graph theory. While graph theory initially focused on regular graphs, since the 1950s large-scale networks with no apparent design principles have been described as random graphs, proposed as the simplest and most straightforward realization of a complex network. Random graphs were first studied by the Hungarian mathematicians Paul Erdős and Alfréd Rényi. According to the Erdős-Rényi model, we start with N nodes and connect every pair of nodes with probability p , creating a graph with approximately $pN(N-1)/2$ edges distributed randomly. This model has guided our thinking about complex networks for decades since its introduction. But the growing interest in complex systems has prompted many scientists to reconsider this modeling paradigm and ask a simple question: are the real networks behind such diverse complex systems as the cell or the Internet fundamentally random? Our intuition clearly indicates that complex systems must display some organizing principles, which should be at some level encoded in their topology. But if the topology of these networks indeed deviates from a random graph, we need to develop tools and measurements to capture in quantitative terms the underlying organizing principles.

In the past few years we have witnessed dramatic advances in this direction, prompted by several parallel developments. First, the computerization of data acquisition in all fields led to the emergence of large databases on the topology of various real networks. Second, the increased computing power allowed us to investigate networks containing millions of nodes, exploring questions that could not be addressed before. Third, the slow but noticeable breakdown of boundaries between disciplines offered researchers access to diverse databases, allowing them to uncover the generic properties of complex networks. Finally, there is an increasingly voiced need to move beyond reductionist approaches and try to understand the behavior of the system as a whole. Along this route, understanding the topology of the interactions between the components, i.e., networks, is unavoidable.

Motivated by these converging developments and circumstances, many new concepts and measures have been proposed and investigated in depth in the past few years. However, three concepts occupy a prominent place in contemporary thinking about complex networks. Here we define and briefly discuss them, a discussion to be expanded in the coming sections.

Small worlds: The small-world concept in simple terms describes the fact that despite their often large size, in most networks there is a relatively short path between any two nodes. The distance between two nodes is defined as the number of edges along the shortest path connecting them. The most popular manifestation of small worlds is the “six degrees of separation” concept, uncovered by the social psychologist Stanley Milgram (1967), who concluded that there was a path of

acquaintances with a typical length of about six between most pairs of people in the United States (Kochen, 1989). The small-world property appears to characterize most complex networks: the actors in Hollywood are on average within three co-stars from each other, or the chemicals in a cell are typically separated by three reactions. The small-world concept, while intriguing, is not an indication of a particular organizing principle. Indeed, as Erdős and Rényi have demonstrated, the typical distance between any two nodes in a random graph scales as the logarithm of the number of nodes. Thus random graphs are small worlds as well.

Clustering: A common property of social networks is that cliques form, representing circles of friends or acquaintances in which every member knows every other member. This inherent tendency to cluster is quantified by the clustering coefficient (Watts and Strogatz, 1998), a concept that has its roots in sociology, appearing under the name “fraction of transitive triples” (Wassermann and Faust, 1994). Let us focus first on a selected node i in the network, having k_i edges which connect it to k_i other nodes. If the nearest neighbors of the original node were part of a clique, there would be $k_i(k_i-1)/2$ edges between them. The ratio between the number E_i of edges that actually exist between these k_i nodes and the total number $k_i(k_i-1)/2$ gives the value of the clustering coefficient of node i ,

$$C_i = \frac{2E_i}{k_i(k_i-1)}. \quad (1)$$

The clustering coefficient of the whole network is the average of all individual C_i 's. An alternative definition of C that is often used in the literature is discussed in Sec. VI.B.2 (Barrat and Weigt, 2000; Newman, Strogatz, and Watts, 2000).

In a random graph, since the edges are distributed randomly, the clustering coefficient is $C=p$ (Sec. III.F). However, in most, if not all, real networks the clustering coefficient is typically much larger than it is in a comparable random network (i.e., having the same number of nodes and edges as the real network).

Degree distribution: Not all nodes in a network have the same number of edges (same *node degree*). The spread in the node degrees is characterized by a distribution function $P(k)$, which gives the probability that a randomly selected node has exactly k edges. Since in a random graph the edges are placed randomly, the majority of nodes have approximately the same degree, close to the average degree $\langle k \rangle$ of the network. The degree distribution of a random graph is a Poisson distribution with a peak at $P(\langle k \rangle)$. One of the most interesting developments in our understanding of complex networks was the discovery that for most large networks the degree distribution significantly deviates from a Poisson distribution. In particular, for a large number of networks, including the World Wide Web (Albert, Jeong, and Barabási, 1999), the Internet (Faloutsos *et al.*, 1999), or metabolic networks (Jeong *et al.*, 2000), the degree distribution has a power-law tail,

$$P(k) \sim k^{-\gamma}. \quad (2)$$

Such networks are called scale free (Barabási and Albert, 1999). While some networks display an exponential tail, often the functional form of $P(k)$ still deviates significantly from the Poisson distribution expected for a random graph.

These discoveries have initiated a revival of network modeling in the past few years, resulting in the introduction and study of three main classes of modeling paradigms. First, random graphs, which are variants of the Erdős-Rényi model, are still widely used in many fields and serve as a benchmark for many modeling and empirical studies. Second, motivated by clustering, a class of models, collectively called small-world models, has been proposed. These models interpolate between the highly clustered regular lattices and random graphs. Finally, the discovery of the power-law degree distribution has led to the construction of various scale-free models that, by focusing on the network dynamics, aim to offer a universal theory of network evolution.

The purpose of this article is to review each of these modeling efforts, focusing on the statistical mechanics of complex networks. Our main goal is to present the theoretical developments in parallel with the empirical data that initiated and support the various models and theoretical tools. To achieve this, we start with a brief description of the real networks and databases that represent the testing ground for most current modeling efforts.

II. THE TOPOLOGY OF REAL NETWORKS: EMPIRICAL RESULTS

The study of most complex networks has been initiated by a desire to understand various real systems, ranging from communication networks to ecological webs. Thus the databases available for study span several disciplines. In this section we review briefly those that have been studied by researchers aiming to uncover the general features of complex networks. Beyond a description of the databases, we shall focus on three robust measures of a network's topology: average path length, clustering coefficient, and degree distribution. Other quantities, as discussed in the following sections, will again be tested on these databases. The properties of the investigated databases, as well as the obtained exponents, are summarized in Tables I and II.

A. World Wide Web

The World Wide Web represents the largest network for which topological information is currently available. The nodes of the network are the documents (web pages) and the edges are the hyperlinks (URL's) that point from one document to another (see Fig. 1). The size of this network was close to one billion nodes at the end of 1999 (Lawrence and Giles, 1998, 1999). The interest in the World Wide Web as a network boomed after it was discovered that the degree distribution of the web pages follows a power law over several orders of magnitude (Albert, Jeong, and Barabási, 1999; Kumar

TABLE I. The general characteristics of several real networks. For each network we have indicated the number of nodes, the average degree $\langle k \rangle$, the average path length ℓ , and the clustering coefficient C . For a comparison we have included the average path length ℓ_{rand} and clustering coefficient C_{rand} of a random graph of the same size and average degree. The numbers in the last column are keyed to the symbols in Figs. 8 and 9.

Network	Size	$\langle k \rangle$	ℓ	ℓ_{rand}	C	C_{rand}	Reference	Nr.
WWW, site level, undir.	153 127	35.21	3.1	3.35	0.1078	0.00023	Adamic, 1999	1
Internet, domain level	3015–6209	3.52–4.11	3.7–3.76	6.36–6.18	0.18–0.3	0.001	Yook <i>et al.</i> , 2001a, Pastor-Satorras <i>et al.</i> , 2001	2
Movie actors	225 226	61	3.65	2.99	0.79	0.00027	Watts and Strogatz, 1998	3
LANL co-authorship	52 909	9.7	5.9	4.79	0.43	1.8×10^{-4}	Newman, 2001a, 2001b, 2001c	4
MEDLINE co-authorship	1 520 251	18.1	4.6	4.91	0.066	1.1×10^{-5}	Newman, 2001a, 2001b, 2001c	5
SPIRES co-authorship	56 627	173	4.0	2.12	0.726	0.003	Newman, 2001a, 2001b, 2001c	6
NCSTRL co-authorship	11 994	3.59	9.7	7.34	0.496	3×10^{-4}	Newman, 2001a, 2001b, 2001c	7
Math. co-authorship	70 975	3.9	9.5	8.2	0.59	5.4×10^{-5}	Barabási <i>et al.</i> , 2001	8
Neurosci. co-authorship	209 293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabási <i>et al.</i> , 2001	9
<i>E. coli</i> , substrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000	10
<i>E. coli</i> , reaction graph	315	28.3	2.62	1.98	0.59	0.09	Wagner and Fell, 2000	11
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06	Montoya and Solé, 2000	12
Silwood Park food web	154	4.75	3.40	3.23	0.15	0.03	Montoya and Solé, 2000	13
Words, co-occurrence	460 902	70.13	2.67	3.03	0.437	0.0001	Ferrer i Cancho and Solé, 2001	14
Words, synonyms	22 311	13.48	4.5	3.84	0.7	0.0006	Yook <i>et al.</i> , 2001b	15
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998	16
<i>C. Elegans</i>	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998	17

et al., 1999). Since the edges of the World Wide Web are directed, the network is characterized by two degree distributions: the distribution of outgoing edges, $P_{out}(k)$, signifies the probability that a document has k outgoing hyperlinks, and the distribution of incoming edges, $P_{in}(k)$, is the probability that k hyperlinks point to a certain document. Several studies have established that both $P_{out}(k)$ and $P_{in}(k)$ have power-law tails:

$$P_{out}(k) \sim k^{-\gamma_{out}} \quad \text{and} \quad P_{in}(k) \sim k^{-\gamma_{in}}. \quad (3)$$

Albert, Jeong, and Barabási (1999) have studied a subset of the World Wide Web containing 325 729 nodes and have found $\gamma_{out}=2.45$ and $\gamma_{in}=2.1$. Kumar *et al.* (1999) used a 40-million-document crawl by Alexa Inc., obtaining $\gamma_{out}=2.38$ and $\gamma_{in}=2.1$ (see also Kleinberg *et al.*, 1999). A later survey of the World Wide Web topology by Broder *et al.* (2000) used two 1999 Altavista crawls containing in total 200 million documents, obtaining $\gamma_{out}=2.72$ and $\gamma_{in}=2.1$ with scaling holding close to five orders of magnitude (Fig. 2). Adamic and Huberman (2000) used a somewhat different representation of the World Wide Web, with each node representing a separate domain name and two nodes being connected if any of the pages in one domain linked to any page in the other. While this method lumped together pages that were on the same domain, representing a nontrivial aggregation of the nodes, the distribution of incoming edges still followed a power law with $\gamma_{in}^{dom}=1.94$.

Note that γ_{in} is the same for all measurements at the document level despite the two-years' time delay between the first and last web crawl, during which the World Wide Web had grown at least five times larger. However, γ_{out} has a tendency to increase with the sample size or time (see Table II).

Despite the large number of nodes, the World Wide Web displays the small-world property. This was first reported by Albert, Jeong, and Barabási (1999), who found that the average path length for a sample of 325 729 nodes was 11.2 and predicted, using finite size scaling, that for the full World Wide Web of 800 million nodes that would be a path length of around 19. Subsequent measurements by Broder *et al.* (2000) found that the average path length between nodes in a 50-million-node sample of the World Wide Web is 16, in agreement with the finite size prediction for a sample of this size. Finally, the domain-level network displays an average path length of 3.1 (Adamic, 1999).

The directed nature of the World Wide Web does not allow us to measure the clustering coefficient using Eq. (1). One way to avoid this difficulty is to make the network undirected, making each edge bidirectional. This was the path followed by Adamic (1999), who studied the World Wide Web at the domain level using a 1997 Alexa crawl of 50 million web pages distributed among 259 794 sites. Adamic removed the nodes that had only one edge, focusing on a network of 153 127 sites. While these modifications are expected to increase the clustering coefficient somewhat, she found $C=0.1078$, orders of magnitude higher than $C_{rand}=0.00023$ corresponding to a random graph of the same size and average degree.

B. Internet

The Internet is a network of physical links between computers and other telecommunication devices (Fig.

TABLE II. The scaling exponents characterizing the degree distribution of several scale-free networks, for which $P(k)$ follows a power law (2). We indicate the size of the network, its average degree $\langle k \rangle$, and the cutoff κ for the power-law scaling. For directed networks we list separately the indegree (γ_{in}) and outdegree (γ_{out}) exponents, while for the undirected networks, marked with an asterisk (*), these values are identical. The columns ℓ_{real} , ℓ_{rand} , and ℓ_{pow} compare the average path lengths of real networks with power-law degree distribution and the predictions of random-graph theory (17) and of Newman, Strogatz, and Watts (2001) [also see Eq. (63) above], as discussed in Sec. V. The numbers in the last column are keyed to the symbols in Figs. 8 and 9.

Network	Size	$\langle k \rangle$	κ	γ_{out}	γ_{in}	ℓ_{real}	ℓ_{rand}	ℓ_{pow}	Reference	Nr.
WWW	325 729	4.51	900	2.45	2.1	11.2	8.32	4.77	Albert, Jeong, and Barabási 1999	1
WWW	4×10^7	7		2.38	2.1				Kumar <i>et al.</i> , 1999	2
WWW	2×10^8	7.5	4000	2.72	2.1	16	8.85	7.61	Broder <i>et al.</i> , 2000	3
WWW, site	260 000				1.94				Huberman and Adamic, 2000	4
Internet, domain*	3015–4389	3.42–3.76	30–40	2.1–2.2	2.1–2.2	4	6.3	5.2	Faloutsos, 1999	5
Internet, router*	3888	2.57	30	2.48	2.48	12.15	8.75	7.67	Faloutsos, 1999	6
Internet, router*	150 000	2.66	60	2.4	2.4	11	12.8	7.47	Govindan, 2000	7
Movie actors*	212 250	28.78	900	2.3	2.3	4.54	3.65	4.01	Barabási and Albert, 1999	8
Co-authors, SPIRES*	56 627	173	1100	1.2	1.2	4	2.12	1.95	Newman, 2001b	9
Co-authors, neuro.*	209 293	11.54	400	2.1	2.1	6	5.01	3.86	Barabási <i>et al.</i> , 2001	10
Co-authors, math.*	70 975	3.9	120	2.5	2.5	9.5	8.2	6.53	Barabási <i>et al.</i> , 2001	11
Sexual contacts*	2810			3.4	3.4				Liljeros <i>et al.</i> , 2001	12
Metabolic, <i>E. coli</i>	778	7.4	110	2.2	2.2	3.2	3.32	2.89	Jeong <i>et al.</i> , 2000	13
Protein, <i>S. cerev.</i> *	1870	2.39		2.4	2.4				Jeong, Mason, <i>et al.</i> , 2001	14
Ythan estuary*	134	8.7	35	1.05	1.05	2.43	2.26	1.71	Montoya and Solé, 2000	14
Silwood Park*	154	4.75	27	1.13	1.13	3.4	3.23	2	Montoya and Solé, 2000	16
Citation	783 339	8.57			3				Redner, 1998	17
Phone call	53×10^6	3.16		2.1	2.1				Aiello <i>et al.</i> , 2000	18
Words, co-occurrence*	460 902	70.13		2.7	2.7				Ferrer i Cancho and Solé, 2001	19
Words, synonyms*	22 311	13.48		2.8	2.8				Yook <i>et al.</i> , 2001b	20

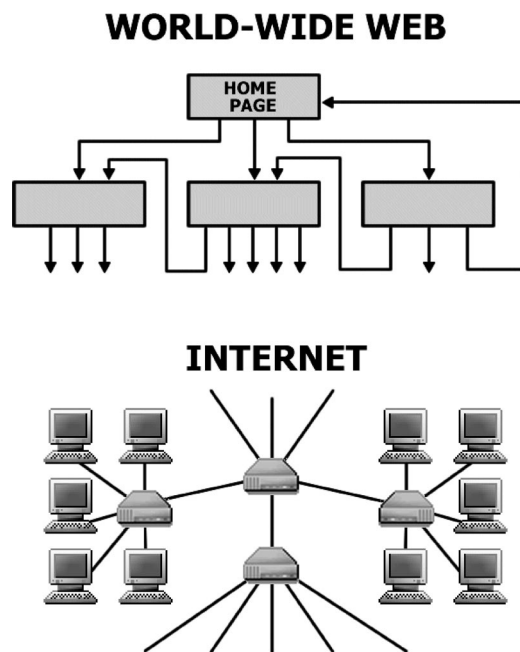


FIG. 1. Network structure of the World Wide Web and the Internet. Upper panel: the nodes of the World Wide Web are web documents, connected with directed hyperlinks (URLs). Lower panel: on the Internet the nodes are the routers and computers, and the edges are the wires and cables that physically connect them. Figure courtesy of István Albert.

1). The topology of the Internet is studied at two different levels. At the router level, the nodes are the routers, and edges are the physical connections between them. At the interdomain (or autonomous system) level, each

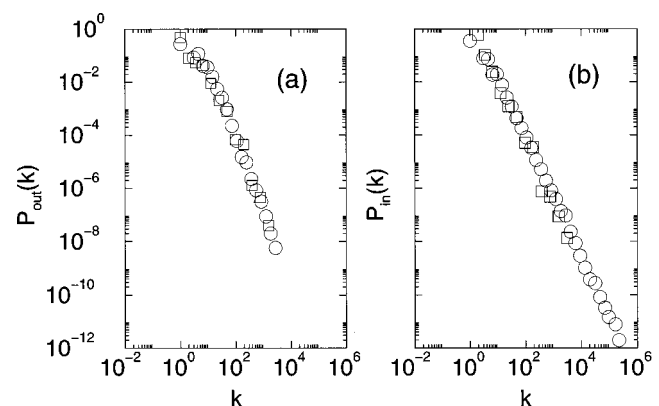


FIG. 2. Degree distribution of the World Wide Web from two different measurements: \square , the 325 729-node sample of Albert *et al.* (1999); \circ , the measurements of over 200 million pages by Broder *et al.* (2000); (a) degree distribution of the outgoing edges; (b) degree distribution of the incoming edges. The data have been binned logarithmically to reduce noise. Courtesy of Altavista and Andrew Tomkins. The authors wish to thank Luis Amaral for correcting a mistake in a previous version of this figure (see Mossa *et al.*, 2001).

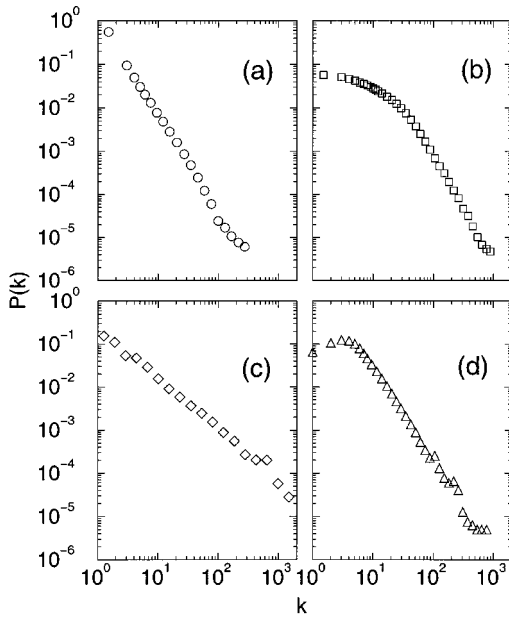


FIG. 3. The degree distribution of several real networks: (a) Internet at the router level. Data courtesy of Ramesh Govindan; (b) movie actor collaboration network. After Barabási and Albert 1999. Note that if TV series are included as well, which aggregate a large number of actors, an exponential cut-off emerges for large k (Amaral *et al.*, 2000); (c) co-authorship network of high-energy physicists. After Newman (2001a, 2001b); (d) co-authorship network of neuroscientists. After Barabási *et al.* (2001).

domain, composed of hundreds of routers and computers, is represented by a single node, and an edge is drawn between two domains if there is at least one route that connects them. Faloutsos *et al.* (1999) have studied the Internet at both levels, concluding that in each case the degree distribution follows a power law. The inter-domain topology of the Internet, captured at three different dates between 1997 and the end of 1998, resulted in degree exponents between $\gamma_I^{as}=2.15$ and $\gamma_I^{as}=2.2$. The 1995 survey of Internet topology at the router level, containing 3888 nodes, found $\gamma_I^r=2.48$ (Faloutsos *et al.*, 1999). Recently Govindan and Tangmunarunkit (2000) mapped the connectivity of nearly 150 000 router interfaces and nearly 200 000 router adjacencies, confirming the power-law scaling with $\gamma_I^r \approx 2.3$ [see Fig. 3(a)].

The Internet as a network does display clustering and small path length as well. Yook *et al.* (2001a) and Pastor-Satorras *et al.* (2001), studying the Internet at the domain level between 1997 and 1999, found that its clustering coefficient ranged between 0.18 and 0.3, to be compared with $C_{rand} \approx 0.001$ for random networks with similar parameters. The average path length of the Internet at the domain level ranged between 3.70 and 3.77 (Pastor-Satorras *et al.*, 2001; Yook *et al.* 2001a) and at the router level it was around 9 (Yook *et al.*, 2001a), indicating its small-world character.

C. Movie actor collaboration network

A much-studied database is the movie actor collaboration network, based on the Internet Movie Database,

which contains all movies and their casts since the 1890s. In this network the nodes are the actors, and two nodes have a common edge if the corresponding actors have acted in a movie together. This is a continuously expanding network, with 225 226 nodes in 1998 (Watts and Strogatz, 1998), which grew to 449 913 nodes by May 2000 (Newman, Strogatz, and Watts, 2000). The average path length of the actor network is close to that of a random graph with the same size and average degree, 3.65 compared with 2.9, but its clustering coefficient is more than 100 times higher than a random graph (Watts and Strogatz, 1998). The degree distribution of the movie actor network has a power-law tail for large k [see Fig. 3(b)], following $P(k) \sim k^{-\gamma_{actor}}$, where $\gamma_{actor} = 2.3 \pm 0.1$ (Barabási and Albert, 1999; Albert and Barabási, 2000; Amaral *et al.*, 2000).

D. Science collaboration graph

A collaboration network similar to that of the movie actors can be constructed for scientists, where the nodes are the scientists and two nodes are connected if the two scientists have written an article together. To uncover the topology of this complex graph, Newman (2001a, 2001b, 2001c) studied four databases spanning physics, biomedical research, high-energy physics, and computer science over a five-year window (1995–1999). All these networks show a small average path length but a high clustering coefficient, as summarized in Table I. The degree distribution of the collaboration network of high-energy physicists is an almost perfect power law with an exponent of 1.2 [Fig. 3(c)], while the other databases display power laws with a larger exponent in the tail.

Barabási *et al.* (2001) investigated the collaboration graph of mathematicians and neuroscientists publishing between 1991 and 1998. The average path length of these networks is around $\ell_{math}=9.5$ and $\ell_{nscl}=6$, their clustering coefficient being $C_{math}=0.59$ and $C_{nscl}=0.76$. The degree distributions of these collaboration networks are consistent with power laws with degree exponents 2.1 and 2.5, respectively [see Fig. 3(d)].

E. The web of human sexual contacts

Many sexually transmitted diseases, including AIDS, spread on a network of sexual relationships. Liljeros *et al.* (2001) have studied the web constructed from the sexual relations of 2810 individuals, based on an extensive survey conducted in Sweden in 1996. Since the edges in this network are relatively short lived, they analyzed the distribution of partners over a single year, obtaining for both females and males a power-law degree distribution with an exponent $\gamma_f = 3.5 \pm 0.2$ and $\gamma_m = 3.3 \pm 0.2$, respectively.

F. Cellular networks

Jeong *et al.* (2000) studied the metabolism of 43 organisms representing all three domains of life, reconstructing them in networks in which the nodes are the

substrates (such as ATP, ADP, H_2O) and the edges represent the predominantly directed chemical reactions in which these substrates can participate. The distributions of the outgoing and incoming edges have been found to follow power laws for all organisms, with the degree exponents varying between 2.0 and 2.4. While due to the network's directedness the clustering coefficient has not been determined, the average path length was found to be approximately the same in all organisms, with a value of 3.3.

The clustering coefficient was studied by Wagner and Fell (2000; see also Fell and Wagner, 2000), focusing on the energy and biosynthesis metabolism of the *Escherichia coli* bacterium. They found that, in addition to the power law degree distribution, the undirected version of this substrate graph has a small average path length and a large clustering coefficient (see Table I).

Another important network characterizing the cell describes protein-protein interactions, where the nodes are proteins and they are connected if it has been experimentally demonstrated that they bind together. A study of these physical interactions shows that the degree distribution of the physical protein interaction map for yeast follows a power law with an exponential cutoff $P(k) \sim (k + k_0)^{-\gamma} e^{-(k+k_0)/k_c}$ with $k_0=1$, $k_c=20$, and $\gamma=2.4$ (Jeong, Mason, *et al.*, 2001).

G. Ecological networks

Food webs are used regularly by ecologists to quantify the interaction between various species (Pimm, 1991). In a food web the nodes are species and the edges represent predator-prey relationships between them. In a recent study, Williams *et al.* (2000) investigated the topology of the seven most documented and largest food webs, namely, those of Skipwith Pond, Little Rock Lake, Bridge Brook Lake, Chesapeake Bay, Ythan Estuary, Coachella Valley, and St. Martin Island. While these webs differ widely in the number of species or their average degree, they all indicate that species in habitats are three or fewer edges from each other. This result was supported by the independent investigations of Montoya and Solé (2000) and Camacho *et al.* (2001a), who showed that food webs are highly clustered as well. The degree distribution was first addressed by Montoya and Solé (2000), focusing on the food webs of Ythan Estuary, Silwood Park, and Little Rock Lake, considering these networks as being nondirected. Although the size of these webs is small (the largest of them has 186 nodes), they appear to share the nonrandom properties of their larger counterparts. In particular, Montoya and Solé (2000) concluded that the degree distribution is consistent with a power law with an unusually small exponent of $\gamma \approx 1.1$. The small size of these webs does leave room, however, for some ambiguity in $P(k)$. Camacho *et al.* (2001a, 2001b) find that for some food webs an exponential fit works equally well. While the well-documented existence of key species that play an important role in food web topology points towards the existence of hubs (a common feature of scale-free networks), an unam-

biguous determination of the network's topology could benefit from larger datasets. Due to the inherent difficulty in the data collection process (Williams *et al.*, 2000), this is not expected anytime soon.

H. Phone call network

A large directed graph has been constructed from long-distance telephone call patterns, where nodes are phone numbers and every completed phone call is an edge, directed from the caller to the receiver. Abello, Pardalos, and Resende (1999) and Aiello, Chung, and Lu (2000) studied the call graph of long-distance telephone calls made during a single day, finding that the degree distributions of the outgoing and incoming edges followed a power law with exponent $\gamma_{out} = \gamma_{in} = 2.1$.

I. Citation networks

A rather complex network is formed by the citation patterns of scientific publications, the nodes standing for published articles and a directed edge representing a reference to a previously published article. Redner (1998), studying the citation distribution of 783 339 papers cataloged by the Institute for Scientific Information and 24 296 papers published in *Physical Review D* between 1975 and 1994, has found that the probability that a paper is cited k times follows a power law with exponent $\gamma_{cite} = 3$, indicating that the incoming degree distribution of the citation network follows a power law. A recent study by Vázquez (2001) extended these studies to the outgoing degree distribution as well, finding that it has an exponential tail.

J. Networks in linguistics

The complexity of human languages offers several possibilities for defining and studying complex networks. Recently Ferrer i Cancho and Solé (2001) have constructed such a network for the English language, based on the British National Corpus, with words as nodes; these nodes are linked if they appear next to or one word apart from each other in sentences. They have found that the resulting network of 440 902 words displays a small average path length $\ell = 2.67$, a high clustering coefficient $C = 0.437$, and a two-regime power-law degree distribution. Words with degree $k \leq 10^3$ decay with a degree exponent $\gamma_{<} = 1.5$, while words with $10^3 < k < 10^5$ follow a power law with $\gamma_{>} \approx 2.7$.

A different study (Yook, Jeong, and Barabási, 2001b) linked words based on their meanings, i.e., two words were connected to each other if they were known to be synonyms according to the Merriam-Webster Dictionary. The results indicate the existence of a giant cluster of 22 311 words from the total of 23 279 words that have synonyms, with an average path length $\ell = 4.5$, and a rather high clustering coefficient $C = 0.7$ compared to $C_{rand} = 0.0006$ for an equivalent random network. In addition, the degree distribution followed had a power-law

tail with $\gamma_{syn}=2.8$. These results indicate that in many respects language also forms a complex network with organizing principles not so different from the examples discussed earlier (see also Steyvers and Tenenbaum, 2001).

K. Power and neural networks

The power grid of the western United States is described by a complex network whose nodes are generators, transformers, and substations, and the edges are high-voltage transmission lines. The number of nodes in the power grid is $N=4941$, and $\langle k \rangle=2.67$. In the tiny ($N=282$) neural network of the nematode worm *C. elegans*, the nodes are the neurons, and an edge joins two neurons if they are connected by either a synapse or a gap junction. Watts and Strogatz (1998) found that, while for both networks the average path length was approximately equal to that of a random graph of the same size and average degree, their clustering coefficient was much higher (Table I). The degree distribution of the power grid is consistent with an exponential, while for the *C. elegans* neural network it has a peak at an intermediate k after which it decays following an exponential (Amaral *et al.*, 2000).

L. Protein folding

During folding a protein takes up consecutive conformations. Representing with a node each distinct state, two conformations are linked if they can be obtained from each other by an elementary move. Scala, Amaral, and Barthélemy (2001) studied the network formed by the conformations of a two-dimensional (2D) lattice polymer, finding that it has small-world properties. Specifically, the average path length increases logarithmically when the size of the polymer (and consequently the size of the network) increases, similarly to the behavior seen in a random graph. The clustering coefficient, however, is much larger than C_{rand} , a difference that increases with the network size. The degree distribution of this conformation network is consistent with a Gaussian (Amaral *et al.*, 2000).

The databases discussed above served as motivation and a source of inspiration for uncovering the topological properties of real networks. We shall refer to them frequently to validate various theoretical predictions or to understand the limitations of the modeling efforts. In the remainder of this review we discuss the various theoretical tools developed to model these complex networks. In this respect, we need to start with the mother of all network models: the random-graph theory of Erdős and Rényi.

III. RANDOM-GRAPH THEORY

In mathematical terms a network is represented by a graph. A graph is a pair of sets $G=\{P, E\}$, where P is a set of N nodes (or vertices or points) P_1, P_2, \dots, P_N and E is a set of edges (or links or lines) that connect two

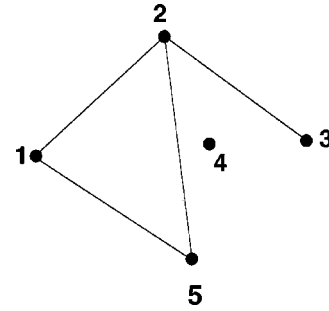


FIG. 4. Illustration of a graph with $N=5$ nodes and $n=4$ edges. The set of nodes is $P=\{1,2,3,4,5\}$ and the edge set is $E=\{\{1,2\}, \{1,5\}, \{2,3\}, \{2,5\}\}$.

elements of P . Graphs are usually represented as a set of dots, each corresponding to a node, two of these dots being joined by a line if the corresponding nodes are connected (see Fig. 4).

Graph theory has its origins in the eighteenth century in the work of Leonhard Euler, the early work concentrating on small graphs with a high degree of regularity. In the twentieth century graph theory has become more statistical and algorithmic. A particularly rich source of ideas has been the study of random graphs, graphs in which the edges are distributed randomly. Networks with a complex topology and unknown organizing principles often appear random; thus random-graph theory is regularly used in the study of complex networks.

The theory of random graphs was introduced by Paul Erdős and Alfréd Rényi (1959, 1960, 1961) after Erdős discovered that probabilistic methods were often useful in tackling problems in graph theory. A detailed review of the field is available in the classic book of Bollobás (1985), complemented by Cohen's (1988) review of the parallels between phase transitions and random-graph theory, and by Karoński and Rućinski's (1997) guide to the history of the Erdős-Rényi approach. Here we briefly describe the most important results of random-graph theory, focusing on the aspects that are of direct relevance to complex networks.

A. The Erdős-Rényi model

In their classic first article on random graphs, Erdős and Rényi define a random graph as N labeled nodes connected by n edges, which are chosen randomly from the $N(N-1)/2$ possible edges (Erdős and Rényi, 1959). In total there are $C_{[N(N-1)/2]}^n$ graphs with N nodes and n edges, forming a probability space in which every realization is equiprobable.

An alternative and equivalent definition of a random graph is the binomial model. Here we start with N nodes, every pair of nodes being connected with probability p (see Fig. 5). Consequently the total number of edges is a random variable with the expectation value $E(n)=p[N(N-1)/2]$. If G_0 is a graph with nodes P_1, P_2, \dots, P_N and n edges, the probability of obtaining it by this graph construction process is $P(G_0)=p^n(1-p)^{N(N-1)/2-n}$.

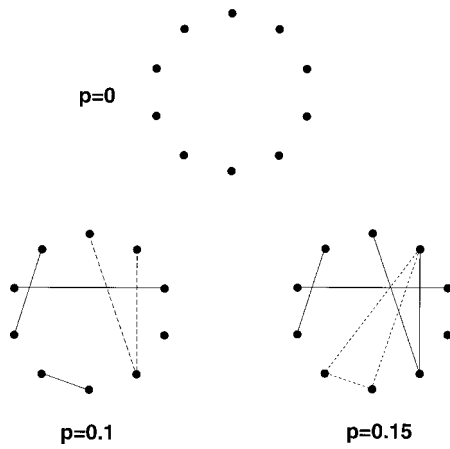


FIG. 5. Illustration of the graph evolution process for the Erdős-Rényi model. We start with $N=10$ isolated nodes (upper panel), then connect every pair of nodes with probability p . The lower panel of the figure shows two different stages in the graph's development, corresponding to $p=0.1$ and $p=0.15$. We can notice the emergence of trees (a tree of order 3, drawn with long-dashed lines) and cycles (a cycle of order 3, drawn with short-dashed lines) in the graph, and a connected cluster that unites half of the nodes at $p=0.15=1.5/N$.

Random-graph theory studies the properties of the probability space associated with graphs with N nodes as $N \rightarrow \infty$. Many properties of such random graphs can be determined using probabilistic arguments. In this respect Erdős and Rényi used the definition that almost every graph has a property Q if the probability of having Q approaches 1 as $N \rightarrow \infty$. Among the questions addressed by Erdős and Rényi, some have direct relevance to an understanding of complex networks as well, such as: Is a typical graph connected? Does it contain a triangle of connected nodes? How does its diameter depend on its size?

In the mathematical literature the construction of a random graph is often called an evolution: starting with a set of N isolated vertices, the graph develops by the successive addition of random edges. The graphs obtained at different stages of this process correspond to larger and larger connection probabilities p , eventually obtaining a fully connected graph [having the maximum number of edges $n=N(N-1)/2$] for $p \rightarrow 1$. The main goal of random-graph theory is to determine at what connection probability p a particular property of a graph will most likely arise. The greatest discovery of Erdős and Rényi was that many important properties of random graphs appear quite suddenly. That is, at a given probability either almost every graph has some property Q (e.g., every pair of nodes is connected by a path of consecutive edges) or, conversely, almost no graph has it. The transition from a property's being very unlikely to its being very likely is usually swift. For many such properties there is a critical probability $p_c(N)$. If $p(N)$ grows more slowly than $p_c(N)$ as $N \rightarrow \infty$, then almost every graph with connection probability $p(N)$ fails to have Q . If $p(N)$ grows somewhat faster than $p_c(N)$, then almost every graph has the property Q . Thus the

probability that a graph with N nodes and connection probability $p=p(N)$ has property Q satisfies

$$\lim_{N \rightarrow \infty} P_{N,p}(Q) = \begin{cases} 0 & \text{if } \frac{p(N)}{p_c(N)} \rightarrow 0 \\ 1 & \text{if } \frac{p(N)}{p_c(N)} \rightarrow \infty. \end{cases} \quad (4)$$

An important note is in order here. Physicists trained in critical phenomena will recognize in $p_c(N)$ the critical probability familiar in percolation. In the physics literature the system is usually viewed at a fixed system size N and then the different regimes in Eq. (4) reduce to the question of whether p is smaller or larger than p_c . The proper value of p_c , that is, the limit $p_c = p_c(N \rightarrow \infty)$, is obtained by finite size scaling. The basis of this procedure is the assumption that this limit exists, reflecting the fact that ultimately the percolation threshold is independent of the system size. This is usually the case in finite-dimensional systems, which include most physical systems of interest for percolation theory and critical phenomena. In contrast, networks are by definition infinite dimensional: the number of neighbors a node can have increases with the system size. Consequently in random-graph theory the occupation probability is defined as a function of the system size: p represents the fraction of the edges that are present from the possible $N(N-1)/2$. Larger graphs with the same p will contain more edges, and consequently properties like the appearance of cycles could occur for smaller p in large graphs than in smaller ones. This means that for many properties Q in random graphs there is no unique, N -independent threshold, but we have to define a threshold function that depends on the system size, and $p_c(N \rightarrow \infty) \rightarrow 0$. However, we shall see that the average degree of the graph

$$\langle k \rangle = 2n/N = p(N-1) \approx pN \quad (5)$$

does have a critical value that is independent of the system size. In the coming subsection we illustrate these ideas by looking at the emergence of various subgraphs in random graphs.

B. Subgraphs

The first property of random graphs to be studied by Erdős and Rényi (1959) was the appearance of subgraphs. A graph G_1 consisting of a set P_1 of nodes and a set E_1 of edges is a subgraph of a graph $G=\{P,E\}$ if all nodes in P_1 are also nodes of P and all edges in E_1 are also edges of E . The simplest examples of subgraphs are cycles, trees, and complete subgraphs (see Fig. 5). A cycle of order k is a closed loop of k edges such that every two consecutive edges and only those have a common node. That is, graphically a triangle is a cycle of order 3, while a rectangle is a cycle of order 4. The average degree of a cycle is equal to 2, since every node has two edges. The opposite of cycles are the trees, which cannot form closed loops. More precisely, a graph is a tree of order k if it has k nodes and $k-1$ edges,

and none of its subgraphs is a cycle. The average degree of a tree of order k is $\langle k \rangle = 2 - 2/k$, approaching 2 for large trees. *Complete subgraphs* of order k contain k nodes and all the possible $k(k-1)/2$ edges—in other words, they are completely connected.

Let us consider the evolution process described in Fig. 5 for a graph $G = G_{N,p}$. We start from N isolated nodes, then connect every pair of nodes with probability p . For small connection probabilities the edges are isolated, but as p , and with it the number of edges, increases, two edges can attach at a common node, forming a tree of order 3. An interesting problem is to determine the critical probability $p_c(N)$ at which almost every graph G contains a tree of order 3. Most generally we can ask whether there is a critical probability that marks the appearance of arbitrary subgraphs consisting of k nodes and l edges.

In random-graph theory there is a rigorously proven answer to this question (Bollobás, 1985). Consider a random graph $G = G_{N,p}$. In addition, consider a small graph F consisting of k nodes and l edges. In principle, the random graph G can contain several such subgraphs F . Our first goal is to determine how many such subgraphs exist. The k nodes can be chosen from the total number of nodes N in C_N^k ways and the l edges are formed with probability p^l . In addition, we can permute the k nodes and potentially obtain $k!$ new graphs (the correct value is $k!/a$, where a is the number of graphs that are isomorphic to each other). Thus the expected number of subgraphs F contained in G is

$$E(X) = C_N^k \frac{k!}{a} p^l \approx \frac{N^k p^l}{a}. \quad (6)$$

This notation suggests that the actual number of such subgraphs, X , can be different from $E(X)$, but in the majority of cases it will be close to it. Note that the subgraphs do not have to be isolated, i.e., there can exist edges with one node inside the subgraph but the other outside of it.

Equation (6) indicates that if $p(N)$ is such that $p(N)N^{k/l} \rightarrow 0$ as $N \rightarrow \infty$, the expected number of subgraphs $E(X) \rightarrow 0$, i.e., almost none of the random graphs contains a subgraph F . However, if $p(N) = cN^{-k/l}$, the mean number of subgraphs is a finite number, denoted by $\lambda = c^l/a$, indicating that this function might be the critical probability. The validity of this finding can be tested by calculating the distribution of subgraph numbers, $P_p(X=r)$, obtaining (Bollobás, 1985)

$$\lim_{N \rightarrow \infty} P_p(X=r) = e^{-\lambda} \frac{\lambda^r}{r!}. \quad (7)$$

The probability that G contains at least one subgraph F is then

$$P_p(G \supset F) = \sum_{r=1}^{\infty} P_p(X=r) = 1 - e^{-\lambda}, \quad (8)$$

which converges to 1 as c increases. For p values satisfying $pN^{k/l} \rightarrow \infty$ the probability $P_p(G \supset F)$ converges to

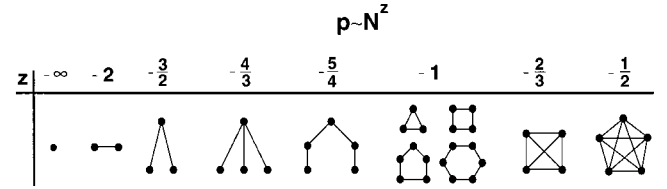


FIG. 6. The threshold probabilities at which different subgraphs appear in a random graph. For $pN^{3/2} \rightarrow 0$ the graph consists of isolated nodes and edges. For $p \sim N^{-3/2}$ trees of order 3 appear, while for $p \sim N^{-4/3}$ trees of order 4 appear. At $p \sim N^{-1}$ trees of all orders are present, and at the same time cycles of all orders appear. The probability $p \sim N^{-2/3}$ marks the appearance of complete subgraphs of order 4 and $p \sim N^{-1/2}$ corresponds to complete subgraphs of order 5. As z approaches 0, the graph contains complete subgraphs of increasing order.

1. Thus, indeed, the critical probability at which almost every graph contains a subgraph with k nodes and l edges is $p_c(N) = cN^{-k/l}$.

A few important special cases directly follow from Eq. (8):

- (a) The critical probability of having a tree of order k is $p_c(N) = cN^{-k/(k-1)}$;
- (b) The critical probability of having a cycle of order k is $p_c(N) = cN^{-1}$;
- (c) The critical probability of having a complete subgraph of order k is $p_c(N) = cN^{-2/(k-1)}$.

C. Graph evolution

It is instructive to look at the results discussed above from a different point of view. Consider a random graph with N nodes and assume that the connection probability $p(N)$ scales as N^z , where z is a tunable parameter that can take any value between $-\infty$ and 0 (Fig. 6). For z less than $-3/2$ almost all graphs contain only isolated nodes and edges. When z passes through $-3/2$, trees of order 3 suddenly appear. When z reaches $-4/3$, trees of order 4 appear, and as z approaches -1 , the graph contains trees of larger and larger order. However, as long as $z < -1$, such that the average degree of the graph $\langle k \rangle = pN \rightarrow 0$ as $N \rightarrow \infty$, the graph is a union of disjoint trees, and cycles are absent. Exactly when z passes through -1 , corresponding to $\langle k \rangle = \text{const}$, even though z is changing smoothly, the asymptotic probability of cycles of all orders jumps from 0 to 1. Cycles of order 3 can also be viewed as complete subgraphs of order 3. Complete subgraphs of order 4 appear at $z = -2/3$, and as z continues to increase, complete subgraphs of larger and larger order continue to emerge. Finally, as z approaches 0, the graph contains complete subgraphs of all finite order.

Further results can be derived for $z = -1$, i.e., when we have $p \propto N^{-1}$ and the average degree of the nodes is $\langle k \rangle = \text{const}$. For $p \propto N^{-1}$ a random graph contains trees and cycles of all order, but so far we have not discussed the size and structure of a typical graph component. A component of a graph is by definition a connected, iso-

lated subgraph, also called a cluster in network research and percolation theory. As Erdős and Rényi (1960) show, there is an abrupt change in the cluster structure of a random graph as $\langle k \rangle$ approaches 1.

If $0 < \langle k \rangle < 1$, almost surely all clusters are either trees or clusters containing exactly one cycle. Although cycles are present, almost all nodes belong to trees. The mean number of clusters is of order $N - n$, where n is the number of edges, i.e., in this range when a new edge is added the number of clusters decreases by 1. The largest cluster is a tree, and its size is proportional to $\ln N$.

When $\langle k \rangle$ passes the threshold $\langle k \rangle_c = 1$, the structure of the graph changes abruptly. While for $\langle k \rangle < 1$ the greatest cluster is a tree, for $\langle k \rangle_c = 1$ it has approximately $N^{2/3}$ nodes and has a rather complex structure. Moreover for $\langle k \rangle > 1$ the greatest (giant) cluster has $[1 - f(\langle k \rangle)]N$ nodes, where $f(x)$ is a function that decreases exponentially from $f(1) = 1$ to 0 for $x \rightarrow \infty$. Thus a finite fraction $S = 1 - f(\langle k \rangle)$ of the nodes belongs to the largest cluster. Except for this giant cluster, all other clusters are relatively small, most of them being trees, the total number of nodes belonging to trees being $Nf(\langle k \rangle)$. As $\langle k \rangle$ increases, the small clusters coalesce and join the giant cluster, the smaller clusters having the higher chance of survival.

Thus at $p_c \approx 1/N$ the random graph changes its topology abruptly from a loose collection of small clusters to a system dominated by a single giant cluster. The beginning of the supercritical phase was studied by Bollobás (1984), Kolchin (1986), and Luczak (1990). Their results show that in this region the largest cluster clearly separates from the rest of the clusters, its size S increasing proportionally with the separation from the critical probability,

$$S \propto (p - p_c). \quad (9)$$

As we shall see in Sec. IV.F, this dependence is analogous to the scaling of the percolation probability in infinite-dimensional percolation.

D. Degree distribution

Erdős and Rényi (1959) were the first to study the distribution of the maximum and minimum degree in a random graph, the full degree distribution being derived later by Bollobás (1981).

In a random graph with connection probability p the degree k_i of a node i follows a binomial distribution with parameters $N-1$ and p :

$$P(k_i = k) = C_{N-1}^k p^k (1-p)^{N-1-k}. \quad (10)$$

This probability represents the number of ways in which k edges can be drawn from a certain node: the probability of k edges is p^k , the probability of the absence of additional edges is $(1-p)^{N-1-k}$, and there are C_{N-1}^k equivalent ways of selecting the k end points for these edges. Furthermore, if i and j are different nodes, $P(k_i = k)$ and $P(k_j = k)$ are close to being independent random variables. To find the degree distribution of the graph, we need to study the number of nodes with de-

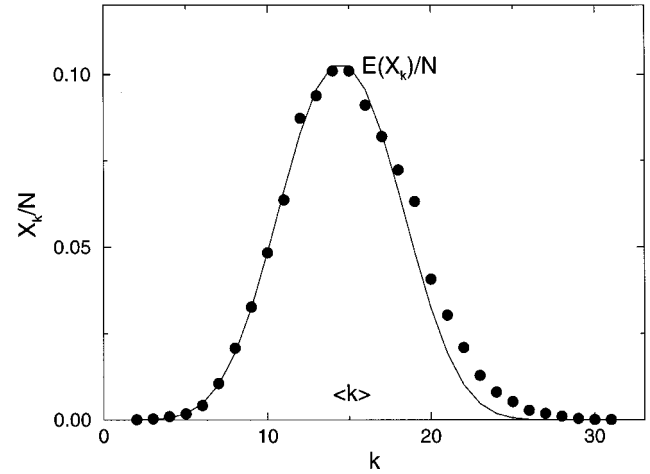


FIG. 7. The degree distribution that results from the numerical simulation of a random graph. We generated a single random graph with $N=10\,000$ nodes and connection probability $p=0.0015$, and calculated the number of nodes with degree k , X_k . The plot compares X_k/N with the expectation value of the Poisson distribution (13), $E(X_k)/N = P(k_i = k)$, and we can see that the deviation is small.

gree k , X_k . Our main goal is to determine the probability that X_k takes on a given value, $P(X_k = r)$.

According to Eq. (10), the expectation value of the number of nodes with degree k is

$$E(X_k) = NP(k_i = k) = \lambda_k, \quad (11)$$

where

$$\lambda_k = NC_{N-1}^k p^k (1-p)^{N-1-k}. \quad (12)$$

As in the derivation of the existence conditions of subgraphs (see Sec. III.B), the distribution of the X_k values, $P(X_k = r)$, approaches a Poisson distribution,

$$P(X_k = r) = e^{-\lambda_k} \frac{\lambda_k^r}{r!}. \quad (13)$$

Thus the number of nodes with degree k follows a Poisson distribution with mean value λ_k . Note that the expectation value of the distribution (13) is the function λ_k given by Eq. (12) and not a constant. The Poisson distribution decays rapidly for large values of r , the standard deviation of the distribution being $\sigma_k = \sqrt{\lambda_k}$. With a bit of simplification we could say that Eq. (13) implies that X_k does not diverge much from the approximative result $X_k = NP(k_i = k)$, valid only if the nodes are independent (see Fig. 7). Thus with a good approximation the degree distribution of a random graph is a binomial distribution,

$$P(k) = C_{N-1}^k p^k (1-p)^{N-1-k}, \quad (14)$$

which for large N can be replaced by a Poisson distribution,

$$P(k) \approx e^{-pN} \frac{(pN)^k}{k!} = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}. \quad (15)$$

Since the pioneering paper of Erdős and Rényi, much work has concentrated on the existence and uniqueness

of the minimum and maximum degree of a random graph. The results indicate that for a large range of p values both the maximum and the minimum degrees are determined and finite. For example, if $p(N) \sim N^{-1-1/k}$ (and thus the graph is a set of isolated trees of order at most $k+1$), almost no graph has nodes with degree higher than k . At the other extreme, if $p = \{\ln(N) + k \ln[\ln(N)] + c\}/N$, almost every random graph has a minimum degree of at least k . Furthermore, for a sufficiently high p , respectively, if $pN/\ln(N) \rightarrow \infty$, the maximum degree of almost all random graphs has the same order of magnitude as the average degree. Thus, despite the fact that the position of the edges is random, a typical random graph is rather homogeneous, the majority of the nodes having the same number of edges.

E. Connectedness and diameter

The diameter of a graph is the maximal distance between any pair of its nodes. Strictly speaking, the diameter of a disconnected graph (i.e., one made up of several isolated clusters) is infinite, but it can be defined as the maximum diameter of its clusters. Random graphs tend to have small diameters, provided p is not too small. The reason for this is that a random graph is likely to be spreading: with large probability the number of nodes at a distance l from a given node is not much smaller than $\langle k \rangle^l$. Equating $\langle k \rangle^l$ with N we find that the diameter is proportional to $\ln(N)/\ln(\langle k \rangle)$; thus it depends only logarithmically on the number of nodes.

The diameter of a random graph has been studied by many authors (see Chung and Lu, 2001). A general conclusion is that for most values of p , almost all graphs with the same N and p have precisely the same diameter. This means that when we consider all graphs with N nodes and connection probability p , the range of values in which the diameters of these graphs can vary is very small, usually concentrated around

$$d = \frac{\ln(N)}{\ln(pN)} = \frac{\ln(N)}{\ln(\langle k \rangle)}. \quad (16)$$

Below we summarize a few important results:

- If $\langle k \rangle = pN < 1$, a typical graph is composed of isolated trees and its diameter equals the diameter of a tree.
- If $\langle k \rangle > 1$, a giant cluster appears. The diameter of the graph equals the diameter of the giant cluster if $\langle k \rangle \geq 3.5$, and is proportional to $\ln(N)/\ln(\langle k \rangle)$.
- If $\langle k \rangle \geq \ln(N)$, almost every graph is totally connected. The diameters of the graphs having the same N and $\langle k \rangle$ are concentrated on a few values around $\ln(N)/\ln(\langle k \rangle)$.

Another way to characterize the spread of a random graph is to calculate the average distance between any pair of nodes, or the average path length. One expects that the average path length scales with the number of nodes in the same way as the diameter,

$$\ell_{rand} \sim \frac{\ln(N)}{\ln(\langle k \rangle)}. \quad (17)$$

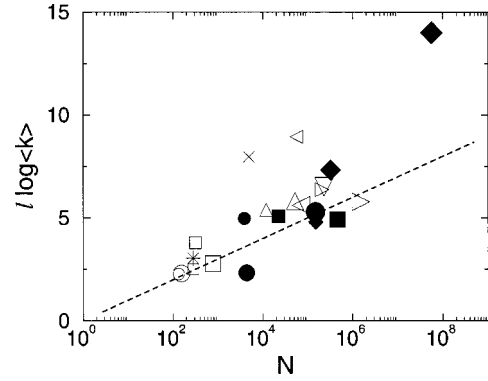


FIG. 8. Comparison between the average path lengths of real networks and the prediction (17) of random-graph theory (dashed line). For each symbol we indicate the corresponding number in Table I or Table II: small \circ , I.12; large \circ , I.13; \star , I.17; small \square , I.10; medium \square , I.11; large \square , II.13; small \bullet , II.6; medium \bullet , I.2; \times , I.16; small \triangle , I.7; small \blacksquare , I.15; large \triangle , I.4; small \triangleleft , I.5; large \triangleleft , I.6; large \bullet , II.6; small \blacklozenge , I.1; small \triangleright , I.7; ∇ , I.3; medium \blacklozenge , II.1; large \blacksquare , I.14; large \triangleright , I.5; large \blacklozenge , II.3.

In Sec. II we presented evidence that the average path length of real networks is close to the average path length of random graphs with the same size. Equation (17) gives us an opportunity to better compare random graphs and real networks (see Newman 2001a, 2001c). According to Eq. (17), the product $\ell_{rand} \ln(\langle k \rangle)$ is equal to $\ln(N)$, so plotting $\ell_{rand} \ln(\langle k \rangle)$ as a function of $\ln(N)$ for random graphs of different sizes gives a straight line of slope 1. In Fig. 8 we plot a similar product for several real networks, $\ell_{real} \log(\langle k \rangle)$, as a function of the network size, comparing it with the prediction of Eq. (17). We can see that the trend of the data is similar to the theoretical prediction, and with several exceptions Eq. (17) gives a reasonable first estimate.

F. Clustering coefficient

As we mentioned in Sec. II, complex networks exhibit a large degree of clustering. If we consider a node in a random graph and its nearest neighbors, the probability that two of these neighbors are connected is equal to the probability that two randomly selected nodes are connected. Consequently the clustering coefficient of a random graph is

$$C_{rand} = p = \frac{\langle k \rangle}{N}. \quad (18)$$

According to Eq. (18), if we plot the ratio $C_{rand}/\langle k \rangle$ as a function of N for random graphs of different sizes, on a log-log plot they will align along a straight line of slope -1 . In Fig. 9 we plot the ratio of the clustering coefficient of real networks and their average degree as a function of their size, comparing it with the prediction of Eq. (18). The plot convincingly indicates that real networks do not follow the prediction of random graphs. The fraction $C/\langle k \rangle$ does not decrease as N^{-1} ; instead, it appears to be independent of N . This property is char-

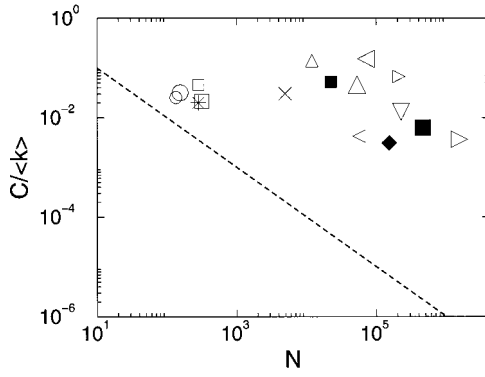


FIG. 9. Comparison between the clustering coefficients of real networks and random graphs. All networks from Table I are included in the figure, the symbols being the same as in Fig. 8. The dashed line corresponds to Eq. (18).

acteristic of large ordered lattices, whose clustering coefficient depends only on the coordination number of the lattice and not their size (Watts and Strogatz, 1998).

G. Graph spectra

Any graph G with N nodes can be represented by its adjacency matrix $A(G)$ with $N \times N$ elements A_{ij} , whose value is $A_{ij} = A_{ji} = 1$ if nodes i and j are connected, and 0 otherwise. The spectrum of graph G is the set of eigenvalues of its adjacency matrix $A(G)$. A graph with N nodes has N eigenvalues λ_j , and it is useful to define its spectral density as

$$\rho(\lambda) = \frac{1}{N} \sum_{j=1}^N \delta(\lambda - \lambda_j), \quad (19)$$

which approaches a continuous function if $N \rightarrow \infty$. The interest in spectral properties is related to the fact that the spectral density can be directly linked to the graph's topological features, since its k th moment can be written as

$$\frac{1}{N} \sum_{j=1}^N (\lambda_j)^k = \frac{1}{N} \sum_{i_1, i_2, \dots, i_k} A_{i_1 i_2} A_{i_2 i_3} \cdots A_{i_k i_1}, \quad (20)$$

i.e., the number of paths returning to the same node in the graph. Note that these paths can contain nodes that were already visited.

Let us consider a random graph $G_{N,p}$ satisfying $p(N) = cN^{-z}$. For $z < 1$ there is an infinite cluster in the graph (see Sec. III.C), and as $N \rightarrow \infty$, any node belongs almost surely to the infinite cluster. In this case the spectral density of the random graph converges to a semicircular distribution (Fig. 10),

$$\rho(\lambda) = \begin{cases} \frac{\sqrt{4Np(1-p) - \lambda^2}}{2\pi Np(1-p)} & \text{if } |\lambda| < 2\sqrt{Np(1-p)} \\ 0 & \text{otherwise.} \end{cases} \quad (21)$$

Known as Wigner's law (see Wigner, 1955, 1957, 1958) or the semicircle law, Eq. (21) has many applications in

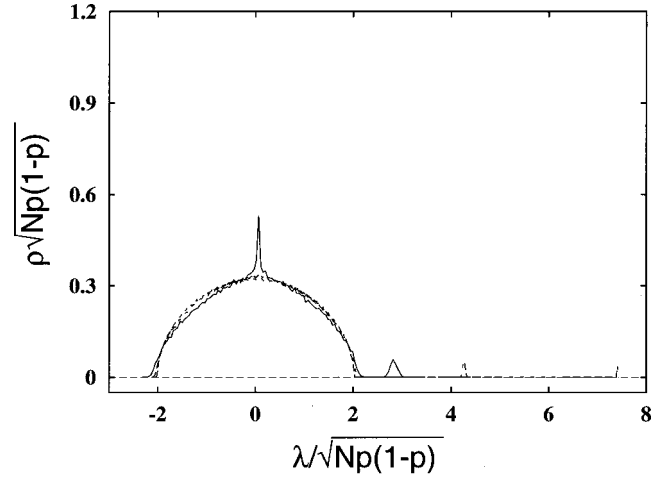


FIG. 10. Rescaled spectral density of three random graphs having $p=0.05$ and size $N=100$ (solid line), $N=300$ (long-dashed line), and $N=1000$ (short-dashed line). The isolated peak corresponds to the principal eigenvalue. After Farkas *et al.* (2001).

quantum, statistical, and solid-state physics (Mehta, 1991; Crisanti *et al.*, 1993; Guhr *et al.*, 1998). The largest (principal) eigenvalue, λ_1 , is isolated from the bulk of the spectrum, and it increases with the network size as pN .

When $z > 1$ the spectral density deviates from the semicircle law. The most striking feature of $\rho(\lambda)$ is that its odd moments are equal to zero, indicating that the only way that a path comes back to the original node is if it returns following exactly the same nodes. This is a salient feature of a tree structure, and, indeed, in Sec. III.B we have seen that in this case the random graph is composed of trees.

IV. PERCOLATION THEORY

One of the most interesting findings of random-graph theory is the existence of a critical probability at which a giant cluster forms. Translated into network language, the theory indicates the existence of a critical probability p_c such that below p_c the network is composed of isolated clusters but above p_c a giant cluster spans the entire network. This phenomenon is markedly similar to a percolation transition, a topic much studied both in mathematics and in statistical mechanics (Stauffer and Aharony, 1992; Bunde and Havlin, 1994, 1996; Grimmett, 1999; ben Avraham and Havlin, 2000). Indeed, a percolation transition and the emergence of a giant cluster are the same phenomenon expressed in different languages. Percolation theory, however, does not simply reproduce the predictions of random-graph theory. Asking questions from a different perspective, it addresses several issues that are crucial for understanding real networks but are not discussed by random graph theory. Consequently it is important to review the predictions of

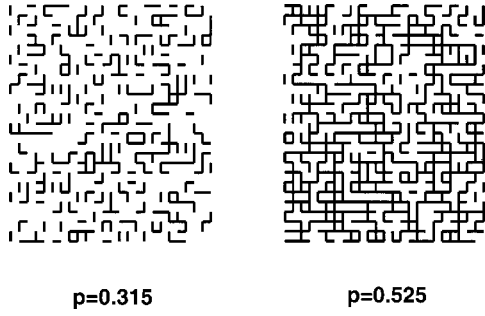


FIG. 11. Illustration of bond percolation in 2D. The nodes are placed on a 25×25 square lattice, and two nodes are connected by an edge with probability p . For $p=0.315$ (left), which is below the percolation threshold $p_c=0.5$, the connected nodes form isolated clusters. For $p=0.525$ (right), which is above the percolation threshold, the largest cluster percolates.

percolation theory relevant to networks, as they are crucial for an understanding of important aspects of the network topology.

A. Quantities of interest in percolation theory

Consider a regular d -dimensional lattice whose edges are present with probability p and absent with probability $1-p$. Percolation theory studies the emergence of paths that percolate through the lattice (starting at one side and ending at the opposite side). For small p only a few edges are present, thus only small clusters of nodes connected by edges can form, but at a critical probability p_c , called the *percolation threshold*, a percolating cluster of nodes connected by edges appears (see Fig. 11). This cluster is also called an infinite cluster, because its size diverges as the size of the lattice increases. There are several much-studied versions of percolation, the one presented above being “bond percolation.” The best-known alternative is site percolation, in which all bonds are present and the nodes of the lattice are occupied with probability p . In a manner similar to bond percolation, for small p only finite clusters of occupied nodes are present, but for $p > p_c$ an infinite cluster appears.

The main quantities of interest in percolation are the following:

- (1) The *percolation probability* P , denoting the probability that a given node belongs to the infinite cluster:

$$P = P_p(|C|=\infty) = 1 - \sum_{s<\infty} P_p(|C|=s), \quad (22)$$

where $P_p(|C|=s)$ denotes the probability that the cluster at the origin has size s . Obviously

$$P = \begin{cases} 0 & \text{if } p < p_c \\ > 0 & \text{if } p > p_c. \end{cases} \quad (23)$$

- (2) The *average cluster size* $\langle s \rangle$, defined as

$$\langle s \rangle = E_p(|C|) = \sum_{s=1}^{\infty} s P_p(|C|=s), \quad (24)$$

giving the expectation value of cluster sizes. Because $\langle s \rangle$ is infinite when $P > 0$, in this case it is useful to

work with the average size of the finite clusters by taking away from the system the infinite ($|C|=\infty$) cluster

$$\langle s \rangle^f = E_p(|C|, |C| < \infty) = \sum_{s<\infty} s P_p(|C|=s). \quad (25)$$

- (3) The *cluster size distribution* n_s , defined as the probability of a node's having a fixed position in a cluster of size s (for example, being its left-hand end, if this position is uniquely defined),

$$n_s = \frac{1}{s} P_p(|C|=s). \quad (26)$$

Note that n_s does not coincide with the probability that a node is part of a cluster of size s . By fixing the position of the node in the cluster we are choosing only one of the s possible nodes, reflected in the fact that $P_p(|C|=s)$ is divided by s , guaranteeing that we count every cluster only once.

These quantities are of interest in random networks as well. There is, however, an important difference between percolation theory and random networks: percolation theory is defined on a regular d -dimensional lattice. In a random network (or graph) we can define a nonmetric distance along the edges, but since any node can be connected by an edge to any other node in the network, there is no regular small-dimensional lattice in which a network can be embedded. However, as we discuss below, random networks and percolation theory meet exactly in the infinite-dimensional limit ($d \rightarrow \infty$) of percolation. Fortunately many results in percolation theory can be generalized to infinite dimensions. Consequently the results obtained within the context of percolation apply directly to random networks as well.

B. General results

1. The subcritical phase ($p < p_c$)

When $p < p_c$, only small clusters of nodes connected by edges are present in the system. The questions asked in this phase are (i) what is the probability that there exists a path $x \leftrightarrow y$ joining two randomly chosen nodes x and y ? and (ii) what is the rate of decay of $P_p(|C|=s)$ when $s \rightarrow \infty$? The first result of this type was obtained by Hammersley (1957), who showed that the probability of a path's joining the origin with a node on the surface, $\partial B(r)$, of a box centered at the origin and with side length $2r$ decays exponentially if $P < \infty$. We can define a correlation length ξ as the characteristic length of the exponential decay

$$P_p[0 \leftrightarrow \partial B(r)] \sim e^{-r/\xi}, \quad (27)$$

where $0 \leftrightarrow \partial B(r)$ means that there is a path from the origin to an arbitrary node on $\partial B(r)$. Equation (27) indicates that the radius of the finite clusters in the subcritical region has an exponentially decaying tail, and the correlation length represents the mean radius of a finite cluster. It was shown (see Grimmett, 1999) that ξ is equal to 0 for $p=0$ and goes to infinity as $p \rightarrow p_c$.

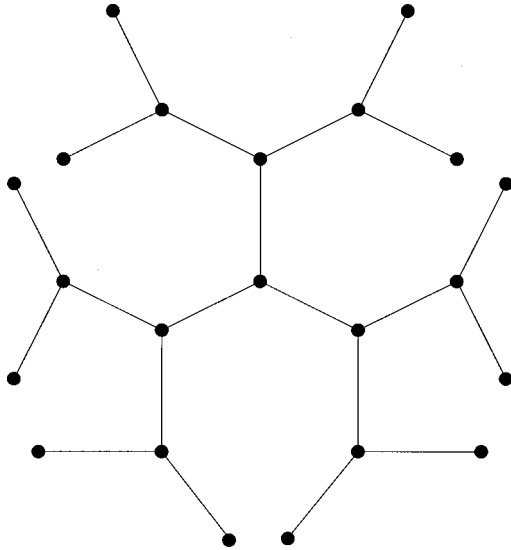


FIG. 12. Example of a Cayley tree with coordination number $z=3$. All of the nodes have three edges, with the exception of those on the surface, which have only one edge. The ratio between the number of nodes on the surface and the total number of nodes approaches a constant, $(z-2)/(z-1)$, a property valid only for infinite-dimensional objects. The average degree approaches $\langle k \rangle = 2$ as the size of the tree goes to infinity, a property held in common with random trees (see Sec. III.B).

The exponential decay of cluster radii implies that the probability that a cluster has size s , $P_p(|C|=s)$, also decays exponentially for large s :

$$P_p(|C|=s) \sim e^{-\alpha(p)s} \quad \text{as } s \rightarrow \infty, \quad (28)$$

where $\alpha(p) \rightarrow \infty$ as $p \rightarrow 0$ and $\alpha(p_c) = 0$.

2. The supercritical phase ($p > p_c$)

For $P > 0$ there is exactly one infinite cluster (Burton and Keane, 1989). In this supercritical phase the previously studied quantities are dominated by the contribution of the infinite cluster; thus it is useful to study the corresponding probabilities in terms of finite clusters. The probability that there is a path from the origin to the surface of a box of edge length $2r$ that is not part of the infinite cluster decays exponentially as

$$P_p[0 \leftrightarrow \partial B(r), |C| < \infty] \sim e^{-r/\xi}. \quad (29)$$

Unlike the subcritical phase, though, the decay of the cluster sizes, $P_p(|C|=s < \infty)$, follows a stretch exponential, $e^{-\beta(p)s^{(d-P)/d}}$, offering the first important quantity that depends on the dimensionality of the lattice, but even this dependence vanishes as $d \rightarrow \infty$, and the cluster size distribution decays exponentially as in the subcritical phase.

C. Exact solutions: Percolation on a Cayley tree

The *Cayley tree* (or Bethe lattice) is a loopless structure (see Fig. 12) in which every node has z neighbors, with the exception of the nodes at the surface. While the

surface and volume of a regular d -dimensional object obey the scaling relation $\text{surface} \propto \text{volume}^{1-1/d}$, and only in the limit $d \rightarrow \infty$ is the surface proportional with the volume, for a Cayley tree the number of nodes on the surface is proportional to the total number of nodes (i.e., the volume of the tree). Thus in this respect a Cayley tree represents an infinite-dimensional object. Another argument for the infinite dimensionality of a Cayley tree is that it has no loops (cycles in graph-theoretic language). Thus, despite its regular topology, the Cayley tree represents a reasonable approximation of the topology of a random network in the subcritical phase, where all the clusters are trees. This is no longer true in the supercritical phase, because at the critical probability $p_c(N)$, cycles of all order appear in the graph (see Sec. III.C).

To investigate percolation on a Cayley tree, we assume that each edge is present with probability p . Next we discuss the main quantities of interest for this system.

- (a) *Percolation threshold*: The condition for the existence of an infinite path starting from the origin is that at least one of the $z-1$ possible outgoing edges of a node is present, i.e., $(z-1)p \geq 1$. Therefore the percolation threshold is

$$p_c = \frac{1}{z-1}. \quad (30)$$

- (b) *Percolation probability*: For a Cayley tree with $z=3$, for which $p_c = 1/2$, the percolation probability is given by (Stauffer and Aharony, 1992)

$$P = \begin{cases} 0 & \text{if } p < p_c = \frac{1}{2} \\ (2p-1)/p^2 & \text{if } p > p_c = \frac{1}{2}. \end{cases} \quad (31)$$

The Taylor series expansion around $p_c = \frac{1}{2}$ gives $P \approx 8(p - \frac{1}{2})$, thus the percolation probability is proportional to the deviation from the percolation threshold

$$P \propto (p - p_c) \quad \text{as } p \rightarrow p_c. \quad (32)$$

- (c) *Mean cluster size*: The average cluster size is given by

$$\langle s \rangle = \sum_{n=1}^{\infty} 3 \times 2^{n-1} p^n = \frac{3}{2} \frac{1}{1-2p} = \frac{3}{4} (p_c - p)^{-1}. \quad (33)$$

Note that $\langle s \rangle$ diverges as $p \rightarrow p_c$, and it depends on p as a power of the distance $p_c - p$ from the percolation threshold. This behavior is an example of critical phenomena: an order parameter goes to zero following a power law in the vicinity of the critical point (Stanley, 1971; Ma, 1976).

- (d) *Cluster size distribution*: The probability of having a cluster of size s is (Durrett, 1985)

$$P_p(|C|=s) = \frac{1}{s} C_{2s}^{s-1} p^{s-1} (1-p)^{s+1}. \quad (34)$$

Here the number of edges surrounding the s nodes is $2s$, from which the $s-1$ inside edges have to be present and the $s+1$ external ones absent. The factor C_{2s}^{s-1} takes into account the different cases that

can be obtained when permuting the edges, and the $1/s$ is a normalization factor. Since $n_s = (1/s) P_p(|C|=s)$, after using Stirling's formula we obtain

$$n_s \propto s^{-5/2} p^{s-1} (1-p)^{s+1}. \quad (35)$$

In the vicinity of the percolation threshold this expression can be approximated as

$$n_s \sim s^{-5/2} e^{-cs} \quad \text{with } c \propto (p-p_c)^2. \quad (36)$$

Thus the cluster size distribution follows a power law with an exponential cutoff: only clusters with size $s < s_\xi = 1/c \propto (p-p_c)^{-2}$ contribute significantly to cluster averages. For these clusters, n_s is effectively equal to $n_s(p_c) \propto s^{-5/2}$. Clusters with $s \gg s_\xi$ are exponentially rare, and their properties are no longer dominated by the behavior at p_c . The notation s_ξ illustrates that as the correlation length ξ is the characteristic length scale for the cluster diameters, s_ξ is an intrinsic characteristic of cluster sizes. The correlation length of a tree is not well defined, but we shall see in the more general cases that s_ξ and ξ are related by a simple power law.

D. Scaling in the critical region

The principal ansatz of percolation theory is that even the most general percolation problem in any dimension obeys a scaling relation similar to Eq. (36) near the percolation threshold. Thus in general the cluster size distribution can be written as

$$n_s(p) \sim \begin{cases} s^{-\tau} f_- (|p-p_c|^{1/\sigma} s) & \text{as } p \leq p_c \\ s^{-\tau} f_+ (|p-p_c|^{1/\sigma} s) & \text{as } p \geq p_c. \end{cases} \quad (37)$$

Here τ and σ are critical exponents whose numerical value needs to be determined, f_- and f_+ are smooth functions on $[0, \infty)$, and $f_-(0) = f_+(0)$. The results of Sec. IV.B suggest that $f_-(x) \approx e^{-Ax}$ and $f_+(x) \approx e^{-Bx^{(d-1)/d}}$ for $x \gg 1$. This ansatz indicates that the role of $s_\xi \propto |p-p_c|^{-1/\sigma}$ as a cutoff is the same as in a Cayley tree. The general form (37) contains as a special case the Cayley tree (36) with $\tau = 5/2$, $\sigma = 1/2$, and $f_\pm(x) = e^{-x}$.

Another element of the scaling hypothesis is that the correlation length diverges near the percolation threshold following a power law:

$$\xi(p) \sim |p-p_c|^{-\nu} \quad \text{as } p \rightarrow p_c. \quad (38)$$

This ansatz introduces the correlation exponent ν and indicates that ξ and s_ξ are related by a power law $s_\xi = \xi^{1/\sigma\nu}$. From these two hypotheses we find that the percolation probability (22) is given by

$$P \sim (p-p_c)^\beta \quad \text{with } \beta = \frac{\tau-2}{\sigma}, \quad (39)$$

which scales as a positive power of $p-p_c$ for $p \geq p_c$; thus it is 0 for $p = p_c$ and increases when $p > p_c$. The average size of finite clusters, $\langle s \rangle^f$, which can be calculated on both sides of the percolation threshold, obeys

$$\langle s \rangle^f \sim |p-p_c|^{-\gamma} \quad \text{with } \gamma = \frac{3-\tau}{\sigma}, \quad (40)$$

diverging for $p \rightarrow p_c$. The exponents β and γ are called the critical exponents of the percolation probability and average cluster size, respectively.

E. Cluster structure

Until now we have discussed cluster sizes and radii, ignoring their internal structure. Let us now consider the *perimeter* of a cluster t denoting the number of nodes situated on the most external edges (the leaf nodes). The perimeter t_s of a very large but finite cluster of size s scales as (Leath, 1976)

$$t_s = s \frac{1-p}{p} + A s^\zeta \quad \text{as } s \rightarrow \infty, \quad (41)$$

where $\zeta = 1$ for $p < p_c$ and $\zeta = 1 - 1/d$ for $p > p_c$. Thus below p_c the perimeter of a cluster is proportional to its volume, a highly irregular property, which is nevertheless true for trees, including the Cayley tree.

Another way of understanding the unusual structure of finite clusters is by looking at the relation between their radii and volume. The correlation length ξ is a measure of the mean cluster radius, and we know that ξ scales with the cutoff cluster size s_ξ as $\xi \propto s_\xi^{1/\nu\sigma}$. Thus finite clusters are fractals (see Mandelbrot, 1982) because their size does not scale as their radius to the d th power, but as

$$s(r) \sim r^{d_f}, \quad (42)$$

where $d_f = 1/\sigma\nu$. It can also be shown that at the percolation threshold an infinite cluster is still a fractal, but for $p > p_c$ it becomes a normal d -dimensional object.

While the cluster radii and the correlation length ξ are defined using Euclidian distances on the lattice, the *chemical distance* is defined as the length of the shortest path between two arbitrary sites on a cluster (Havlin and Nossal, 1984). Thus the chemical distance is the equivalent of the distance on random graphs. The number of nodes within chemical distance ℓ scales as

$$s(\ell) \sim \ell^{d_\ell}, \quad (43)$$

where d_ℓ is called the *graph dimension* of the cluster. While the fractal dimension d_f of the Euclidian distances has been related to the other critical exponents, no such relation has yet been found for the graph dimension d_ℓ .

F. Infinite-dimensional percolation

Percolation is known to have a critical dimension d_c , below which some exponents depend on d , but for any dimension above d_c the exponents are the same. While it is generally believed that the critical dimension of percolation is $d_c = 6$, the dimension independence of the critical exponents is proven rigorously only for $d \geq 19$ (see Hara and Slade, 1990). Thus for $d > d_c$ the results of infinite-dimensional percolation theory apply, which predict that

- $P \sim (p - p_c)$ as $p \rightarrow p_c$;
- $\langle s \rangle \sim (p_c - p)^{-1}$ as $p \rightarrow p_c$;
- $n_s \sim s^{-5/2} e^{-|p - p_c|^{1/2} s}$ as $p \rightarrow p_c$;
- $\xi \sim |p - p_c|^{-1/2}$ as $p \rightarrow p_c$.

Consequently the critical exponents of infinite-dimensional percolation are $\tau_\infty = 5/2$, $\sigma_\infty = 1/2$, and $\nu_\infty = 1/2$. The fractal dimension of an infinite cluster at the percolation threshold is $d_f = 4$, while the graph dimension is $d_g = 2$ (Bunde and Havlin, 1996). Thus the characteristic chemical distance on a finite cluster or infinite cluster at the percolation threshold scales with its size as

$$\ell \sim s^{2/d_f} = s^{1/2}. \quad (44)$$

G. Parallels between random-graph theory and percolation

In random-graph theory we study a graph of N nodes, each pair of nodes being connected with probability p . This corresponds to percolation in at most N dimensions, such that each two connected nodes are neighbors, and the edges between graph nodes are the edges in the percolation problem. Since random-graph theory investigates the $N \rightarrow \infty$ regime, it is analogous to infinite-dimensional percolation.

We have seen in Sec. IV.C that infinite-dimensional percolation is similar to percolation on a Cayley tree. The percolation threshold of a Cayley tree is $p_c = 1/(z - 1)$, where z is the coordination number of the tree. In a random graph of N nodes the coordination number is $N - 1$; thus the “percolation threshold,” denoting the connection probability at which a giant cluster appears, should be $p_c \approx 1/N$. Indeed, this is exactly the probability at which the phase transition leading to a giant component appears in random graphs, as Erdős and Rényi showed (see Sec. III.C).

Compare the predictions of random-graph theory and infinite-dimensional percolation, some of which reflect a complete analogy:

(1) For $p < p_c = 1/N$

- The probability of a giant cluster in a graph, and of an infinite cluster in percolation, is equal to 0.
- The clusters of a random graph are trees, while the clusters in percolation have a fractal structure and a perimeter proportional with their volume.
- The largest cluster in a random graph is a tree with $\ln(N)$ nodes, while in general for percolation $P_p(|C| = s) \sim e^{-s/s_\xi}$ [see Eq. (28) in Sec. IV.B], suggesting that the size of the largest cluster scales as $\ln(N)$.

(2) For $p = p_c = 1/N$

- A unique giant cluster or an infinite cluster appears.
- The size of the giant cluster is $N^{2/3}$; while for infinite-dimensional percolation $P_p(|C| = s) \sim s^{-3/2}$, thus the size of the largest cluster scales as $N^{2/3}$.

(3) For $p > p_c = 1/N$

- The size of the giant cluster is $(f(p_c N) - f(p N))N$, where f is an exponentially decreasing function with $f(1) = 1$. The size of the infinite cluster is $PN \propto (p - p_c)N$.

- The giant cluster has a complex structure containing cycles, while the infinite cluster is no longer fractal, but compact.

All these correspondences indicate that the phase transition in random graphs belongs in the same universality class as mean-field percolation. Numerical simulations of random graphs (see, for example, Christensen *et al.*, 1998) have confirmed that the critical exponents of the phase transition are equal to the critical exponents of infinite-dimensional percolation. The equivalence of these two theories is very important because it offers us different perspectives on the same problem. For example, it is often of interest to look at the cluster size distribution of a random network with a fixed number of nodes. This question is answered in a simpler way in percolation theory. However, random-graph theory answers questions of major importance for networks, such as the appearance of trees and cycles, which are largely ignored by percolation theory.

In some cases there is an apparent discrepancy between the predictions of random-graph theory and percolation theory. For example, percolation theory predicts that the chemical distance between two nodes in an infinite cluster scales as a power of the size of the cluster [see Eq. (44)]. However, random-graph theory predicts [Eq. (16)] that the diameter of an infinite cluster scales logarithmically with its size (see Chung and Lu, 2001). The origin of the apparent discrepancy is that these two predictions refer to different regimes. While Eq. (44) is valid only when the infinite cluster is barely formed [i.e., $p = p_c$ and $\langle k \rangle = 1$] and is still a fractal, the prediction of random-graph theory is valid only well beyond the percolation transition, when $\langle k \rangle \gg 1$. Consequently, by using these two limits we can address the evolution of the chemical distance in an infinite cluster (see Cohen *et al.*, 2001). Thus for a full characterization of random networks we need to be aware of both of these complementary approaches.

V. GENERALIZED RANDOM GRAPHS

In Sec. II we have seen that real networks differ from random graphs in that often their degree distribution follows a power law $P(k) \sim k^{-\gamma}$. Since power laws are free of a characteristic scale, these networks are called “scale-free networks” (Barabási and Albert, 1999; Barabási, Albert, and Jeong, 1999). As random graphs do not capture the scale-free character of real networks, we need a different model to describe these systems. One approach is to generalize random graphs by constructing a model that has the degree distribution as an input but is random in all other respects. In other words, the edges connect randomly selected nodes, with the constraint that the degree distribution is restricted to a power law. The theory of such semirandom graphs should answer similar questions to those asked by Erdős and Rényi and percolation theory (see Secs. III, IV): Is there a threshold at which a giant cluster appears? How do the size and topology of the clusters evolve? When does the

graph become connected? In addition, we need to determine the average path length and clustering coefficient of such graphs.

The first step in developing such a theory is to identify the relevant parameter that, together with the network size, gives a statistically complete characterization of the network. In the case of random graphs this parameter is the connection probability (see Sec. III.A); for percolation theory it is the bond occupation probability (see Sec. IV). Since the only restriction for these graphs is that their degree distribution follow a power law, the exponent γ of the degree distribution could play the role of the control parameter. Accordingly, we study scale-free random networks by systematically varying γ and see if there is a threshold value of γ at which the networks' important properties abruptly change.

We start by sketching a few intuitive expectations. Consider a large network with degree distribution $P(k) \sim k^{-\gamma}$, in which γ decreases from ∞ to 0. The average degree of the network, or equivalently, the number of edges, increases as γ decreases, since $\langle k \rangle \sim k_{\max}^{-\gamma+2}$, where $k_{\max} < N$ is the maximum degree of the graph. This is very similar to the graph evolution process described by Erdős and Rényi (see Sec. III.C). Consequently we expect that, while at large γ the network consists of isolated small clusters, there is a critical value of γ at which a giant cluster forms, and at an even smaller γ the network becomes completely connected.

The theory of random graphs with given degree sequence is relatively recent. One of the first results is due to Luczak (1992), who showed that almost all random graphs with a fixed degree distribution and no nodes of degree smaller than 2 have a unique giant cluster. Molloy and Reed (1995, 1998) have proven that for a random graph with degree distribution $P(k)$ an infinite cluster emerges almost surely when

$$Q \equiv \sum_{k \geq 1} k(k-2)P(k) > 0, \quad (45)$$

provided that the maximum degree is less than $N^{1/4}$. The method of Molloy and Reed was applied to random graphs with power-law degree distributions by Aiello, Chung, and Lu (2000). As we show next, their results are in excellent agreement with the expectations outlined above.

A. Thresholds in a scale-free random graph

Aiello, Chung, and Lu (2000) introduce a two-parameter random-graph model $P(\alpha, \gamma)$ defined as follows: Let N_k be the number of nodes with degree k . $P(\alpha, \gamma)$ assigns uniform probability to all graphs with $N_k = e^{\alpha} k^{-\gamma}$. Thus in this model it is not the total number of nodes that is specified—along with the exponent γ —from the beginning, but the number of nodes with degree 1. Nevertheless the number of nodes and edges in the graph can be deduced, noting that the maximum degree of the graph is $e^{\alpha/\gamma}$. To find the condition for the appearance of a giant cluster in this model, we insert

$P(\alpha, \gamma)$ into Eq. (45), finding as a solution $\gamma_0 = 3.47875 \dots$. Thus when $\gamma > \gamma_0$ the random graph almost surely has no infinite cluster. On the other hand, when $\gamma < \gamma_0$ there is almost surely a unique infinite cluster.

An important question is whether the graph is connected or not. Certainly for $\gamma > \gamma_0$ the graph is disconnected as it is made of independent finite clusters. In the $0 < \gamma < \gamma_0$ regime Aiello, Chung, and Lu (2000) study the size of the second-largest cluster, finding that for $2 \leq \gamma \leq \gamma_0$ the second-largest cluster almost surely has a size of the order of $\ln N$; thus it is relatively small. However, for $1 < \gamma < 2$ almost surely every node with degree greater than $\ln(N)$ belongs to the infinite cluster. The second-largest cluster has a size of order 1, i.e., its size does not increase as the size of the graph goes to infinity. This means that the fraction of nodes in the infinite cluster approaches 1 as the system size increases; thus the graph becomes totally connected in the limit of infinite system size. Finally, for $0 < \gamma < 1$ the graph is almost surely connected.

B. Generating function formalism

A general approach to random graphs with given degree distribution was developed by Newman, Strogatz, and Watts (2001) using a generating function formalism (Wilf, 1990). The generating function of the degree distribution,

$$G_0(x) = \sum_{k=0}^{\infty} P(k)x^k, \quad (46)$$

encapsulates all the information contained in $P(k)$, since

$$P(k) = \frac{1}{k!} \left. \frac{d^k G_0}{dx^k} \right|_{x=0}. \quad (47)$$

An important quantity for studying cluster structure is the generating function for the degree distribution of the nearest neighbors of a randomly selected node. This can be obtained in the following way: a randomly selected edge reaches a node with degree k with probability proportional to $kP(k)$ (i.e., it is easier to find a well-connected node). If we start from a randomly chosen node and follow each of the edges starting from it, then the nodes we visit have their degree distribution generated by $kP(k)$. In addition, the generating function will contain a term x^{k-1} [instead of x^k as in Eq. (46)] because we have to discount the edge through which we reached the node. Thus the distribution of outgoing edges is generated by the function

$$G_1(x) = \frac{\sum_k kP(k)x^{k-1}}{\sum_k kP(k)} = \frac{1}{\langle k \rangle} G'_0(x). \quad (48)$$

The average number of first neighbors is equal to the average degree of the graph,

$$z_1 = \langle k \rangle = \sum_k kP(k) = G'_0(1). \quad (49)$$

1. Component sizes and phase transitions

When we identify a cluster using a burning (breadth-first-search) algorithm, we start from an arbitrary node and follow its edges until we reach its nearest neighbors. We record these nodes as part of the cluster, then follow their outside edges (avoiding the already recorded nodes) and record the nodes we arrive at as next-nearest neighbors of the starting node. This process is repeated until no new nodes are found, the set of identified nodes forming an isolated cluster. This algorithm is implicitly incorporated into the generating function method. The generating function, $H_1(x)$, for the size distribution of the clusters reached by following a random edge satisfies the iterative equation

$$H_1(x) = \frac{\sum_k kP(k)[H_1(x)]^k}{\sum_k kP(k)} = xG_1[H_1(x)]. \quad (50)$$

Here $kP(k)$ is proportional to the probability that a random edge arrives at a node with degree k , and $[H_1(x)]^k$ represents the k ways in which the cluster can be continued recursively (i.e., by finding the nearest neighbors of a previously found node). If we start at a randomly chosen node then we have one such cluster at the end of each edge leaving that node, and hence the generating function for the size of the whole cluster is

$$H_0(x) = x \sum_k P(k)[H_1(x)]^k = xG_0[H_1(x)]. \quad (51)$$

When there is no giant cluster present in the graph, the average cluster size is given by

$$\langle s \rangle = H'_0(1) = 1 + \frac{G'_0(1)}{1 - G'_1(1)}. \quad (52)$$

This expression diverges when $G'_1(1) = 1$, indicating the appearance of a giant cluster. Substituting the definition of $G_0(x)$ we can write the condition of the emergence of the giant cluster as

$$\sum_k k(k-2)P(k) = 0, \quad (53)$$

identical to Eq. (45) derived by Molloy and Reed (1995). Equation (53) gives an implicit relation for the critical degree distribution of a random graph: For any degree distribution for which the sum on the left-hand side is negative, no giant cluster is present in the graph, but degree distributions that give a positive sum lead to the appearance of a giant cluster.

When a giant cluster is present, $H_0(x)$ generates the probability distribution of the finite clusters. This means that $H_0(1)$ is no longer unity but instead takes on the value $1 - S$, where S is the fraction of nodes in the giant cluster. We can use this to calculate the size of the giant cluster S as (Molloy and Reed, 1998)

$$S = 1 - G_0(u), \quad (54)$$

where u is the smallest non-negative real solution of the equation $u = G_1(u)$.

Since we are dealing with random graphs (although with an arbitrary degree distribution), percolation theory (see Sec. IV) indicates that close to the phase transition the tail of cluster size distribution, n_s , behaves as

$$n_s \sim s^{-\tau} e^{-s/s_\xi}. \quad (55)$$

The characteristic cluster size s_ξ can be related to the first singularity of $H_0(x)$, x^* , and at the phase transition $x^* = 1$ and $s_\xi \rightarrow \infty$. Using a Taylor expansion around the critical point, we find that $H_0(x)$ scales as

$$H_0(x) \sim (1-x)^\alpha \quad \text{as } x \rightarrow 1, \quad (56)$$

with $\alpha = \frac{1}{2}$. This exponent can be related to the exponent τ by using the connection between n_s and $H_0(x)$, obtaining $\tau = \alpha + 2 = \frac{5}{2}$, regardless of degree distribution. Thus close to the critical point the cluster size distribution follows $n_s^c \sim s^{-5/2}$, as predicted by infinite-dimensional percolation (Sec. IV.F), but now extended to a large family of random graphs with arbitrary degree distribution.

2. Average path length

Extending the method of calculating the average number of nearest neighbors, we find the average number of m th neighbors,

$$z_m = [G'_1(1)]^{m-1} G'_0(1) = \left[\frac{z_2}{z_1} \right]^{m-1} z_1, \quad (57)$$

where z_1 and z_2 are the numbers of nearest and next-nearest neighbors. Using this expression, we can derive an approximative relation for the average path length of the graph. Let us start from a given node and find the number of its nearest, next-nearest, \dots , m th neighbors. Assuming that all nodes in the graph can be reached within l steps, we have

$$1 + \sum_{m=1}^l n(m) = N, \quad (58)$$

where $n(m)$ is the number of m th neighbors of the initial node. To estimate the average path length, we can replace $n(m)$ with z_m , obtaining

$$1 + \sum_{m=1}^l z_m = N. \quad (59)$$

As for most graphs $N \gg z_1$ and $z_2 \gg z_1$, we obtain

$$l = \frac{\ln(N/z_1)}{\ln(z_2/z_1)} + 1. \quad (60)$$

A more rigorous method exists in the case of connected tree graphs (Ambjorn, Durhuus, and Jonsson, 1990; Burda, Correia, and Krzywicki, 2001), yielding that the average pathlength of connected trees with power-law degree distribution scales as $N^{(\gamma-2)/(\gamma-1)}$, where γ is the degree exponent. Although this scaling has a different functional form, for γ approaching 2 the dependence on the system size becomes very weak and practically indistinguishable from a logarithmic dependence.

C. Random graphs with power-law degree distribution

As an application of the generating function formalism, Newman, Strogatz, and Watts (2001) consider the case of a degree distribution of type

$$P(k) = Ck^{-\gamma}e^{-k/\kappa} \quad \text{for } k \geq 1, \quad (61)$$

where C , γ , and κ are constants. The exponential cutoff, present in some social and biological networks (see Amaral *et al.*, 2000; Jeong, Mason, *et al.*, 2001; Newman 2001a), has the technical advantage of making the distribution normalizable for all γ , not just $\gamma \geq 2$, as in the case of a pure power law. The constant C is fixed by normalization, giving $C = [Li_\gamma(e^{-1/\kappa})]^{-1}$, where $Li_n(x)$ is the n th polylogarithm of x . Thus the degree distribution is characterized by two independent parameters, the exponent γ and the cutoff κ . Following the formalism described above, we find that the size of an infinite cluster is

$$S = 1 - \frac{Li_\gamma(ue^{-1/\kappa})}{Li_\gamma(e^{-1/\kappa})}, \quad (62)$$

where u is the smallest non-negative real solution of the equation $u = Li_{\gamma-1}(ue^{-1/\kappa})/[uLi_{\gamma-1}(e^{-1/\kappa})]$. For graphs with purely power-law distribution ($\kappa \rightarrow \infty$), the above equation becomes $u = Li_{\gamma-1}(u)/[u\zeta(\gamma-1)]$, where $\zeta(x)$ is the Riemann ζ function. For all $\gamma \leq 2$ this gives $u=0$, and hence $S=1$, implying that a randomly chosen node belongs to the giant cluster with probability converging to 1 as $\kappa \rightarrow \infty$. For graphs with $\gamma > 2$ this is never the case, even for infinite κ , indicating that such a graph contains finite clusters, i.e., it is not connected, in agreement with the conclusions of Aiello, Chung, and Lu (2000).

The average path length is

$$\ell = \frac{\ln N + \ln[Li_\gamma(e^{-1/\kappa})/Li_{\gamma-1}(e^{-1/\kappa})]}{\ln[Li_{\gamma-2}(e^{-1/\kappa})/Li_{\gamma-1}(e^{-1/\kappa}) - 1]} + 1, \quad (63)$$

which in the limit $\kappa \rightarrow \infty$ becomes

$$\ell = \frac{\ln N + \ln[\zeta(\gamma)/\zeta(\gamma-1)]}{\ln[\zeta(\gamma-2)/\zeta(\gamma-1) - 1]} + 1. \quad (64)$$

Note that this expression does not have a finite positive real value for any $\gamma < 3$, indicating that one must specify a finite cutoff κ for the degree distribution to get a well-defined average path length. Equations (60) and (63) reproduce the result of finite size scaling simulations of the World Wide Web, indicating that its average path length scales logarithmically with its size (Albert, Jeong, and Barabási, 1999). But do they offer a good estimate for the average path lengths of real networks? In Sec. II we saw that the prediction of random-graph theory is in qualitative agreement with the average path lengths of real networks, but that there also are significant deviations from it. It is thus important to see if taking into account the correct degree distribution gives a better fit.

In Fig. 13 we compare the prediction of Eq. (63) with the average path length of a real network by plotting $A(\ell-1)-B$ as a function of the network size

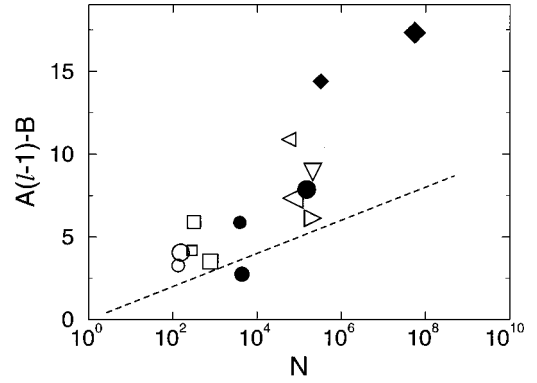


FIG. 13. Comparison between the average path lengths of real scale-free networks and the prediction (63) of scale-free random graphs (dashed line). For each network we have plotted $A(\ell-1)-B$ as a function of N , where A and B are given in the text. The networks included in the figure, indicated by their number in Table I or Table II, are small \circ , I.12; large \circ , I.13; small \square , I.10; medium \square , I.11; large \square , II.13; small \bullet , II.6; medium \bullet , I.2; small \triangle , I.6; large \triangle , I.8; large \bullet , II.7; ∇ , I.9; \blacklozenge , I.3; medium \blacklozenge , II.1; large \blacklozenge , II.3.

N , where $A = \log[Li_{\gamma-2}(e^{-1/\kappa})/Li_{\gamma-1}(e^{-1/\kappa}) - 1]$ and $B = \log[Li_\gamma(e^{-1/\kappa})/Li_{\gamma-1}(e^{-1/\kappa})]$, and we use the cutoff length κ as obtained from the empirical degree distributions. For directed networks we used the γ_{out} values. For random networks with the same N , γ , and κ as the real networks, the $A\ell-B$ values would align along a straight line with slope 1 in a log-linear plot, given by the dashed line on the figure. The actual values for real networks obey the trend, but they seem to be systematically larger than the prediction of Eq. (63), indicating that the average path lengths of real networks are larger than those of random graphs with power-law degree distribution. This conclusion is further supported by the last three columns of Table II, which directly compare the average path lengths of real networks with power-law degree distribution ℓ_{real} , with the estimates of random-graph theory ℓ_{rand} , and with scale-free random-graph theory ℓ_{pow} . We can see that the general trend is for ℓ_{real} to be larger than both ℓ_{pow} and ℓ_{rand} , an indication of the nonrandom aspects of the topology of real networks.

D. Bipartite graphs and the clustering coefficient

The clustering coefficient of a scale-free random graph has not yet been calculated in the literature, but we can get some idea of its general characteristics if we take into account that scale-free random graphs are similar to Erdős-Rényi random graphs in the sense that their edges are distributed randomly. Consequently the clustering coefficients of scale-free random graphs converge to 0 as the network size increases.

It is worth noting, however, that some of the real-world networks presented in Sec. II, for example, the collaboration networks, can be more completely described by bipartite graphs (Newman, Strogatz, and Watts, 2001). In a bipartite graph there are two kinds of

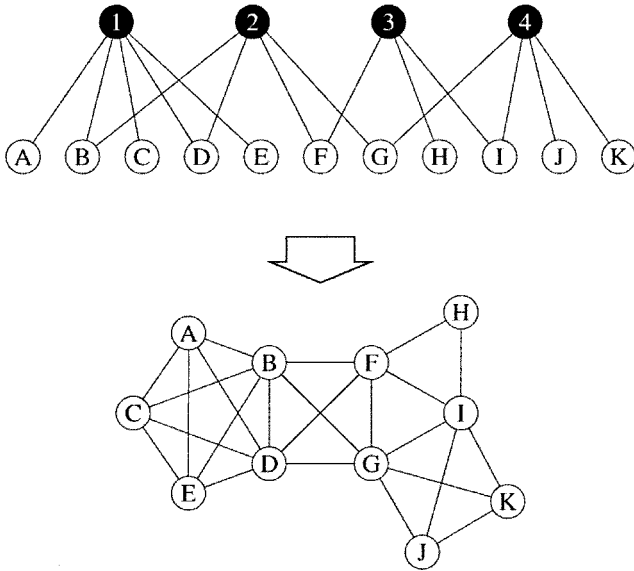


FIG. 14. A schematic representation of a bipartite graph, such as the graph of movies and the actors who have appeared in them. In this small graph we have four movies, labeled 1 to 4, and eleven actors, labeled A to K, with edges joining each movie to the actors in its cast. The bottom figure shows the one-mode projection of the graph for the eleven actors. After Newman, Strogatz, and Watts (2001).

nodes, and edges connect only nodes of different kinds. For example, the collaboration network of movie actors is in fact a projection of a bipartite actor-movie graph, in which the two types of nodes are the actors and movies, and an edge connects each movie with the actors playing in it (see Fig. 14). The same approach is applicable to the collaborations between scientists (where scientists and papers are the two types of nodes) and metabolic networks (where nodes can be the substrates or reactions). The generating function method can be generalized to bipartite graphs (see Newman, Strogatz, and Watts, 2001), and it results in a nonvanishing clustering coefficient inherent to the bipartite structure,

$$C = \frac{1}{1 + \frac{(\mu_2 - \mu_1)(\nu_2 - \nu_1)^2}{\mu_1 \nu_1 (2\nu_1 - 3\nu_2 + \nu_3)}}, \quad (65)$$

where $\mu_n = \sum_k k^n P_a(k)$ and $\nu_n = \sum_k k^n P_m(k)$. In the actor-movie framework, $P_a(k)$ represents the fraction of actors who appeared in k movies, while $P_m(k)$ means the fraction of movies in which k actors have appeared.

The prediction of Eq. (65) has been tested for several collaboration graphs (Newman, Strogatz, and Watts, 2001). In some cases there is excellent agreement, but in others it deviates by a factor of 2 from the clustering coefficient of the real network. Consequently we can conclude that the order present in real networks is not due solely to the definition of the network, but an as yet unknown organizing principle.

VI. SMALL-WORLD NETWORKS

In Secs. II and III.A we saw (Table I, Figs. 8 and 9) that real-world networks have a small-world character

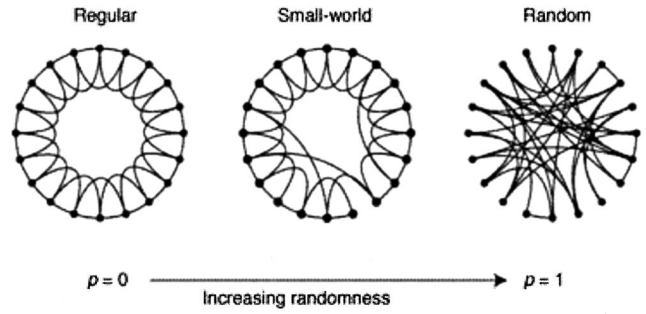


FIG. 15. The random rewiring procedure of the Watts-Strogatz model, which interpolates between a regular ring lattice and a random network without altering the number of nodes or edges. We start with $N=20$ nodes, each connected to its four nearest neighbors. For $p=0$ the original ring is unchanged; as p increases the network becomes increasingly disordered until for $p=1$ all edges are rewired randomly. After Watts and Strogatz, 1998.

like random graphs, but they have unusually large clustering coefficients. Furthermore, as Fig. 9 demonstrates, the clustering coefficient appears to be independent of the network size. This latter property is characteristic of ordered lattices, whose clustering coefficient is size independent and depends only on the coordination number. For example, in a one-dimensional lattice with periodic boundary conditions (i.e., a ring of nodes), in which each node is connected to the K nodes closest to it (see Fig. 15), most of the immediate neighbors of any site are also neighbors of one another, i.e., the lattice is clustered. For such a lattice the clustering coefficient is

$$C = \frac{3(K-2)}{4(K-1)}, \quad (66)$$

which converges to $3/4$ in the limit of large K . Such low-dimensional regular lattices, however, do not have short path lengths: for a d -dimensional hypercubic lattice the average node-node distance scales as $N^{1/d}$, which increases much faster with N than the logarithmic increase observed for random and real graphs. The first successful attempt to generate graphs with high clustering coefficients and small ℓ is that of Watts and Strogatz (1998).

A. The Watts-Strogatz model

Watts and Strogatz (1998) proposed a one-parameter model that interpolates between an ordered finite-dimensional lattice and a random graph. The algorithm behind the model is the following (Fig. 15):

(1) *Start with order:* Start with a ring lattice with N nodes in which every node is connected to its first K neighbors ($K/2$ on either side). In order to have a sparse but connected network at all times, consider $N \gg K \gg \ln(N) \gg 1$.

(2) *Randomize:* Randomly rewire each edge of the lattice with probability p such that self-connections and duplicate edges are excluded. This process introduces $pNK/2$ long-range edges which connect nodes that oth-

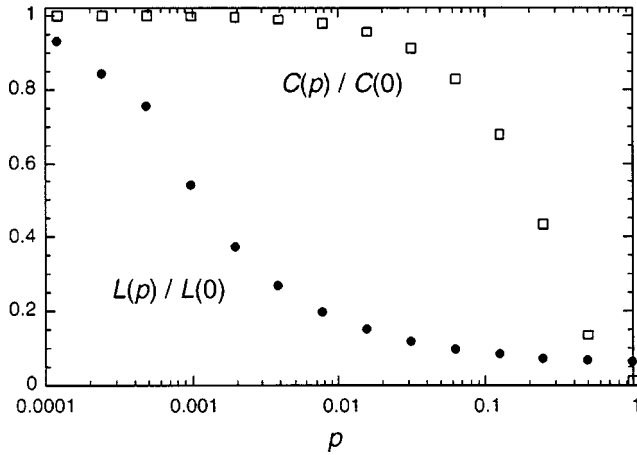


FIG. 16. Characteristic path length $\ell(p)$ and clustering coefficient $C(p)$ for the Watts-Strogatz model. The data are normalized by the values $\ell(0)$ and $C(0)$ for a regular lattice. A logarithmic horizontal scale resolves the rapid drop in $\ell(p)$, corresponding to the onset of the small-world phenomenon. During this drop $C(p)$ remains almost constant, indicating that the transition to a small world is almost undetectable at the local level. After Watts and Strogatz, 1998.

erwise would be part of different neighborhoods. By varying p one can closely monitor the transition between order ($p=0$) and randomness ($p=1$).

This model has its roots in social systems in which most people are friends with their immediate neighbors—neighbors on the same street, colleagues, people their friends introduce them to. However, everybody has one or two friends who are a long way away—people in other countries, old acquaintances—who are represented by the long-range edges obtained by rewiring in the Watts-Strogatz model.

To understand the coexistence of small path length and clustering, we study the behavior of the clustering coefficient $C(p)$ and the average path length $\ell(p)$ as a function of the rewiring probability p . For a ring lattice $\ell(0) \approx N/2K \gg 1$ and $C(0) \approx 3/4$; thus ℓ scales linearly with the system size, and the clustering coefficient is large. On the other hand, for $p \rightarrow 1$ the model converges to a random graph for which $\ell(1) \sim \ln(N)/\ln(K)$; and $C(1) \sim K/N$; thus ℓ scales logarithmically with N , and the clustering coefficient decreases with N . These limiting cases might suggest that large C is always associated with large ℓ , and small C with small ℓ . On the contrary, Watts and Strogatz (1998) found that there is a broad interval of p over which $\ell(p)$ is close to $\ell(1)$ yet $C(p) \gg C(1)$ (Fig. 16). This regime originates in a rapid drop of $\ell(p)$ for small values of p , while $C(p)$ stays almost unchanged, resulting in networks that are clustered but have a small characteristic path length. This coexistence of small ℓ and large C is in excellent agreement with the characteristics of real networks discussed in Sec. II, prompting many to call such systems small-world networks.

B. Properties of small-world networks

The pioneering article of Watts and Strogatz started an avalanche of research on the properties of small-

world networks and the Watts-Strogatz (WS) model. A much-studied variant of the WS model was proposed by Newman and Watts (1999a, 1999b), in which edges are added between randomly chosen pairs of sites, but no edges are removed from the regular lattice. This model is somewhat easier to analyze than the original Watts-Strogatz model because it does not lead to the formation of isolated clusters, whereas this can happen in the original model. For sufficiently small p and large N this model is equivalent to the WS model. In the following we shall summarize the main results regarding the properties of small-world models.

1. Average path length

As we discussed above, in the Watts-Strogatz model there is a change in the scaling of the characteristic path length ℓ as the fraction p of the rewired edges is increased. For small p , ℓ scales linearly with the system size, while for large p the scaling is logarithmic. As discussed by Watts (1999) and Pandit and Amritkar (1999), the origin of the rapid drop in ℓ is the appearance of shortcuts between nodes. Every shortcut, created at random, is likely to connect widely separated parts of the graph, and thus has a significant impact on the characteristic path length of the entire graph. Even a relatively low fraction of shortcuts is sufficient to drastically decrease the average path length, yet locally the network remains highly ordered.

An important question regarding the average path length is whether the onset of small-world behavior is dependent on the system size. It was Watts (1999) who first noticed that ℓ does not begin to decrease until $p \geq 2/NK$, guaranteeing the existence of at least one shortcut. This implies that the transition p depends on the system size, or conversely, there exists a p -dependent crossover length (size) N^* such that if $N < N^*$, $\ell \sim N$, but if $N > N^*$, $\ell \sim \ln(N)$. The concept of the crossover size was introduced by Barthélemy and Amaral (1999), who conjectured that the characteristic path length scales as (see Fig. 17)

$$\ell(N, p) \sim N^* F\left(\frac{N}{N^*}\right), \quad (67)$$

where

$$F(u) = \begin{cases} u & \text{if } u \leq 1 \\ \ln(u) & \text{if } u \geq 1. \end{cases} \quad (68)$$

Numerical simulations and analytical arguments (Barat 1999; Barthélemy and Amaral, 1999; Newman and Watts, 1999a; Argollo de Menezes *et al.*, 2000; Barrat and Weigt, 2000) concluded that the crossover size N^* scales with p as $N^* \sim p^{-\tau}$, where $\tau = 1/d$ and d is the dimension of the original lattice to which the random edges are added (Fig. 18). Thus for the original WS model, defined on a circle ($d=1$), we have $\tau=1$, the onset of small-world behavior taking place at the rewiring probability $p^* \sim 1/N$.

It is now widely accepted that the characteristic path length obeys the general scaling form

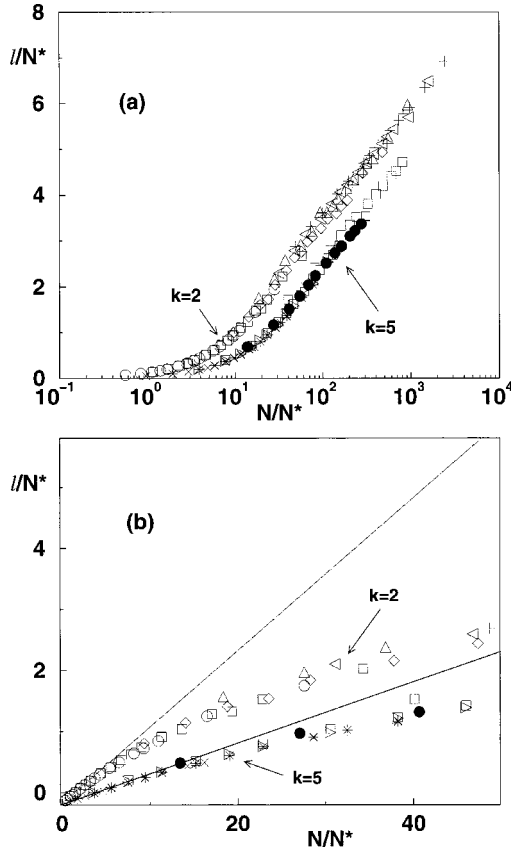


FIG. 17. Data collapse $l(N,p)/N^*(p)$ versus $N/N^*(p)$ for two different values of K : (a) log-linear scale showing the logarithmic behavior at large N/N^* ; (b) linear scale showing the linear behavior $l(N,p) \sim N/(4K)$ at small N/N^* . After Barrat and Weigt (2000).

$$l(N,p) \sim \frac{N^{1/d}}{K} f(pKN), \quad (69)$$

where $f(u)$ is a universal scaling function that obeys

$$f(u) = \begin{cases} \text{const} & \text{if } u \ll 1 \\ \ln(u)/u & \text{if } u \gg 1. \end{cases} \quad (70)$$

Newman, Moore, and Watts (2000) have calculated the form of the scaling function $f(u)$ for the one-dimensional small-world model using a mean-field method that is exact for small or large values of u , but not in the regime in which $u \approx 1$, obtaining

$$f(u) = \frac{4}{\sqrt{u^2 + 4u}} \tanh^{-1} \frac{u}{\sqrt{u^2 + 4u}}. \quad (71)$$

They also solved for the complete distribution of path lengths within this mean-field approximation.

The scaling relation (69) has been confirmed by extensive numerical simulations (Newman and Watts, 1999a; Argollo de Menezes *et al.*, 2000), renormalization-group techniques (Newman and Watts, 1999a), and series expansions (Newman and Watts, 1999b). Equation (69) tells us that although the average path length in a small-world model appears at first glance to depend on three parameters— p , K , and N —it is in fact entirely deter-

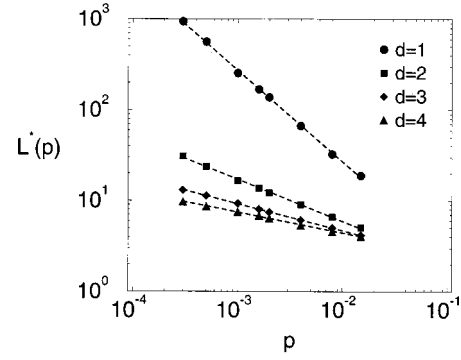


FIG. 18. The dependence of the crossover size N^* on the rewiring probability in one to four dimensions. The dashed lines represent the scaling relation $N^* \sim p^{-1/d}$. After Argollo de Menezes *et al.* (2000).

mined by a single scalar function $f(u)$ of a single scalar variable. Note that both the scaling function $f(u)$ and the scaling variable $u = pKN^d$ have simple physical interpretations. The variable u is two times the average number of random links (shortcuts) on the graph for a given p , and $f(u)$ is the average of the fraction by which the distance between two nodes is reduced for a given u .

Several attempts have been made to calculate exactly the distribution of path lengths and the average path length ℓ . Dorogovtsev and Mendes (2000a) studied a simpler model that contains a ring lattice with directed edges of length 1 and a central node that is connected with probability p to the nodes of the lattice by undirected edges of length 0.5. They calculated exactly the distribution of path lengths for this model, showing that ℓ/N depends only on the scaling variable pN , and the functional form of this dependence is similar to the numerically obtained $\ell(p)$ in the WS model. Kulkarni *et al.* (1999) calculated the probability $P(m|n)$ that two nodes separated by a Euclidian distance n have a path length m . They have shown that the average path length ℓ is simply related to the mean $\langle s \rangle$ and the mean square $\langle s^2 \rangle$ of the shortest distance between two diametrically opposite nodes (i.e., separated by the largest Euclidian distance), according to

$$\frac{\ell}{N} = \frac{\langle s \rangle}{N-1} - \frac{\langle s^2 \rangle}{L(N-1)}. \quad (72)$$

Unfortunately calculating the shortest distance between opposite nodes is just as difficult as determining ℓ directly.

2. Clustering coefficient

In addition to a short average path length, small-world networks have a relatively high clustering coefficient. The WS model displays this duality for a wide range of the rewiring probabilities p . In a regular lattice ($p=0$) the clustering coefficient does not depend on the size of the lattice but only on its topology. As the edges of the network are randomized, the clustering coefficient remains close to $C(0)$ up to relatively large values of p .

The dependence of $C(p)$ on p can be derived using a slightly different but equivalent definition of C , introduced by Barrat and Weigt (2000). According to this definition, $C'(p)$ is the fraction between the mean number of edges between the neighbors of a node and the mean number of possible edges between those neighbors. In a more graphic formulation (Newman, Strogatz, and Watts, 2001),

$$C' = \frac{3 \times \text{number of triangles}}{\text{number of connected triples}}. \quad (73)$$

Here triangles are trios of nodes in which each node is connected to both of the others, and connected triples are trios in which at least one is connected to both others, the factor 3 accounting for the fact that each triangle contributes to three connected triples. This definition corresponds to the concept of the “fraction of transitive triples” used in sociology (see Wasserman and Faust, 1994).

To calculate $C'(p)$ for the WS model, let us start with a regular lattice with a clustering coefficient $C(0)$. For $p > 0$, two neighbors of a node i that were connected at $p = 0$ are still neighbors of i and connected by an edge with probability $(1-p)^3$, since there are three edges that need to remain intact. Consequently $C'(p) \approx C(0)(1-p)^3$. Barrat and Weigt (2000) have verified that the deviation of $C(p)$ from this expression is small and goes to zero as $N \rightarrow \infty$. The corresponding expression for the Newman-Watts model is (Newman, 2001e)

$$C'(p) = \frac{3K(K-1)}{2K(2K-1) + 8pK^2 + 4p^2K^2}. \quad (74)$$

3. Degree distribution

In the WS model for $p = 0$ each node has the same degree K . Thus the degree distribution is a delta function centered at K . A nonzero p introduces disorder in the network, broadening the degree distribution while maintaining the average degree equal to K . Since only a single end of every edge is rewired ($pNK/2$ edges in total), each node has at least $K/2$ edges after the rewiring process. Consequently for $K > 2$ there are no isolated nodes and the network is usually connected, unlike a random graph which consists of isolated clusters for a wide range of connection probabilities.

For $p > 0$, the degree k_i of a vertex i can be written as (Barrat and Weigt, 2000) $k_i = K/2 + c_i$, where c_i can be divided into two parts: $c_i^1 \leq K/2$ edges have been left in place (with probability $1-p$), while $c_i^2 = c_i - c_i^1$ edges have been rewired towards i , each with probability $1/N$. The probability distributions of c_i^1 and c_i^2 are

$$P_1(c_i^1) = C_{K/2}^{c_i^1} (1-p)^{c_i^1} p^{K/2 - c_i^1} \quad (75)$$

and

$$\begin{aligned} P_2(c_i^2) &= C_{pNK/2}^{c_i^2} \left(\frac{1}{N} \right)^{c_i^2} \left(1 - \frac{1}{N} \right)^{pNK/2 - c_i^2} \\ &\approx \frac{(pK/2)^{c_i^2}}{c_i^2!} e^{-pK/2} \end{aligned} \quad (76)$$

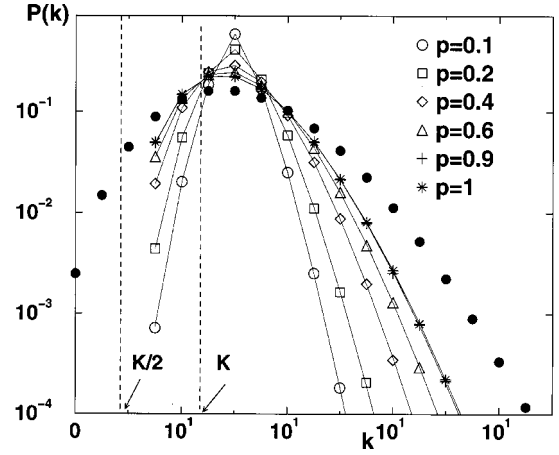


FIG. 19. Degree distribution of the Watts-Strogatz model for $K=3$ and various p . We can see that only $k \geq K/2$ values are present, and the mean degree is $\langle k \rangle = K$. The symbols are obtained from numerical simulations of the Watts-Strogatz model with $N=1000$, and the lines correspond to Eq. (77). As a comparison, the degree distribution of a random graph with the same parameters is plotted with filled symbols. After Barrat and Weigt (2000).

for large N . Combining these two factors, the degree distribution follows

$$P(k) = \sum_{n=0}^{f(k,K)} C_{K/2}^n (1-p)^n p^{K/2-n} \frac{(pK/2)^{k-K/2-n}}{(k-K/2-n)!} e^{-pK/2} \quad (77)$$

for $k \geq K/2$, where $f(k,K) = \min(k-K/2, K/2)$.

The shape of the degree distribution is similar to that of a random graph. It has a pronounced peak at $\langle k \rangle = K$ and decays exponentially for large k (Fig. 19). Thus the topology of the network is relatively homogeneous, all nodes having approximately the same number of edges.

4. Spectral properties

As discussed in Sec. III.G, the spectral density $\rho(\lambda)$ of a graph reveals important information about its topology. Specifically, we have seen that for large random graphs $\rho(\lambda)$ converges to a semicircle. It comes as no surprise that the spectrum of the Watts-Strogatz model depends on the rewiring probability p (Farkas *et al.*, 2001). For $p=0$ the network is regular and periodical; consequently $\rho(\lambda)$ contains numerous singularities [Fig. 20(a)]. For intermediate values of p these singularities become blurred, but $\rho(\lambda)$ retains a strong skewness [Figs. 20(b) and (c)]. Finally, as $p \rightarrow 1$, $\rho(\lambda)$ approaches the semicircle law characterizing random graphs [Fig. 20(d)]. While the details of the spectral density change considerably with p , the third moment of $\rho(\lambda)$ is consistently high, indicating a high number of triangles in the network. Thus the results summarized in Fig. 20 allow us to conclude that a high number of triangles is a basic property of the WS model (see also Gleis *et al.*, 2000). The high regularity of small-world models for a

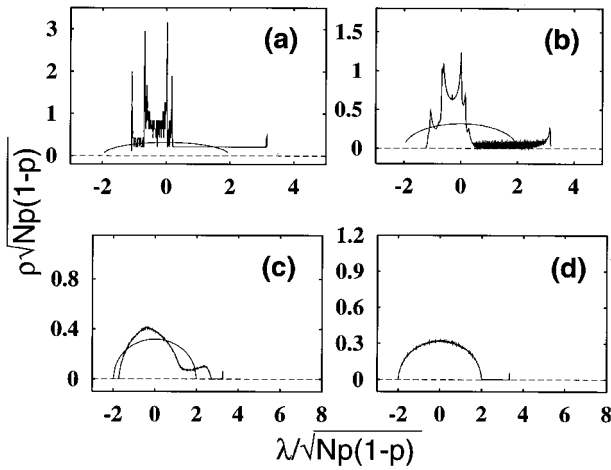


FIG. 20. Spectral density of small-world networks, compared to the semicircle law corresponding to random graphs (solid line). The rewiring probabilities are (a) $p=0$; (b) $p=0.01$; (c) $p=0.3$; and (d) $p=1$. After Farkas *et al.* (2001).

broad range of p is underlined by the results concerning the spectral properties of the Laplacian operator, which tell us about the time evolution of a diffusive field on the graph (Monasson, 2000).

VII. SCALE-FREE NETWORKS

The empirical results discussed in Sec. II demonstrate that many large networks are scale free, that is, their degree distribution follows a power law for large k . Furthermore, even for those networks for which $P(k)$ has an exponential tail, the degree distribution significantly deviates from a Poisson distribution. We have seen in Secs. III.D and VI.B.3 that random-graph theory and the WS model cannot reproduce this feature. While it is straightforward to construct random graphs that have a power-law degree distribution (Sec. V), these constructions only postpone an important question: what is the mechanism responsible for the emergence of scale-free networks? We shall see in this section that answering this question will require a shift from modeling network topology to modeling the network assembly and evolution. While at this point these two approaches do not appear to be particularly distinct, we shall find that there is a fundamental difference between the modeling approach we took in random graphs and the small-world models, and the one required to reproduce the power-law degree distribution. While the goal of the former models is to construct a graph with correct topological features, the modeling of scale-free networks will put the emphasis on capturing the network dynamics. That is, the underlying assumption behind evolving or dynamic networks is that if we capture correctly the processes that assembled the networks that we see today, then we will obtain their topology correctly as well. Dynamics takes the driving role, topology being only a by-product of this modeling philosophy.

A. The Barabási-Albert model

The origin of the power-law degree distribution observed in networks was first addressed by Barabási and Albert (1999), who argued that the scale-free nature of real networks is rooted in two generic mechanisms shared by many real networks. The network models discussed thus far assume that we start with a fixed number N of vertices that are then randomly connected or rewired, without modifying N . In contrast, most real-world networks describe open systems that *grow* by the continuous addition of new nodes. Starting from a small nucleus of nodes, the number of nodes increases throughout the lifetime of the network by the subsequent addition of new nodes. For example, the World Wide Web grows exponentially in time by the addition of new web pages, and the research literature constantly grows by the publication of new papers.

Second, network models discussed so far assume that the probability that two nodes are connected (or their connection is rewired) is independent of the nodes' degree, i.e., new edges are placed randomly. Most real networks, however, exhibit *preferential attachment*, such that the likelihood of connecting to a node depends on the node's degree. For example, a web page will more likely include hyperlinks to popular documents with already high degrees, because such highly connected documents are easy to find and thus well known, or a new manuscript is more likely to cite well-known and thus much-cited publications than less-cited and consequently less-known papers.

These two ingredients, growth and preferential attachment, inspired the introduction of the Barabási-Albert model, which led for the first time to a network with a power-law degree distribution. The algorithm of the Barabási-Albert model is the following:

(1) *Growth*: Starting with a small number (m_0) of nodes, at every time step, we add a new node with m ($\leq m_0$) edges that link the new node to m different nodes already present in the system.

(2) *Preferential attachment*: When choosing the nodes to which the new node connects, we assume that the probability Π that a new node will be connected to node i depends on the degree k_i of node i , such that

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}. \quad (78)$$

After t time steps this procedure results in a network with $N=t+m_0$ nodes and mt edges. Numerical simulations indicated that this network evolves into a scale-invariant state with the probability that a node has k edges following a power law with an exponent $\gamma_{BA}=3$ (see Fig. 21). The scaling exponent is independent of m , the only parameter in the model.

B. Theoretical approaches

The dynamical properties of the scale-free model can be addressed using various analytic approaches. The continuum theory proposed by Barabási and Albert

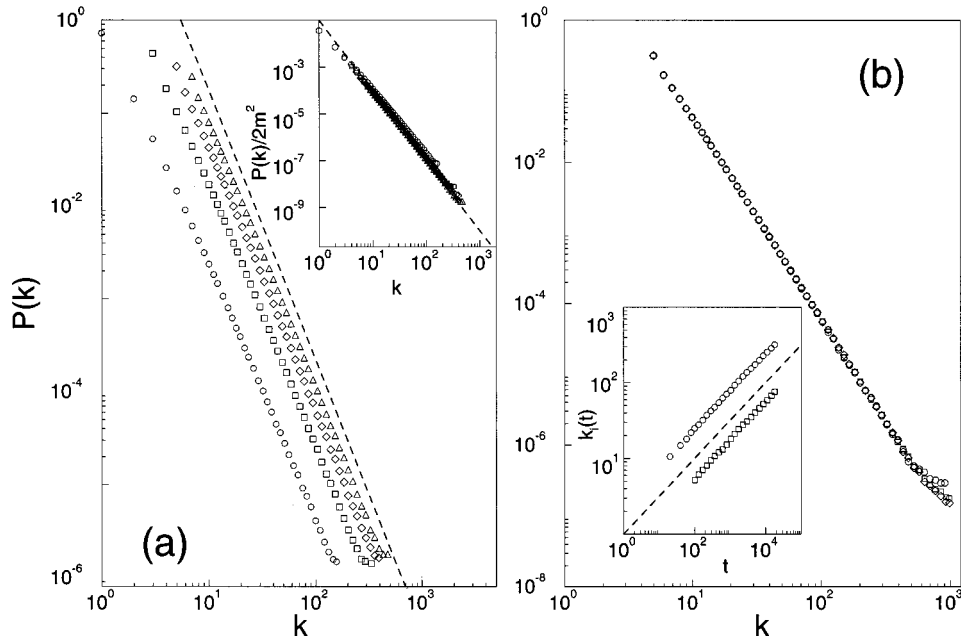


FIG. 21. Numerical simulations of network evolution: (a) Degree distribution of the Barabási-Albert model, with $N = m_0 + t = 300\,000$ and \circ , $m_0 = m = 1$; \square , $m_0 = m = 3$; \diamond , $m_0 = m = 5$; and \triangle , $m_0 = m = 7$. The slope of the dashed line is $\gamma = 2.9$, providing the best fit to the data. The inset shows the rescaled distribution (see text) $P(k)/2m^2$ for the same values of m , the slope of the dashed line being $\gamma = 3$; (b) $P(k)$ for $m_0 = m = 5$ and various system sizes, \circ , $N = 100\,000$; \square , $N = 150\,000$; \diamond , $N = 200\,000$. The inset shows the time evolution for the degree of two vertices, added to the system at $t_1 = 5$ and $t_2 = 95$. Here $m_0 = m = 5$, and the dashed line has slope 0.5, as predicted by Eq. (81). After Barabási, Albert, and Jeong (1999).

(1999) focuses on the dynamics of node degrees, followed by the master-equation approach of Dorogovtsev, Mendes, and Samukhin (2000a) and the rate-equation approach introduced by Krapivsky, Redner, and Leyvraz (2000). As these methods are often used interchangeably in the subsequent section, we briefly review each of them.

Continuum theory: The continuum approach introduced by Barabási and Albert (1999) and Barabási, Albert, and Jeong (1999) calculates the time dependence of the degree k_i of a given node i . This degree will increase every time a new node enters the system and links to node i , the probability of this process being $\Pi(k_i)$. Assuming that k_i is a continuous real variable, the rate at which k_i changes is expected to be proportional to $\Pi(k_i)$. Consequently k_i satisfies the dynamical equation

$$\frac{\partial k_i}{\partial t} = m \Pi(k_i) = m \frac{k_i}{\sum_{j=1}^{N-1} k_j}. \quad (79)$$

The sum in the denominator goes over all nodes in the system except the newly introduced one; thus its value is $\sum_j k_j = 2mt - m$, leading to

$$\frac{\partial k_i}{\partial t} = \frac{k_i}{2t}. \quad (80)$$

The solution of this equation, with the initial condition that every node i at its introduction has $k_i(t_i) = m$, is

$$k_i(t) = m \left(\frac{t}{t_i} \right)^\beta \quad \text{with} \quad \beta = \frac{1}{2}. \quad (81)$$

Equation (81) indicates that the degree of all nodes evolves the same way, following a power law, the only difference being the intercept of the power law.

Using Eq. (81), one can write the probability that a node has a degree $k_i(t)$ smaller than k , $P[k_i(t) < k]$, as

$$P[k_i(t) < k] = P\left(t_i > \frac{m^{1/\beta} t}{k^{1/\beta}}\right). \quad (82)$$

Assuming that we add the nodes at equal time intervals to the network, the t_i values have a constant probability density

$$P(t_i) = \frac{1}{m_0 + t}. \quad (83)$$

Substituting this into Eq. (82) we obtain

$$P\left(t_i > \frac{m^{1/\beta} t}{k^{1/\beta}}\right) = 1 - \frac{m^{1/\beta} t}{k^{1/\beta} (t + m_0)}. \quad (84)$$

The degree distribution $P(k)$ can be obtained using

$$P(k) = \frac{\partial P[k_i(t) < k]}{\partial k} = \frac{2m^{1/\beta} t}{m_0 + t} \frac{1}{k^{1/\beta+1}}, \quad (85)$$

predicting that asymptotically ($t \rightarrow \infty$)

$$P(k) \sim 2m^{1/\beta} k^{-\gamma} \quad \text{with} \quad \gamma = \frac{1}{\beta} + 1 = 3 \quad (86)$$

being independent of m , in agreement with the numerical results.

As the power law observed for real networks describes systems of rather different sizes, it is expected that a correct model should provide a time-independent degree distribution. Indeed, Eq. (85) predicts that asymptotically the degree distribution of the Barabási-Albert model is independent of time (and subsequently independent of the system size $N=m_0+t$), indicating that, despite its continuous growth, the network reaches a stationary scale-free state. Furthermore, Eq. (85) also indicates that the coefficient of the power-law distribution is proportional to m^2 . All these predictions are confirmed by numerical simulations (see Fig. 21).

Master-equation approach: The method introduced by Dorogovtsev, Mendes, and Samukhin (2000a; see also Kullmann and Kertész, 2001) studies the probability $p(k, t_i, t)$ that at time t a node i introduced at time t_i has a degree k . In the Barabási-Albert model, when a new node with m edges enters the system, the degree of node i increases by 1 with a probability $m\Pi(k)=k/2t$; otherwise it stays the same. Consequently the master equation governing $p(k, t_i, t)$ for the Barabási-Albert model has the form

$$p(k, t_i, t+1) = \frac{k-1}{2t} p(k-1, t_i, t) + \left(1 - \frac{k}{2t}\right) p(k, t_i, t). \quad (87)$$

The degree distribution can be obtained as

$$P(k) = \lim_{t \rightarrow \infty} \left(\sum_{t_i} p(k, t_i, t) \right) / t. \quad (88)$$

Equation (87) implies that $P(k)$ is the solution of the recursive equation

$$P(k) = \begin{cases} \frac{k-1}{k+2} P(k-1) & \text{for } k \geq m+1 \\ 2/(m+2) & \text{for } k = m, \end{cases} \quad (89)$$

giving

$$P(k) = \frac{2m(m+1)}{k(k+1)(k+2)}, \quad (90)$$

very close to Eq. (86) obtained using the continuum theory.

Rate-equation approach: The rate-equation approach, introduced by Krapivsky, Redner, and Leyvraz (2000), focuses on the average number $N_k(t)$ of nodes with k edges at time t . When a new node enters the network in the scale-free model, $N_k(t)$ changes as

$$\frac{dN_k}{dt} = m \frac{(k-1)N_{k-1}(t) - kN_k(t)}{\sum_k kN_k(t)} + \delta_{k,m}. \quad (91)$$

Here the first term accounts for the new edges that connect to nodes with $k-1$ edges, thus increasing their degree to k . The second term describes the new edges connecting to nodes with k edges turning them into nodes with $k+1$ edges, decreasing the number of nodes with k edges. The third term accounts for the new nodes with m edges. In the asymptotic limit $N_k(t)=tP(k)$ and

$\sum_k kN_k(t)=2mt$, leading to the same recursive equation (89), as predicted by the master-equation approach.

The master-equation and rate-equation approaches are completely equivalent and offer the same asymptotic results as the continuum theory. Thus for calculating the scaling behavior of the degree distribution they can be used interchangeably. In addition, these methods, not using a continuum assumption, appear more suitable for obtaining exact results in more challenging network models.

C. Limiting cases of the Barabási-Albert model

The power-law scaling in the Barabási-Albert model indicates that growth and preferential attachment play important roles in network development. But are both of them necessary for the emergence of power-law scaling? To address this question, two limiting cases of the Barabási-Albert model have been investigated, which contain only one of these two mechanisms (Barabási and Albert, 1999; Barabási, Albert, and Jeong, 1999).

Model A keeps the growing character of the network without preferential attachment. Starting with a small number of nodes (m_0), at every time step we add a new node with $m(\leq m_0)$ edges. We assume that the new node connects with equal probability to the nodes already present in the system, i.e., $\Pi(k_i)=1/(m_0+t-1)$, independent of k_i .

The continuum theory predicts that $k_i(t)$ follows a logarithmic time dependence, and for $t \rightarrow \infty$ the degree distribution decays exponentially, following [Fig. 22(a)]

$$P(k) = \frac{e}{m} \exp\left(-\frac{k}{m}\right). \quad (92)$$

The exponential character of the distribution indicates that the absence of preferential attachment eliminates the scale-free character of the resulting network.

Model B starts with N nodes and no edges. At each time step a node is selected randomly and connected with probability $\Pi(k_i)=k_i/\sum_j k_j$ to a node i in the system. Consequently model B eliminates the growth process, the number of nodes being kept constant during the network evolution. Numerical simulations indicate that while at early times the model exhibits power-law scaling, $P(k)$ is not stationary (Fig. 22). Since N is constant and the number of edges increases with time, after $T \approx N^2$ time steps the system reaches a state in which all nodes are connected.

The time evolution of the individual degrees can be calculated analytically using the continuum theory, indicating that

$$k_i(t) \approx \frac{2}{N} t, \quad (93)$$

assuming $N \gg 1$, in agreement with the numerical results [Fig. 22(b)].

Since the continuum theory predicts that after a transient period the average degree of all nodes should have the same value given by Eq. (93), we expect that the

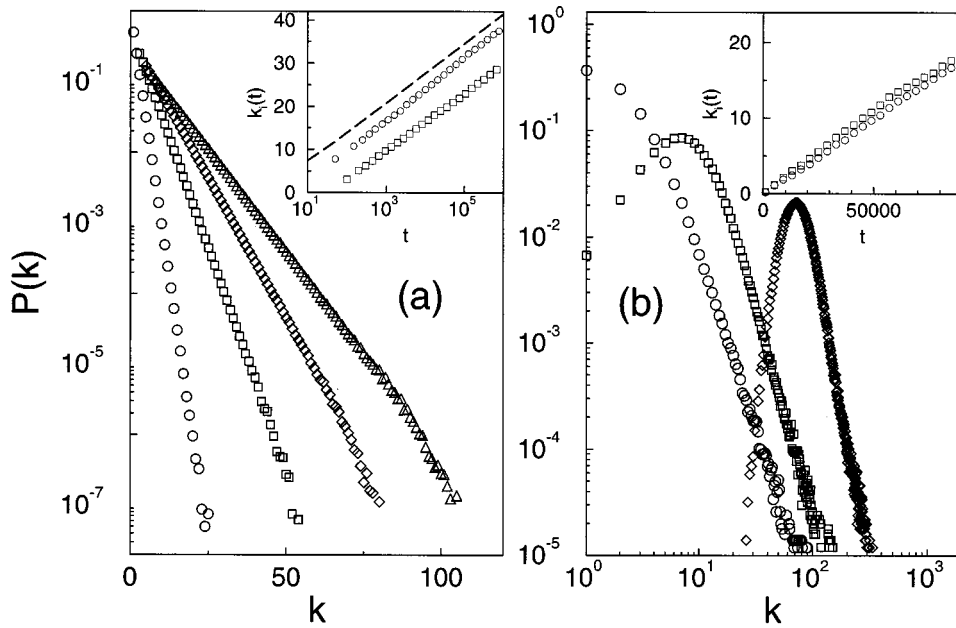


FIG. 22. Degree distribution for two models: (a) Degree distribution for model A: \circ , $m_0 = m = 1$; \square , $m_0 = m = 3$; \diamond , $m_0 = m = 5$; \triangle , $m_0 = m = 7$. The size of the network is $N = 800\,000$. Inset: time evolution for the degree of two vertices added to the system at $t_1 = 7$ and $t_2 = 97$. Here $m_0 = m = 3$. The dashed line follows $k_i(t) = m \ln(m_0 + t - 1)$; (b) the degree distribution for model B for $N = 10\,000$: \circ , $t = N$; \square , $t = 5N$; and \diamond , $t = 40N$. Inset: time dependence of the degrees of two vertices. The system size is $N = 10\,000$. After Barabási, Albert, and Jeong (1999).

degree distribution becomes a Gaussian around its mean value. Indeed, Fig. 22(b) shows that the shape of $P(k)$ changes from the initial power law to a Gaussian.

Motivated by correlations between stocks in financial markets and airline route maps, a prior model incorporating preferential attachment while keeping N constant was independently proposed and studied by Amaral *et al.* (1999).

The failure of models A and B to lead to a scale-free distribution indicates that growth and preferential attachment are needed simultaneously to reproduce the stationary power-law distribution observed in real networks.

D. Properties of the Barabási-Albert model

While the Barabási-Albert model captures the power-law tail of the degree distribution, it has other properties that may or may not agree with empirical results on real networks. As we discussed in Sec. I, a characteristic feature of real networks is the coexistence of clustering and short path lengths. Thus we need to investigate whether the network generated by the model has a small-world character.

1. Average path length

Figure 23 shows the average path length of a Barabási-Albert network with average degree $\langle k \rangle = 4$ as a function of the network size N , compared with the average path length of a random graph with the same size and average degree. The figure indicates that the average path length is smaller in the Barabási-Albert network than in a random graph for any N , indicating that the heterogeneous scale-free topology is more efficient in bringing the nodes close than is the homogeneous topology of random graphs. We find that the average path length of the Barabási-Albert network

increases approximately logarithmically with N , the best fit following a generalized logarithmic form

$$\ell = A \ln(N - B) + C. \quad (94)$$

Recent analytical results indicate that there is a double logarithmic correction to the logarithmic N dependence, i.e., $\ell \sim \ln(N)/\ln \ln(N)$ (Bollobás and Riordan, 2001).

In Fig. 23 we also show the prediction of Eq. (60) for these networks, using the numerically determined number of nearest and next-nearest neighbors. While the fit is good for the random graph, Eq. (60) systematically underestimates the average path length of the Barabási-

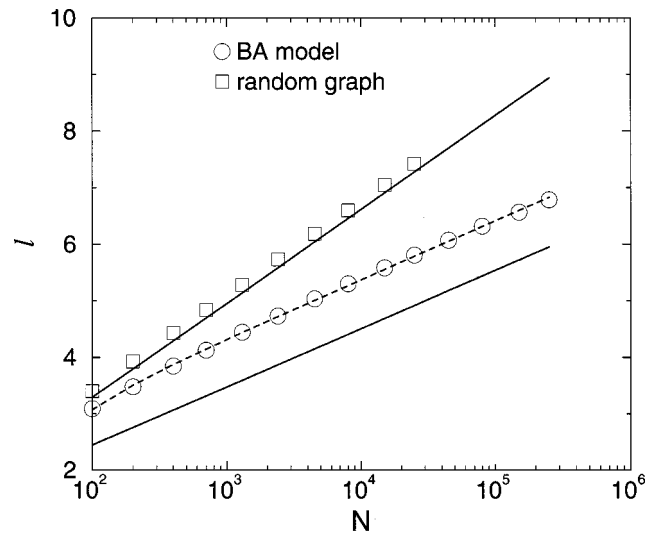


FIG. 23. Characteristic path length ℓ versus network size N in a Barabási-Albert (BA) network with $\langle k \rangle = 4$ (\circ), compared with a random graph of the same size and average degree generated with the algorithm described in Sec. III.A (\square). The dashed line follows Eq. (94), and the solid lines represent Eq. (60) with $z_1 = \langle k \rangle$ and z_2 the numerically obtained number of next-nearest neighbors in the respective networks.

Albert network, as it does the average path length of real networks (see Table II, last three columns).

The failure of Eq. (60) underlies the fact that the topology of the network generated by the Barabási-Albert model is different from the topology of a random network with power-law degree distribution (Sec. V). The dynamical process that generates the network introduces nontrivial correlations that affect all topological properties.

2. Node degree correlations

In random-graph models with arbitrary degree distribution (see Aiello *et al.*, 2000 and Newman, Strogatz, and Watts, 2001), the node degrees are uncorrelated. Krapivsky and Redner (2001) have shown that in the Barabási-Albert model correlations develop spontaneously between the degrees of connected nodes.

Let us consider all node pairs with degree k and l connected by an edge. Without loss of generality we assume that the node with degree k was added later to the system, implying that $k < l$ since, according to Eq. (81), older nodes have higher degree than younger ones, and for simplicity we use $m=1$. Denoting by $N_{kl}(t)$ the number of connected pairs of nodes with degree k and l , we have

$$\begin{aligned} \frac{dN_{kl}}{dt} = & \frac{(k-1)N_{k-1,l} - kN_{kl}}{\sum_k kN(k)} + \frac{(l-1)N_{k,l-1} - lN_{kl}}{\sum_k kN(k)} \\ & + (l-1)N_{l-1}\delta_{k1}. \end{aligned} \quad (95)$$

The first term on the right-hand side accounts for the change in N_{kl} due to the addition of an edge to a node of degree $k-1$ or k that is connected to a node of degree l . Since the addition of a new edge increases the node's degree by 1, the first term in the numerator corresponds to a gain in N_{kl} , while the second corresponds to a loss. The second term on the right-hand side incorporates the same effects as the first applied to the other node. The last term takes into account the possibility that $k=1$; thus the edge that is added to the node with degree $l-1$ is the same edge that connects the two nodes.

This equation can be transformed into a time-independent recursion relation using the hypotheses $\sum_k kN(k) \rightarrow 2t$ and $N_{kl}(t) \rightarrow tn_{kl}$. Solving for n_{kl} we obtain

$$\begin{aligned} n_{kl} = & \frac{4(l-1)}{k(k+1)(k+l)(k+l+1)(k+l+2)} \\ & + \frac{12(l-1)}{k(k+l-1)(k+l)(k+l+1)(k+l+2)}. \end{aligned} \quad (96)$$

For a network with an arbitrary degree distribution, if the edges are placed randomly, $n_{kl} = n_k n_l$. The most important feature of the result (96) is that the joint distribution does not factorize, i.e., $n_{kl} \neq n_k n_l$. This indicates the spontaneous appearance of correlations between the

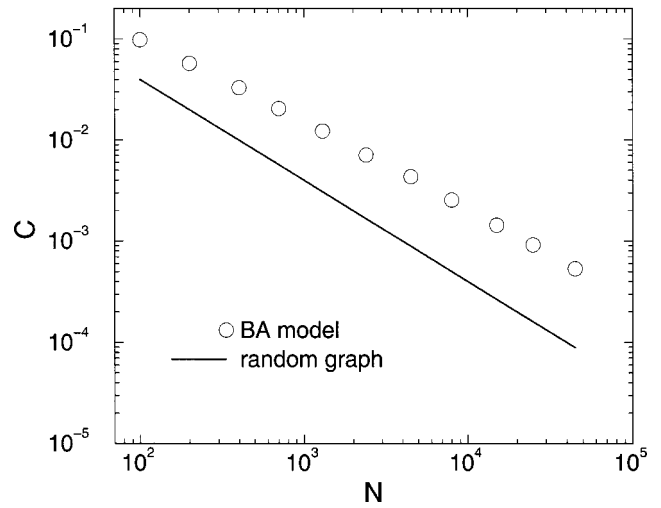


FIG. 24. Clustering coefficient versus size of the Barabási-Albert (BA) model with $\langle k \rangle = 4$, compared with the clustering coefficient of a random graph, $C_{rand} \approx \langle k \rangle / N$.

degrees of the connected nodes. The only case in which n_{kl} can be simplified to a factorized expression is when $1 \ll k \ll l$, and n_{kl} becomes

$$n_{kl} \approx k^{-2} l^{-2}, \quad (97)$$

but even then it is different from $n_{kl} = k^{-3} l^{-3}$, as expected if correlations are absent from the network. This result offers the first explicit proof that the dynamical process that creates a scale-free network builds up nontrivial correlations between the nodes that are not present in the uncorrelated models discussed in Sec. V.

3. Clustering coefficient

While the clustering coefficient has been much investigated for the Watts-Strogatz model (Sec. VI.B.2), there is no analytical prediction for the Barabási-Albert model. Figure 24 shows the clustering coefficient of a Barabási-Albert network with average degree $\langle k \rangle = 4$ and different sizes, compared with the clustering coefficient $C_{rand} = \langle k \rangle / N$ of a random graph. We find that the clustering coefficient of the scale-free network is about five times higher than that of the random graph, and this factor slowly increases with the number of nodes. However, the clustering coefficient of the Barabási-Albert model decreases with the network size, following approximately a power law $C \sim N^{-0.75}$, which, while a slower decay than the $C = \langle k \rangle N^{-1}$ decay observed for random graphs, is still different from the behavior of the small-world models, where C is independent of N .

4. Spectral properties

The spectral density of the Barabási-Albert model is continuous, but it has a markedly different shape from the semicircular spectral density of random graphs (Farkas *et al.*, 2001; Goh, Kahng, and Kim, 2001). Numerical simulations indicate that the bulk of $\rho(\lambda)$ has a triangle-like shape with the top lying well above the semicircle and edges decaying as a power law (Fig. 25). This power-

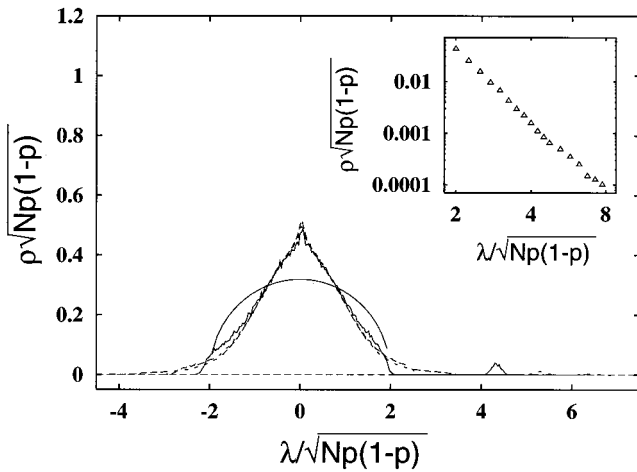


FIG. 25. Rescaled spectral density of three Barabási-Albert networks having $m=m_0=5$ and various sizes N : solid line, $N=100$; long-dashed line, $N=300$; short-dashed line, $N=1000$. The semicircle law corresponding to random graphs is drawn for comparison. The isolated peak corresponds to the largest eigenvalue, which increases as $N^{1/4}$. Inset: the edge of the spectral density decays as a power law. After Farkas *et al.* (2001).

law decay is due to the eigenvectors localized on the nodes with the highest degree. As in the case of random graphs (and unlike small-world networks), the principal eigenvalue, λ_1 , is clearly separated from the bulk of the spectrum. A lower bound for λ_1 can be given as the square root of the network's largest degree k_1 . The node degrees in the Barabási-Albert model increase as $N^{1/2}$; hence λ_1 increases approximately as $N^{1/4}$. Numerical results indicate that λ_1 deviates from the expected behavior for small network sizes, reaching it asymptotically for $N \rightarrow \infty$. This crossover indicates the presence of correlations between the longest row vectors, offering additional evidence for correlations in the Barabási-Albert model.

The principal eigenvalue plays an important role in the moments of $\rho(\lambda)$, determining the loop structure of the network. In contrast with the subcritical random graph (i.e., $p < 1/N$), where the fraction of loops becomes negligible, in a Barabási-Albert network the fraction of loops with more than four edges increases with N , and the growth rate of the loops increases with their size. Note that the fraction of triangles decreases as $N \rightarrow \infty$ (Bianconi, 2000b; Gleiss *et al.*, 2001).

While for random graphs $\rho(\lambda)$ follows the semicircle law (Wigner, 1955, 1957, 1958), deriving a similarly simple expression for small-world (see Sec. VI.B.4) and scale-free networks remains a considerable challenge.

VIII. THE THEORY OF EVOLVING NETWORKS

The Barabási-Albert model discussed in the previous section is a minimal model that captures the mechanisms responsible for the power-law degree distribution. Compared to real networks, it has evident limitations: it predicts a power-law degree distribution with a fixed exponent, while the exponents measured for real networks

vary between 1 and 3 (see Table II). In addition, the degree distribution of real networks can have non-power-law features such as exponential cutoffs (see Amaral *et al.*, 2000; Jeong, Mason, *et al.* 2001; Newman 2001b, 2001c) or saturation for small k . These discrepancies between the model and real networks led to a surge of interest in addressing several basic questions of network evolution: How can we change the scaling exponents? Are there universality classes similar to those seen in critical phenomena, characterized by unique exponents? How do various microscopic processes, known to be present in real networks, influence the network topology? Are there quantities beyond the degree distribution that could help in classifying networks? While the community is still in the process of answering these questions, several robust results are already available. These results signal the emergence of a self-consistent theory of evolving networks, offering unprecedented insights into network evolution and topology.

A. Preferential attachment $\Pi(k)$

A central ingredient of all models aiming to generate scale-free networks is preferential attachment, i.e., the assumption that the likelihood of receiving new edges increases with the node's degree. The Barabási-Albert model assumes that the probability $\Pi(k)$ that a node attaches to node i is proportional to the degree k of node i [see Eq. (78)]. This assumption involves two hypotheses: first, that $\Pi(k)$ depends on k , in contrast to random graphs in which $\Pi(k)=p$, and second, that the functional form of $\Pi(k)$ is linear in k . The precise form of $\Pi(k)$ is more than a purely academic question, as recent studies have demonstrated that the degree distribution depends strongly on $\Pi(k)$. To review these developments we start by discussing the empirical results on the functional form of $\Pi(k)$, followed by the theoretical work predicting the effect of $\Pi(k)$ on the network topology.

1. Measuring $\Pi(k)$ for real networks

The functional form of $\Pi(k)$ can be determined for networks for which we know the time at which each node joined the network (Jeong, Nédá, and Barabási, 2001; Newman 2001d; Pastor-Satorras *et al.*, 2001). Such dynamical data are available for the co-authorship network of researchers, the citation network of articles, the actor collaboration network, and the Internet at the domain level (see Sec. II).

Consider the state of the network at a given time, and record the number of “old” nodes present in the network and their degrees. Next measure the increase in the degree of the “old” nodes over a time interval ΔT , much shorter than the age of the network. Then, according to Eq. (78), plotting the relative increase $\Delta k_i / \Delta k$ as a function of the earlier degree k_i for every node gives the $\Pi(k)$ function. Here Δk is the number of edges

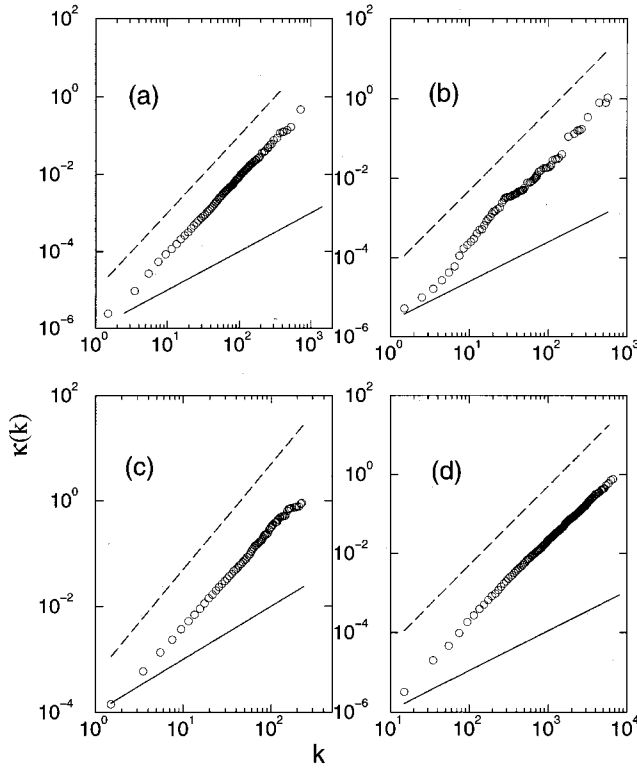


FIG. 26. Cumulative preferential attachment for (a) the citation network; (b) the Internet; (c) the neuroscience scientific collaboration network; (d) the actor collaboration network. In all panels the dashed line corresponds to linear preferential attachment, and the solid line to no preferential attachment. After Jeong, Néda, and Barabási (2001).

added to the network in the time ΔT . We can reduce fluctuations in the data by plotting the cumulative distribution

$$\kappa(k) = \sum_{k_i=0}^k \Pi(k_i). \quad (98)$$

As Fig. 26 shows, the obtained $\Pi(k)$ supports the existence of preferential attachment. Furthermore, it appears that in each case $\Pi(k)$ follows a power law, i.e.,

$$\Pi(k) \sim k^\alpha. \quad (99)$$

In some cases, such as the Internet (Jeong, Néda, and Barabási, 2001; Pastor-Satorras *et al.*, 2001), the citation network (Jeong, Néda, and Barabási, 2001), Medline, and the Los Alamos archive (Newman, 2001d) we have $\alpha \approx 1$, i.e., $\Pi(k)$ depends linearly on k as assumed in the Barabási-Albert model. For other networks the dependence is sublinear, with $\alpha = 0.8 \pm 0.1$ for the neuroscience co-authorship and the actor collaboration networks (Jeong, Néda, and Barabási, 2001).

2. Nonlinear preferential attachment

The effect of a nonlinear $\Pi(k)$ on the network dynamics and topology was explained by Krapivsky, Redner, and Leyvraz (2000). Replacing linear preferential attachment [Eq. (78)] with Eq. (99) in a directed network model, Krapivsky, Redner, and Leyvraz calculate

the average number $N_k(t)$ of nodes with $k-1$ incoming edges at time t by the rate-equation approach (see Sec. VII.B). The time evolution of $N_k(t)$ follows

$$\frac{dN_k}{dt} = \frac{1}{M_\alpha} [(k-1)^\alpha N_{k-1} - k^\alpha N_k] + \delta_{k1}, \quad (100)$$

where $M_\alpha(t) = \sum k^\alpha N_k(t)$ is the α th moment of $N_k(t)$. In Eq. (100) the first term accounts for new nodes that connect to nodes with $k-1$ edges, thus increasing their degree to k . The second term describes new nodes connecting to nodes with k edges, turning them into nodes with $k+1$ edges and hence decreasing the number of nodes with k edges. The third term accounts for the continuous introduction of new nodes with a single outgoing edge.

Depending on the value of α , distinct phases have been identified:

- (a) *Sublinear case* ($\alpha < 1$): In this regime in the long-time limit $M_\alpha(t)$ satisfies $M_\alpha(t) = \mu t$, with a prefactor $1 \leq \mu = \mu(\alpha) \leq 2$. Substituting $M_\alpha(t)$ and N_k into Eq. (100), we obtain the degree distribution

$$P(k) = \frac{\mu}{k^\alpha} \prod_{j=1}^k \left(1 + \frac{\mu}{j^\alpha} \right)^{-1}. \quad (101)$$

This product can be expanded in series, and the result is a stretch exponential in which a new term arises whenever α decreases below $1/l$, where l is an arbitrary positive integer.

- (b) *Superlinear preferential attachment* ($\alpha > 1$): In this regime Eq. (100) has no analytical solution, but its discretized version can be used to determine recursively the leading behavior of each N_k as $t \rightarrow \infty$. For $\alpha > 2$ a “winner-takes-all” phenomenon emerges, such that almost all nodes have a single edge, connecting them to a “gel” node that has the rest of the edges of the network. For $3/2 < \alpha < 2$ the number of nodes with two edges grows as $t^{2-\alpha}$, while the number of nodes with more than two edges is again finite. Again, the rest of the edges belong to the gel node. In general for $(l+1)/l < \alpha < l/(l-1)$ the number of nodes with more than l edges is finite even in infinite systems, while $N_k \sim t^{k-(k-1)\alpha}$ for $k \leq l$.

In conclusion, the analytical calculations of Krapivsky, Redner, and Leyvraz demonstrate that the scale-free nature of the network is destroyed for nonlinear preferential attachment. The only case in which the topology of the network is scale free is that in which the preferential attachment is asymptotically linear, i.e., $\Pi(k_i) \sim a_\infty k_i$ as $k_i \rightarrow \infty$. In this case the rate equation leads to

$$P(k) \sim k^{-\gamma} \quad \text{with} \quad \gamma = 1 + \frac{\mu}{a_\infty}. \quad (102)$$

This way the exponent of the degree distribution can be tuned to any value between 2 and ∞ .

3. Initial attractiveness

Another general feature of $\Pi(k)$ in real networks is that $\Pi(0) \neq 0$, i.e., there is a nonzero probability that a

new node attaches to an isolated node (Jeong, Nédá, and Barabási, 2001). Thus in general $\Pi(k)$ has the form

$$\Pi(k) = A + k^\alpha, \quad (103)$$

where A is the initial attractiveness of the node i (Dorogovtsev, Mendes, and Samukhin, 2000a). Indeed, if $A = 0$, a node that has $k = 0$ can never increase its connectivity according to Eq. (78). However, in real networks every node has a finite chance to be “discovered” and linked to, even if it has no edges to start with. Thus the parameter A describes the likelihood that an isolated node will be discovered, such as a new article’s being cited the first time.

Dorogovtsev, Mendes, and Samukhin (2000a) gave an exact solution for a class of growing network models using the master-equation approach (see Sec. VII.B). In their model at every time step a new node is added to the network, followed by the addition of m directed edges pointing from any node in the network to preferentially chosen nodes. The probability that a node will receive an incoming edge is proportional to the sum of an initial attractiveness and the number of incoming edges, i.e., $\Pi(k_{in}) = A + k_{in}$. The calculations indicate that the degree distribution follows $P(k) \sim k^{-\gamma}$ with $\gamma = 2 + A/m$. Consequently initial attractiveness does not destroy the scale-free nature of the degree distribution; it only changes the degree exponent. These results agree with the conclusion of Krapivsky, Redner, and Leyvraz (2000), who find that the power law $P(k)$ is preserved for a shifted linear $\Pi(k)$, since the effect of the initial attractiveness diminishes as $k \rightarrow \infty$. A generalization of the Dorogovtsev-Mendes-Samukhin model (Dorogovtsev, Mendes, and Samukhin, 2000b) allows for the random distribution of n_r edges and an initial degree n of every new node. These changes do not modify the asymptotically linear scaling of the preferential attachment; thus this model also gives a power-law degree distribution with $\gamma = 2 + (n_r + n + A)/m$.

B. Growth

In the Barabási-Albert model the number of nodes and edges increases linearly in time, and consequently the average degree of the network is constant. In this section we discuss the effect of nonlinear growth rates on the network dynamics and topology.

1. Empirical results

The ability of networks to follow different growth patterns is supported by several recent measurements. For example, the average degree of the Internet in November of 1997 was 3.42, but it increased to 3.96 by December of 1998 (Faloutsos *et al.*, 1999). Similarly, the World Wide Web has increased its average degree from 7.22 to 7.86 in the five months between the measurements of Broder *et al.* (2000). The average degree of the co-authorship network of scientists has been found to continuously increase over an eight-year period (Barabási *et al.*, 2001). Finally, comparison of metabolic networks of organisms of different sizes indicates that the average

degree of the substrates increases approximately linearly with the number of substrates involved in the metabolism (Jeong *et al.*, 2000). The increase of the average degree indicates that in many real systems the number of edges increases faster than the number of nodes, supporting the presence of a phenomenon called *accelerated growth*.

2. Analytical results

Dorogovtsev and Mendes (2001a) studied analytically the effect of accelerated growth on the degree distribution, generalizing the directed model with the asymptotically linear preferential attachment of Dorogovtsev, Mendes, and Samukhin (2000a; see also Sec. VIII.A). In this model, at every step a new node is added to the network, which receives n incoming edges from random nodes in the system. Additionally $c_0 t^\theta$ new edges are distributed, each of them being directed from a randomly selected node to a node with high incoming degree, with asymptotically linear preferential attachment $\Pi(k_{in}) \propto A + k_{in}$. The authors show that accelerated growth, controlled by the exponent θ , does not change the scale-free nature of the degree distribution, but it modifies the degree exponent, which now becomes

$$\gamma = 1 + \frac{1}{1 + \theta}. \quad (104)$$

While the model of Dorogovtsev and Mendes (2001a) is based on a directed network, Barabási *et al.* (2001) discuss an undirected model motivated by measurements on the evolution of the co-authorship network. In the model new nodes are added to the system with a constant rate, and these new nodes connect to b nodes already in the system with preferential attachment

$$P_i = b \frac{k_i}{\sum_j k_j}. \quad (105)$$

Additionally, at every time step a linearly increasing number of edges (constituting a fraction a of the nodes that are present in the network) are distributed between the nodes, the probability that an edge is added between nodes i and j being

$$P_{ij} = \frac{k_i k_j}{\sum_{s,l} k_s k_l} N(t) a. \quad (106)$$

Here $N(t)$ is the number of nodes in the system and the summation goes over all nonequal values of s and l . As a result of these two processes the average degree of the network increases linearly in time, following $\langle k \rangle = at + 2b$, in agreement with the measurements on the real co-author network. The continuum theory predicts that the time-dependent degree distribution displays a crossover at a critical degree,

$$k_c = \sqrt{b^2 t (2 + 2at/b)^{3/2}}, \quad (107)$$

such that for $k \ll k_c$, $P(k)$ follows a power law with exponent $\gamma=1.5$ and for $k \gg k_c$ the exponent is $\gamma=3$. This result explains the fast-decaying tail of the degree distributions measured by Newman (2001a), and it indicates that as time increases the scaling behavior with $\gamma=1.5$ becomes increasingly visible. An equivalent model, proposed by Dorogovtsev and Mendes (2001c), was able to reproduce the two separate power-law regimes in the distribution of word combinations (Ferrer i Cancho and Solé, 2001).

C. Local events

The Barabási-Albert model incorporates only one mechanism for network growth: the addition of new nodes that connect to the nodes already in the system. In real systems, however, a series of microscopic events shape the network evolution, including the addition or rewiring of new edges or the removal of nodes or edges. Lately several models have been proposed to investigate the effect of selected processes on the scale-free nature of the degree distribution, offering a more realistic description of various real networks. Any local change in the network topology can be obtained through a combination of four elementary processes: addition or removal of a node and addition or removal of an edge. But in reality these events come jointly; for example, the rewiring of an edge is a combination of an edge removal and addition. Next we briefly review several studies that address in general terms the effects of local events on network topology.

1. Internal edges and rewiring

A model that incorporates new edges between existing nodes and the rewiring of edges was discussed by Albert and Barabási (2000). Starting with m_0 isolated nodes, at each time step we perform one of the following three operations:

- (i) With probability p we add $m(m \leq m_0)$ new edges. One end of a new edge is selected randomly, the other with probability

$$\Pi(k_i) = \frac{k_i + 1}{\sum_j (k_j + 1)}. \quad (108)$$
- (ii) With probability q we rewire m edges. For this we randomly select a node i and remove an edge l_{ij} connected to it, replacing it with a new edge $l_{ij'}$ that connects i with node j' chosen with probability $\Pi(k_j')$ given by Eq. (108).
- (iii) With probability $1-p-q$ we add a new node. The new node has m new edges that with probability $\Pi(k_i)$ are connected to nodes i already present in the system.

In the continuum theory the growth rate of the degree of a node i is given by

$$\frac{\partial k_i}{\partial t} = (p-q)m \frac{1}{N} + m \frac{k_i + 1}{\sum_j (k_j + 1)}. \quad (109)$$

The first term on the right-hand side corresponds to the random selection of node i as a starting point of a new edge (with probability p) or as the end point from which an edge is disconnected (with probability q). The second term corresponds to the selection of node i as an end point of an edge with the preferential attachment present in all three of the possible processes.

The solution of Eq. (109) has the form

$$k_i(t) = [A(p, q, m) + m + 1] \left(\frac{t}{t_i} \right)^{1/B(p, q, m)} - A(p, q, m) - 1, \quad (110)$$

where

$$A(p, q, m) = (p-q) \left(\frac{2m(1-q)}{1-p-q} + 1 \right),$$

$$B(p, q, m) = \frac{2m(1-q) + 1 - p - q}{m}. \quad (111)$$

The corresponding degree distribution has the generalized power-law form

$$P(k) \propto [k + \kappa(p, q, m)]^{-\gamma(p, q, m)}, \quad (112)$$

where $\kappa(p, q, m) = A(p, q, m) + 1$ and $\gamma(p, q, m) = B(p, q, m) + 1$.

Equation (112) is valid only when $A(p, q, m) + m + 1 > 0$, which, for fixed p and m , translates into $q < q_{\max} = \min\{1-p, (1-p+m)/(1+2m)\}$. Thus the (p, q) phase diagram separates into two regions: For $q < q_{\max}$ the degree distribution is given by Eq. (112), following a generalized power law. For $q > q_{\max}$, however, Eq. (112) is not valid, but numerical simulations indicate that $P(k)$ approaches an exponential.

While a power-law tail is present in any point of the scale-free regime, for small k the probability saturates at $P[\kappa(p, q, m)]$, a feature seen in many real networks [Figs. 3(b) and (d)]. In addition, the exponent $\gamma(p, q, m)$ characterizing the tail of $P(k)$ for $k \gg \kappa(p, q)$ changes continuously with p , q , and m , predicting a range of exponents between 2 and ∞ . The realistic nature of $P(k)$ was confirmed by successfully fitting it to the degree distribution of the actor collaboration network (Albert and Barabási, 2000).

2. Internal edges and edge removal

Dorogovtsev and Mendes (2000c) consider a class of undirected models in which new edges are added between old sites and existing edges can be removed. In the first variant of the model, called a developing network, c new edges are introduced at every time step, which connect two unconnected nodes i and j with a probability proportional to the product of their degrees [as in Eq. (106)], an assumption confirmed by empirical measurements on the co-authorship network (Barabási

et al., 2001). It is assumed that c can be tuned continuously, such that $c > 0$ for a developing and $c < 0$ for a decaying network. The continuum theory predicts that the rate of change of the node degrees has the form

$$\frac{\partial k_i}{\partial t} = \frac{k_i(t)}{\int_0^t k_j(t) dt_j} + 2c \frac{k_i(t) \left[\int_0^t k_j(t) dt_j - k_i(t) \right]}{\left[\int_0^t k_j(t) dt_j \right]^2 - \int_0^t k_j^2(t) dt_j}, \quad (113)$$

where the summation over all nodes $\sum_j k_j$ has been approximated by an integral over all introduction times t_j . The first term on the right-hand side incorporates linear preferential attachment, while the second term corresponds to the addition of c new edges. Every node can be at either end of the new edge, and the probability of a node i becoming an end of the new edge is proportional to the product of its degree k_i and the sum of the degrees k_j of all other nodes. The normalization factor is the sum of all products $k_i k_j$ with i different from j .

In the asymptotic limit the second term can be neglected compared with the first term in both the numerator and denominator, and Eq. (113) becomes

$$\frac{\partial k_i}{\partial t} = (1 + 2c) \frac{k_i(t)}{\int_0^t k_j(t) dt_j}, \quad (114)$$

which predicts the dynamic exponent (81) as

$$\beta = \frac{1 + 2c}{2(1 + c)} \quad (115)$$

and the degree exponent as

$$\gamma = 2 + \frac{1}{1 + 2c}. \quad (116)$$

The limiting cases of this developing network are $c = 0$ when the familiar Barabási-Albert values $\beta = 1/2$ and $\gamma = 3$ are obtained, and $c \rightarrow \infty$, when $\beta \rightarrow 1$ and $\gamma \rightarrow 2$.

In the decaying network at every time step $|c|$ edges are removed randomly. The decrease in the node degrees due to this process is proportional to their current value, so Eq. (114) applies here as well, the only difference being that now $c < 0$. A more rigorous calculation accounting for the fact that only existing edges can be removed confirms that the end result is identical with Eqs. (115) and (116), only with negative c . The limiting value of c is -1 , since the rate of removal of edges cannot be higher than the rate of addition of new nodes and edges, leading to the limit exponents $\beta \rightarrow -\infty$ and $\gamma \rightarrow \infty$.

D. Growth constraints

For many real networks the nodes have a finite life-time (for example, in social networks) or a finite edge capacity (Internet routers or nodes in the electrical power grid). Recently several groups have addressed the degree to which such constraints affect the degree distribution.

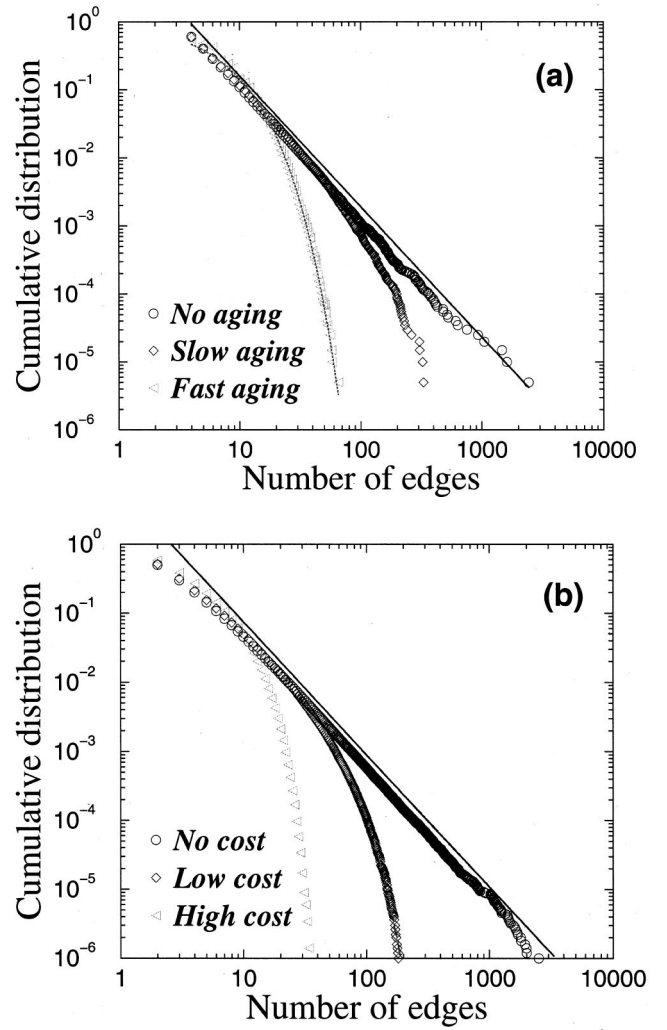


FIG. 27. Deviation from a power law of the degree distribution due to adding (a) age and (b) capacity constraints to the Barabási-Albert model. The constraints result in cutoffs of the power-law scaling. After Amaral *et al.* (2000).

1. Aging and cost

Amaral *et al.* (2000) suggested that while several networks do show deviations from the power-law behavior, they are far from being random networks. For example, the degree distribution of the electric power grid of southern California and of the neural network of the worm *C. elegans* is more consistent with a single-scale exponential distribution. Other networks, like the extended actor collaboration network, in which TV films and series are included, have a degree distribution in which power-law scaling is followed by an exponential cutoff for large k . In all these examples there are constraints limiting the addition of new edges. For example, the actors have a finite active period during which they are able to collect new edges, while for the electrical power grid or neural networks there are constraints on the total number of edges a particular node can have, driven by economic, physical, or evolutionary reasons. Amaral *et al.* propose that in order to explain these deviations from a pure power law we need to incorporate

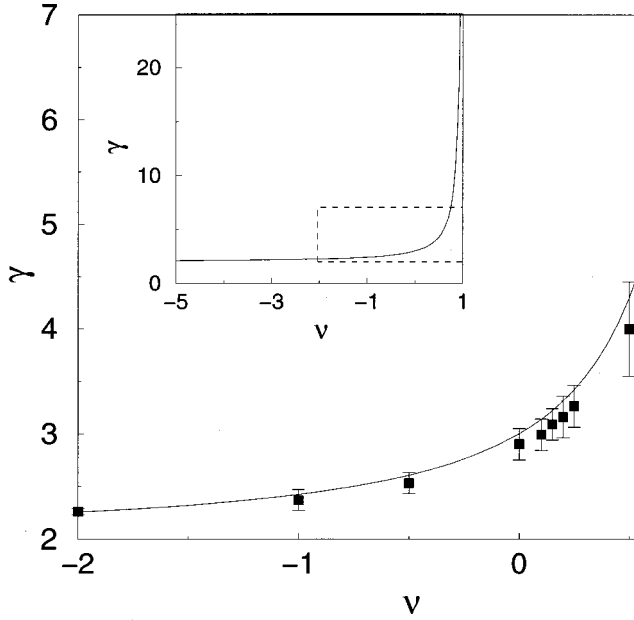


FIG. 28. The dependence of the degree exponent γ on the aging exponent ν in the model of Dorogovtsev and Mendes (2000b). The points are obtained from simulations, while the solid line is the prediction of the continuum theory. After Dorogovtsev and Mendes (2000b).

aging and cost or capacity constraints. The model studied by them evolves following growth and preferential attachment, but when a node reaches a certain age (aging) or has more than a critical number of edges (capacity constraints), new edges cannot connect to it. In both cases numerical simulations indicate that while for small k the degree distribution still follows a power law, for large k an exponential cutoff develops (Fig. 27).

2. Gradual aging

Dorogovtsev and Mendes (2000b) propose that in some systems the probability that a new node connects to a node i is not only proportional to the degree k_i of node i , but it also depends on its age, decaying as $(t - t_i)^{-\nu}$, where ν is a tunable parameter. Papers or actors gradually lose their ability to attract more edges, the model assuming that this phaseout follows a power law. The calculations predict that the degree distribution depends on the exponent ν : power-law scaling is present only for $\nu < 1$, and the degree exponent depends on ν (Fig. 28). Moreover, when $\nu > 1$ power-law scaling completely disappears, the degree distribution approaching an exponential.

E. Competition in evolving networks

The Barabási-Albert model assumes that all nodes increase their degree following a power-law time dependence with the same dynamic exponent $\beta = 1/2$ [Eq. (81)]. As a consequence, the oldest nodes have the highest number of edges, since they had the longest lifetime to accumulate them. However, numerous examples indi-

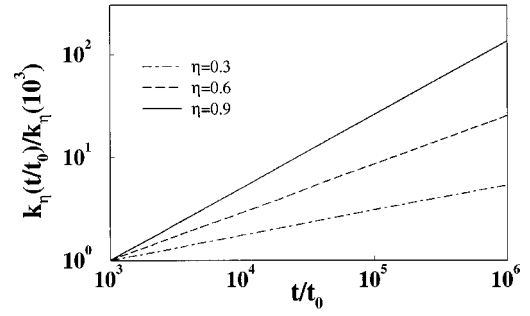


FIG. 29. Time dependence of the degree $k_\eta(t)$, for nodes with fitness $\eta = 0.3, 0.6$, and 0.9 . Note that $k_\eta(t)$ follows a power law in each case and the dynamic exponent $\beta(\eta)$, given by the slope of $k(t)$, increases with η . After Bianconi and Barabási (2000a).

cate that in real networks a node's degree and growth rate do not depend on age alone. For example, on the World Wide Web some documents acquire a large number of edges in a very short time through a combination of good content and marketing (Adamic and Huberman, 2000), and some research papers acquire many more citations than their peers. Several studies have offered models that address this phenomenon.

1. Fitness model

Bianconi and Barabási (2001a) argue that real networks have a competitive aspect, as each node has an intrinsic ability to compete for edges at the expense of other nodes. They propose a model in which each node is assigned a fitness parameter η_i which does not change in time. Thus at every time step a new node j with a fitness η_j is added to the system, where η_j is chosen from a distribution $\rho(\eta)$. Each new node connects with m edges to the nodes already in the network, and the probability of connecting to a node i is proportional to the degree and the fitness of node i ,

$$\Pi_i = \frac{\eta_i k_i}{\sum_j \eta_j k_j}. \quad (117)$$

This generalized preferential attachment ensures that even a relatively young node with a few edges can acquire edges at a high rate if it has a high fitness parameter. The continuum theory predicts that the rate of change of the degree of node i is

$$\frac{\partial k_i}{\partial t} = m \frac{\eta_i k_i}{\sum_k \eta_k k_k}. \quad (118)$$

Assuming that the time evolution of k_i follows Eq. (81) with a fitness-dependent $\beta(\eta)$,

$$k_{\eta_i}(t, t_i) = m \left(\frac{t}{t_i} \right)^{\beta(\eta_i)}, \quad (119)$$

the dynamic exponent satisfies

$$\beta(\eta) = \frac{\eta}{C} \quad \text{with} \quad C = \int \rho(\eta) \frac{\eta}{1 - \beta(\eta)} d\eta. \quad (120)$$

Thus β is described by a spectrum of values governed by the fitness distribution (Fig. 29). Equation (119) indicates that nodes with higher fitness increase their degree faster than those with lower fitness. Thus the fitness model allows for late but fit nodes to take a central role in the network topology. The degree distribution of the model is a weighted sum of different power laws,

$$P(k) \sim \int \rho(\eta) \frac{C}{\eta} \left(\frac{m}{k} \right)^{C/\eta + 1}, \quad (121)$$

which depends on the choice of the fitness distribution (see Sec. VIII.G.2). For example, for a uniform fitness distribution Eq. (120) gives $C=1.255$ and $\beta(\eta) = \eta/1.255$, and the degree distribution is

$$P(k) \sim \frac{k^{-C-1}}{\ln(k)}, \quad (122)$$

i.e., a power law with a logarithmic correction. The fitness model can be extended to incorporate additional processes, such as internal edges, which affect the exponents, a problem studied by Ergün and Rodgers (2001).

2. Edge inheritance

A different mechanism that gives individuality to the new nodes is proposed by Dorogovtsev, Mendes, and Samukhin (2000c). They build on the evolving directed-network algorithm introduced in their earlier paper (Dorogovtsev, Mendes, and Samukhin, 2000a), this time assuming that the degree of the new nodes is not constant but depends on the state of the network at the time the new node is added to the system. Specifically, every new node is assumed to be an “heir” of a randomly chosen old node, and it inherits a fraction c of the old node’s incoming edges (i.e., a fraction c of the nodes that point to the parent node will also point to the heir). The parameter c is assumed to be distributed with a probability density $h(c)$.

The time-dependent degree distribution for uniformly distributed c indicates that the fraction of nodes with no incoming edges increases and tends to 1 asymptotically. The distribution of nonzero incoming edges tends to a distribution

$$P(k_{in}, k_{in} \neq 0) = \frac{d}{k_{in}^2} \ln(ak_{in}), \quad (123)$$

where $d \approx 0.174$ and $a \approx 0.84$.

F. Alternative mechanisms for preferential attachment

It is now established that highly connected nodes have better chances of acquiring new edges than their less-connected counterparts. The Barabási-Albert model reflects this fact by incorporating it explicitly through preferential attachment (78). But where does preferential attachment come from? We do not yet have a universal answer to this question, and there is a growing suspicion

that the mechanisms responsible for preferential attachment are system dependent. However, recently several papers have offered promising proposals and models that shed some light on this issue. The unifying theme of these models is that while a preferential attachment is not explicitly introduced, the mechanisms used to place nodes and edges effectively induce one. The diversity of the proposals vividly illustrates the wide range of microscopic mechanisms that could effect the evolution of growing networks and still lead to the observed scale-free topologies.

1. Copying mechanism

Motivated by the desire to explain the power-law degree distribution of the World Wide Web, Kleinberg *et al.* (1999) and Kumar *et al.* (2000a, 2000b) assume that new Web pages dedicated to a certain topic copy links from existing pages on the same topic. In this model, at each time step a new node is added to the network, which connects to the rest of the nodes with a constant number of directed edges. At the same time a “prototype” node is chosen randomly from the nodes already in the system. The outgoing edges of the new node are distributed in the following way: with probability p the destination of the i th edge is selected randomly, and with probability $1-p$ it is taken to be the destination of the i th edge of the prototype node. This second process increases the probability of high-degree nodes’ receiving new incoming edges. In fact, since the prototype nodes are selected randomly, the probability that a Web page with degree k will receive a new hyperlink is proportional to $(1-p)k$, indicating that the copying mechanism effectively amounts to a linear preferential attachment. Kumar *et al.* prove that the expectation of the incoming degree distribution is

$$P(k_{in}) = k^{-(2-p)/(1-p)}, \quad (124)$$

thus $P(k)$ follows a power law with an exponent that varies between 2 (for $p \rightarrow 0$) and ∞ (for $p \rightarrow 1$).

2. Edge redirection

Although inspired by a different mechanism, the growing network with the redirection model of Krapivsky and Redner (2001) is mathematically equivalent with the model of Kumar *et al.* (2000a, 2000b) discussed above. In this model at every time step a new node is added to the system and an earlier node i is selected uniformly as a possible target for attachment. With probability $1-r$ a directed edge from the new node to i is created; however, with probability r the edge is redirected to the ancestor node j of node i (i.e., the node at which i attached when it was first added to the network).

When the rate-equation approach (Sec. VII.B) is applied, the number of nodes $N(k)$ with degree k evolves as

$$\frac{dN(k)}{dt} = \delta_{k1} + \frac{1-r}{M_0}(N_{k-1} - N_k) + \frac{r}{M_0}[(k-2)N_{k-1} - (k-1)N_k]. \quad (125)$$

The first term corresponds to nodes that are just added to the network. The second term indicates the random selection of a node to which the new node will attach. This process affects $N(k)$ if this node has a degree $k-1$ [in which case its degree will become k , increasing $N(k)$] or k [in which case $N(k)$ decreases by one]. The normalization factor M_0 is the sum of all degrees. The third term corresponds to the rewiring process. Since the initial node is chosen uniformly, if redirection does occur, the probability that a node with $k-1$ preexisting edges will receive the redirected edge is proportional to $k-2$, the number of preexisting incoming edges. Thus redirection also leads to a linear preferential attachment.

This rate equation is equivalent with Eq. (100) with an asymptotically linear attachment $\Pi(k) \sim k-2+1/r$. Thus this model leads to a power-law degree distribution with degree exponent $\gamma=1+1/r$, which can be tuned to any value larger than 2.

3. Walking on a network

The walking mechanism proposed by Vázquez (2000) was inspired by citation networks. Entering a new field, we are usually aware of a few important papers and follow the references included in these to find other relevant articles. This process is continued recursively, such that a manuscript will contain references to papers discovered this way. Vázquez formulates the corresponding network algorithm in the following way: We start with an isolated node. At every time step a new node is added that links with a directed edge to a randomly selected node, and then it follows the edges starting from this node and links to their end points with probability p . This last step is repeated starting from the nodes to which connections were established, until no new target node is found. In fact, this algorithm is similar to the breadth-first search used in determining the cluster structure of a network, with the exception that not all edges are followed, but only a fraction equal to p . In the special case of $p=1$ one can see that nodes of high degree will be more likely to acquire new incoming edges, leading to a preferential attachment $\Pi(k)=(1+k)/N$. Consequently, the degree distribution follows a power law with $\gamma=2$. If p varies between 0 and 1, numerical simulations indicate a phase transition: for $p < p_c \approx 0.4$ the degree distribution decays exponentially, while for $p > p_c$ it has a power-law tail with γ very close to 2, the value corresponding to $p=1$. Thus, while the model does not explicitly include preferential attachment, the mechanism responsible for creation of the edges induces one.

4. Attaching to edges

Perhaps the simplest model of a scale-free network without explicit preferential attachment was proposed

by Dorogovtsev, Mendes, and Samukhin (2001a). In this model at every time step a new node connects to both ends of a randomly selected edge. Consequently the probability that a node will receive a new edge is directly proportional to its degree; in other words, this model has exactly the same preferential attachment as the Barabási-Albert model. It readily follows that the degree distribution has the same asymptotic form as the Barabási-Albert model, i.e., $P(k) \sim k^{-3}$.

The evolving network models presented in this section attempt to capture the mechanisms that govern the evolution of network topology (see Table III), guided by the information contained in the degree distribution. Less is known, however, about the clustering coefficients of these models. Notable exceptions are the models of Barabási *et al.* (2001; see also Sec. VIII.B) and Dorogovtsev, Mendes, and Samukhin (2001a; see also Sec. VIII.F). The clustering coefficient of the model of Barabási *et al.* displays a complex behavior as the network increases, first decreasing, going through a minimum, then increasing again, while the model of Dorogovtsev, Mendes, and Samukhin (2000d) has a constant asymptotic clustering coefficient. These results suggest that evolving network models can capture the high clustering coefficients of real networks.

G. Connection to other problems in statistical mechanics

The modeling of complex networks has offered fertile ground for statistical mechanics. Indeed, many advances in our understanding of the scaling properties of both small-world and evolving networks have benefited from concepts ranging from critical phenomena to nucleation theory and gelation. On the other hand, there appears to be another close link between statistical mechanics and evolving networks: the continuum theories proposed to predict the degree distribution can be mapped, often exactly, onto some well-known problems investigated in statistical physics. In the following we shall discuss two such mappings, relating evolving networks to the Simon model (Simon, 1955; see Amaral *et al.*, 2000; Bornholdt and Ebel, 2001) and to a Bose gas (Bianconi and Barabási, 2001b).

1. The Simon model

Aiming to account for the wide range of empirical distributions following a power law, such as the frequency of word occurrences (Zipf, 1949), the number of articles published by scientists (Lotka, 1926), the populations of cities (Zipf, 1949), or the distribution of incomes (Pareto, 1898), Simon (1955) proposed a class of stochastic models that result in a power-law distribution function. The simplest variant of the Simon model, described in terms of word frequencies, has the following algorithm: Consider a book that is being written and has reached a length of N words. Denote by $f_N(i)$ the number of different words that each occurred exactly i times in the text. Thus $f_N(1)$ denotes the number of different words that have occurred only once. The text is contin-

TABLE III. Summary of the mechanisms behind the current evolving network models. For each model (beyond the Barabási-Albert model) we list the concept or mechanism deviating from linear growth and preferential attachment, the two basic ingredients of the Barabási-Albert model, and the interval in which the exponent γ of the degree distribution can vary.

New concept or mechanism	Limits of γ	Reference
Linear growth, linear pref. attachment	$\gamma=3$	Barabási and Albert, 1999
Nonlinear preferential attachment $\Pi(k_i) \sim k_i^\alpha$	no scaling for $\alpha \neq 1$	Krapivsky, Redner, and Leyvraz, 2000
Asymptotically linear pref. attachment $\Pi(k_i) \sim a_\infty k_i$ as $k_i \rightarrow \infty$	$\gamma \rightarrow 2$ if $a_\infty \rightarrow \infty$ $\gamma \rightarrow \infty$ if $a_\infty \rightarrow 0$	Krapivsky, Redner, and Leyvraz, 2000
Initial attractiveness $\Pi(k_i) \sim A + k_i$	$\gamma=2$ if $A=0$ $\gamma \rightarrow \infty$ if $A \rightarrow \infty$	Dorogovtsev, Mendes, and Samukhin, 2000a, 2000b
Accelerating growth $\langle k \rangle \sim t^\theta$ constant initial attractiveness	$\gamma=1.5$ if $\theta \rightarrow 1$ $\gamma \rightarrow 2$ if $\theta \rightarrow 0$	Dorogovtsev and Mendes, 2001a
Accelerating growth $\langle k \rangle = at + 2b$	$\gamma=1.5$ for $k \ll k_c(t)$ $\gamma=3$ for $k \gg k_c(t)$	Barabási <i>et al.</i> , 2001 Dorogovtsev and Mendes, 2001c
Internal edges with probab. p	$\gamma=2$ if $q = \frac{1-p+m}{1+2m}$	
Rewiring of edges with probab. q	$\gamma \rightarrow \infty$ if $p, q, m \rightarrow 0$	Albert and Barabási, 2000
c internal edges or removal of c edges	$\gamma \rightarrow 2$ if $c \rightarrow \infty$ $\gamma \rightarrow \infty$ if $c \rightarrow 1$	Dorogovtsev and Mendes, 2000c
Gradual aging $\Pi(k_i) \sim k_i(t-t_i)^{-\nu}$	$\gamma \rightarrow 2$ if $\nu \rightarrow -\infty$ $\gamma \rightarrow \infty$ if $\nu \rightarrow 1$	Dorogovtsev and Mendes, 2000b
Multiplicative node fitness $\Pi_i \sim \eta_i k_i$	$P(k) \sim \frac{k^{-1-c}}{\ln(k)}$	Bianconi and Barabási, 2001a
Additive-multiplicative fitness $\Pi_i \sim \eta_i(k_i - 1) + \zeta_i$	$P(k) \sim \frac{k^{-1-m}}{\ln(k)}$ $1 \leq m \leq 2$	Ergün and Rodgers, 2001 Dorogovtsev, Mendes, and Samukhin, 2000c
Edge inheritance $P(k_{in}) = \frac{d}{k_{in}^2} \ln(ak_{in})$		
Copying with probab. p	$\gamma = (2-p)/(1-p)$	Kumar <i>et al.</i> , 2000a, 2000b
Redirection with probab. r	$\gamma = 1 + 1/r$	Krapivsky and Redner, 2001
Walking with probab. p	$\gamma \approx 2$ for $p > p_c$	Vázquez, 2000
Attaching to edges	$\gamma=3$	Dorogovtsev, Mendes, and Samukhin, 2001a
p directed internal edges $\Pi(k_i, k_j) \propto (k_i^{in} + \lambda)(k_j^{out} + \mu)$	$\gamma_{in} = 2 + p\lambda$ $\gamma_{out} = 1 + (1-p)^{-1} + \mu p/(1-p)$	Krapivsky, Rodgers, and Redner, 2001
$1-p$ directed internal edges Shifted linear pref. activity	$\gamma_{in} = 2 + p$ $\gamma_{out} \approx 2 + 3p$	Tadić, 2001a

ued by adding a new word. With probability p , this is a new word. However, with probability $1-p$, this word is already present. In this case Simon assumes that the probability that the $(N+1)$ th word has already appeared i times is proportional to $if_N(i)$, i.e., the total number of words that have occurred i times.

As noticed by Bornholdt and Ebel (2001), the Simon model can be mapped exactly onto the following network model: Starting from a small seed network, we record the number of nodes that have exactly k incoming edges, N_k . At every time step one of two processes can happen:

- (a) With probability p a new node is added, and a randomly selected node will point to the new node.
- (b) With probability $1-p$ a directed edge between two existing nodes is added. The starting point of this edge is selected randomly, while its end point is selected such that the probability that a node belonging to the N_k nodes with k incoming edges will be chosen is

$$\Pi(\text{class } k) \propto k N_k. \quad (126)$$

To appreciate the nature of this mapping, we need to clarify several issues:

- (1) While Eq. (126) represents a form of the “rich-get-richer” phenomenon, it does not imply preferential attachment [Eq. (78)] as used in various evolving network models. However, Eq. (78) implies Eq. (126). Thus the Simon model describes a general class of stochastic processes that can result in a power-law distribution, appropriate to capture Pareto and Zipf’s laws.
- (2) The interest in evolving network models comes from their ability to describe the topology of complex networks. The Simon model does not have an underlying network structure, as it was designed to describe events whose frequency follows a power law. Thus network measures going beyond the degree distribution, such as the average path length, spectral properties, or clustering coefficient, cannot be obtained from this mapping.
- (3) The mapping described above leads to a directed network with internal edges, different from the Barabási-Albert model. However, it is close to the model proposed by Dorogovtsev, Mendes, and Samukhin (2000a, 2000b) discussed in Sec. VIII.A.3, with the only difference being that here the initial attractiveness is present only for the isolated nodes. Since Eq. (126) corresponds to an asymptotically linear preferential attachment, a correspondence can be made with the model of Krapivsky, Redner, and Leyvraz (2000) as well.

2. Bose-Einstein condensation

Bianconi and Barabási (2001b) show the existence of a close link between evolving networks and an equilibrium Bose gas. Starting with the fitness model introduced in Sec. VIII.E.1, the mapping to a Bose gas can be done by assigning an energy ϵ_i to each node, determined by its fitness through the relation

$$\epsilon_i = -\frac{1}{\beta} \ln \eta_i, \quad (127)$$

where $\beta = 1/T$ plays the role of inverse temperature. An edge between two nodes i and j , having energies ϵ_i and ϵ_j , corresponds to two noninteracting particles, one on each energy level (see Fig. 30). Adding a new node l to the network corresponds to adding a new energy level ϵ_l and $2m$ new particles to the system. Half of these particles are deposited on the level ϵ_l (since all new edges

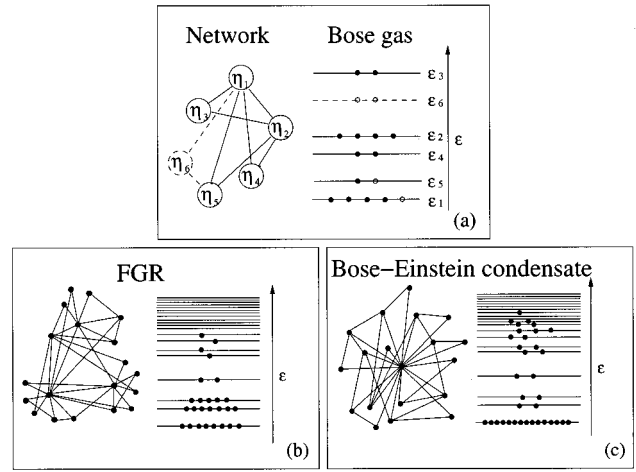


FIG. 30. Fitness and mapping between the network model and the Bose gas: (a) On the left we have a network of five nodes, each characterized by a fitness η_i . Equation (127) assigns an energy ϵ_i to each η_i (right). An edge from node i to node j corresponds to a particle at level ϵ_i and one at ϵ_j . The network evolves by adding a new node (dashed circle, η_6), which connects to $m=2$ other nodes (dashed lines), chosen following Eq. (78). In the gas this results in the addition of a new energy level (ϵ_6 , dashed) populated by $m=2$ new particles (\circ), and the deposition of $m=2$ other particles to energy levels to which the new node is connected (ϵ_2 and ϵ_5). (b) In the fit-get-rich (FGR) phase we have a continuous degree distribution, the several high-degree nodes linking the low-degree nodes together. In the energy diagram this corresponds to a decreasing occupation number with increasing energy. (c) In the Bose-Einstein condensate the fittest node attracts a finite fraction of all edges, corresponding to a highly populated ground level and sparsely populated higher energies. After Bianconi and Barabási (2001b).

start from the new node), while the other half are distributed between the energy levels of the end points of the new edges, the probability that a particle lands on level i being given by

$$\Pi_i = \frac{e^{-\beta \epsilon_i} k_i}{\sum_j e^{-\beta \epsilon_j} k_j}. \quad (128)$$

The continuum theory predicts that the rate at which particles accumulate on energy level ϵ_i is given by

$$\frac{\partial k_i(\epsilon_i, t, t_i)}{\partial t} = m \frac{e^{-\beta \epsilon_i} k_i(\epsilon_i, t, t_i)}{Z_t}, \quad (129)$$

where $k_i(\epsilon_i, t, t_i)$ is the occupation number of level i and Z_t is the partition function, defined as $Z_t = \sum_{j=1}^{\infty} t e^{-\beta \epsilon_j} k_j(\epsilon_j, t, t_j)$. The solution of Eq. (129) is

$$k_i(\epsilon_i, t, t_i) = m \left(\frac{t}{t_i} \right)^{f(\epsilon_i)}, \quad (130)$$

where the dynamic exponent $f(\epsilon)$ satisfies $f(\epsilon) = e^{-\beta(\epsilon - \mu)}$, μ plays the role of the chemical potential, satisfying the equation

$$\int \deg(\epsilon) \frac{1}{e^{\beta(\epsilon-\mu)} - 1} = 1, \quad (131)$$

and $\deg(\epsilon)$ is the degeneracy of the energy level ϵ . Equation (131) suggests that in the $t \rightarrow \infty$ limit the occupation number, giving the number of particles with energy ϵ , follows the familiar Bose statistics

$$n(\epsilon) = \frac{1}{e^{\beta(\epsilon-\mu)} - 1}. \quad (132)$$

The existence of the solution (130) depends on the functional form of the distribution $g(\epsilon)$ of the energy levels, determined by the $\rho(\eta)$ fitness distribution (see Sec. VIII.E.1). Specifically, if Eq. (131) has no non-negative solution for a given $g(\epsilon)$ and β , we can observe a Bose-Einstein condensation, indicating that a finite fraction of the particles condenses on the lowest energy level [see Fig. 30(c)].

This mapping to a Bose gas predicts the existence of two distinct phases as a function of the energy distribution. In the fit-get-rich phase, describing the case of uniform fitness discussed in Sec. VIII.E.1, the fitter nodes acquire edges at a higher rate than older but less fit nodes. In the end the fittest node will have the most edges, but the richest node is not an absolute winner, since its share of the edges (i.e., the ratio of its edges and the total number of edges in the network) decays to zero for large system sizes [Fig. 30(b)]. The unexpected outcome of this mapping is the possibility of Bose-Einstein condensation for $T < T_{BE}$, when the fittest node acquires a finite fraction of the edges and maintains this share of edges over time [Fig. 30(c)]. A representative fitness distribution that leads to condensation is $\rho(\eta) = (1 - \eta)^\lambda$ with $\lambda > 1$.

The temperature in Eq. (127) plays the role of a dummy variable, since if we define a fixed distribution $\rho(\eta)$, the existence of Bose-Einstein condensation or the fit-get-rich phase depends only on the functional form of $\rho(\eta)$ and is independent of β . Indeed, β falls out at the end from all topologically relevant quantities. As Dorogovtsev and Mendes (2001b) have subsequently shown, the existence of Bose-Einstein condensation can be derived directly from the fitness model, without employing the mapping to a Bose gas. While the condensation phenomenon appears to be similar to the gelation process observed by Krapivsky, Redner, and Leyvraz, (2000) in the case of superlinear preferential attachment, it is not clear at this point if this similarity is purely accidental or if there is a deeper connection between the fitness model and the fitness-free superlinear model.

IX. ERROR AND ATTACK TOLERANCE

Many complex systems display a surprising degree of tolerance for errors (Albert, Jeong, and Barabási, 2000). For example, relatively simple organisms grow, persist, and reproduce despite drastic pharmaceutical or environmental interventions, an error tolerance attributed to the robustness of the underlying metabolic and genetic

network (Jeong *et al.*, 2000; Jeong, Mason, *et al.*, 2001). Complex communication networks display a high degree of robustness: while key components regularly malfunction, local failures rarely lead to loss of the global information-carrying ability of the network. The stability of these and other complex systems is often attributed to the redundant wiring of their underlying network structure. But could the network topology, beyond redundancy, play a role in the error tolerance of such complex systems?

While error tolerance and robustness almost always have a dynamical component, here we shall focus only on the topological aspects of robustness, caused by edge and/or node removal. The first results regarding network reliability when subjected to edge removal came from random-graph theory (Moore and Shannon, 1956a, 1956b; Margulis, 1974; Bollobás, 1985). Consider an arbitrary connected graph H_N of N nodes, and assume that a p fraction of the edges have been removed. What is the probability that the resulting subgraph is connected, and how does it depend on the removal probability p ? For a broad class of starting graphs H_N (Margulis, 1974) there exists a threshold probability $p_c(N)$ such that if $p < p_c(N)$ the subgraph is connected, but if $p > p_c(N)$ it is disconnected. This phenomenon is in fact an inverse bond percolation problem defined on a graph, with the slight difference (already encountered in the evolution of a random graph) that the critical probability depends on N .

As the removal of a node implies the malfunctioning of all of its edges as well, node removal inflicts more damage than edge removal. Does a threshold phenomenon appear for node removal too? And to what degree does the topology of the network determine the network's robustness? In the following we shall call a network error tolerant (or robust) if it contains a giant cluster comprised of most of the nodes even after a fraction of its nodes are removed. The results indicate a strong correlation between robustness and network topology. In particular, scale-free networks are more robust than random networks against random node failures, but are more vulnerable when the most connected nodes are targeted (Albert, Jeong, and Barabási, 2000).

A. Numerical results

In the first study comparing the robustness of the Erdős-Rényi random graph and a scale-free network generated by the Barabási-Albert model, Albert, Jeong, and Barabási (2000) investigated networks that have the same number of nodes and edges, differing only in the degree distribution. Two types of node removal were considered. Random perturbations can cause the failure of some nodes; thus the first mechanism studied was the removal of randomly selected nodes. The second mechanism, in which the most highly connected nodes are removed at each step, was selected because it is the most damaging to the integrity of the system. This second choice emulates an intentional attack on the network.

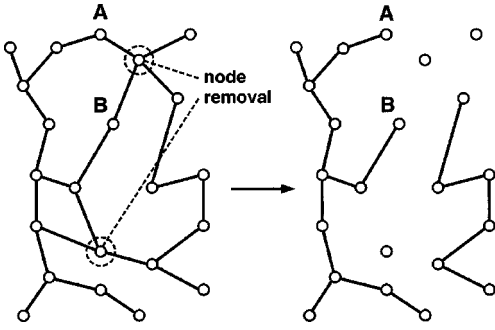


FIG. 31. Illustration of the effects of node removal on an initially connected network. In the unperturbed state the distance between nodes A and B is 2, but after two nodes are removed from the system, it increases to 6. At the same time the network breaks into five isolated clusters.

Let us start from a connected network, and at each time step remove a node. The disappearance of the node implies the removal of all edges that connect to it, disrupting some of the paths between the remaining nodes (Fig. 31). One way to monitor the disruption of an initially connected network is to study the relative size of the largest cluster that remains connected, S , and the average path length ℓ of this cluster, as a function of the fraction f of the nodes removed from the system. We expect that the size of the largest cluster will decrease and its average path length increase as an increasing number of nodes are removed from the network.

1. Random network, random node removal

We start by investigating the response of a random network to the random removal of its nodes [see Fig.

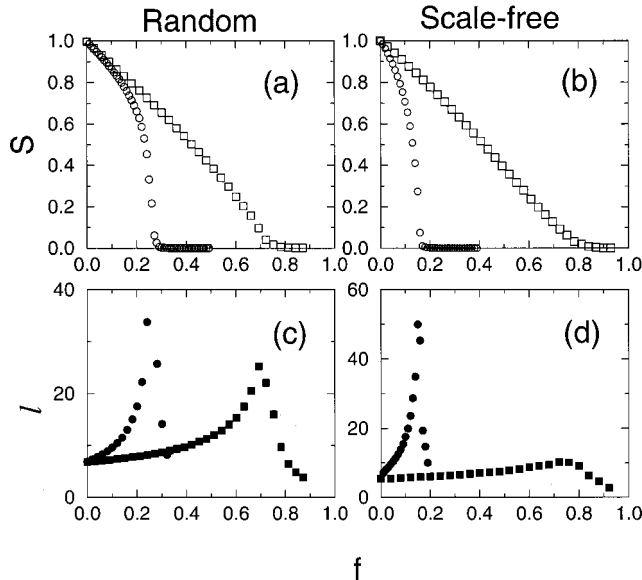


FIG. 32. The relative size S (a),(b) and average path length ℓ (c),(d) of the largest cluster in an initially connected network when a fraction f of the nodes are removed. (a),(c) Erdős-Rényi random network with $N=10\,000$ and $\langle k \rangle=4$; (b),(d) scale-free network generated by the Barabási-Albert model with $N=10\,000$ and $\langle k \rangle=4$. \square , random node removal; \circ , preferential removal of the most connected nodes. After Albert, Jeong, and Barabási (2000).

32(a), \square], looking at the changes in the relative size of the largest cluster S (i.e., the fraction of nodes contained in the largest cluster) and its average path length ℓ as an increasing number of nodes are randomly removed.

As expected, for a random network, the size S of the largest cluster decreases from $S=1$ as f increases. If only the removed nodes were missing from the largest cluster, S would follow the diagonal corresponding to $S=1$ for $f=0$ and $S=0$ for $f=1$. While for small f , S follows this line, as f increases the decrease becomes more rapid, indicating that clusters of nodes become isolated from the main cluster. At a critical fraction f_c , S drops to 0, indicating that the network breaks into tiny isolated clusters. These numerical results indicate an inverse percolation transition. Indeed, percolation theory can be used to calculate the critical fraction f_c (Sec. IX.B).

The behavior of the average path length ℓ also confirms this percolationlike transition: it starts from a value characteristic of an unperturbed random graph, increases with f as paths are disrupted in the network, and peaks at f_c [Fig. 32(c), filled squares]. After the network breaks into isolated clusters, ℓ decreases as well, since in this regime the size of the largest cluster decreases very rapidly.

When f is small we can use the prediction of random-graph theory, Eq. (16), indicating that ℓ scales as $\ln(SN)/\ln(\langle k \rangle)$, where $\langle k \rangle$ is the average degree of the largest cluster (Sec. IV.G). Since the number of edges decreases more rapidly than the number of nodes during the disruption of several edges, $\langle k \rangle$ decreases faster with increasing f than SN , and consequently ℓ increases. However, for $f=f_c$ the prediction of percolation theory becomes valid, and Eq. (44) indicates that ℓ no longer depends on $\langle k \rangle$ and decreases with S .

2. Scale-free network, random node removal

While a random network undergoes an inverse percolation transition when a critical fraction of its nodes are randomly removed, the situation is dramatically different for a Barabási-Albert network [Figs. 32(b) and (d), square datapoints]. Simulations indicate that while the size of the largest cluster decreases, it reaches 0 at a higher f . At the same time, ℓ increases much more slowly than in the random case, and its peak is much less prominent. The behavior of the system still suggests a percolation transition, but analytical calculations indicate that this is merely a finite size effect, and $f_c \rightarrow 1$ for a scale-free network as the size of the network increases (Sec. IX.B). In simple terms, scale-free networks display an exceptional robustness against random node failures.

3. Preferential node removal

In the case of an intentional attack, when the nodes with the highest number of edges are targeted, the network breaks down faster than in the case of random node removal. The general breakdown scenario again follows an inverse percolation transition, but now the critical fraction is much lower than in the random case.

This is understandable, since at every step the highest possible number of edges are removed from the system. Again, the two network topologies respond differently to attacks (Fig. 32, circular datapoints): the scale-free network, due to its reliance on the highly connected nodes, breaks down earlier than the random network.

In conclusion, numerical simulations indicate that scale-free networks display a topological robustness against random node failures. The origin of this error tolerance lies in their heterogeneous topology: low-degree nodes are far more abundant than nodes with high degree, so random node selection will more likely affect the nodes that play a marginal role in the overall network topology. But the same heterogeneity makes scale-free networks fragile to intentional attacks, since the removal of the highly connected nodes has a dramatic disruptive effect on the network.

B. Error tolerance: analytical results

The first analytical approach to calculating the critical threshold for fragmentation, f_c , of a network under random node failures was developed by Cohen *et al.* (2000). An alternative approach was proposed independently by Callaway *et al.* (2000). Cohen *et al.* (2000) argue that for a random network with a given degree distribution, f_c can be determined using the following criterion: a giant cluster, with size proportional to the size of the original network, exists if an arbitrary node i , connected to a node j in the giant cluster, is also connected to at least one other node. If i is connected only to j , the network is fragmented. If we assume that loops can be neglected (true for large fragmented systems) and use the Bayesian rules for conditional probabilities (see Cohen *et al.*, 2000), this criterion can be written as

$$\frac{\langle k^2 \rangle}{\langle k \rangle} = 2. \quad (133)$$

Consider a node with initial degree k_0 chosen from an initial distribution $P(k_0)$. After the random removal of a fraction f of the nodes, the probability that the degree of that node becomes k is $C_{k_0}^k (1-f)^k f^{k_0-k}$, and the new degree distribution is

$$P(k) = \sum_{k_0=k}^{\infty} P(k_0) C_{k_0}^k (1-f)^k f^{k_0-k}. \quad (134)$$

Thus the average degree and its second moment for the new system follows $\langle k \rangle = \langle k_0 \rangle (1-f)$ and $\langle k^2 \rangle = \langle k_0^2 \rangle (1-f)^2 + \langle k_0 \rangle f (1-f)$, allowing us to rewrite the criterion (133) for criticality as

$$f_c = 1 - \frac{1}{\frac{\langle k_0^2 \rangle}{\langle k_0 \rangle} - 1}, \quad (135)$$

where f_c is the critical fraction of removed nodes and $\langle k_0^2 \rangle, \langle k_0 \rangle$ are computed from the original distribution before the node removal.

As a test of the applicability of Eq. (133), let us remove a fraction f of the nodes from a random graph.

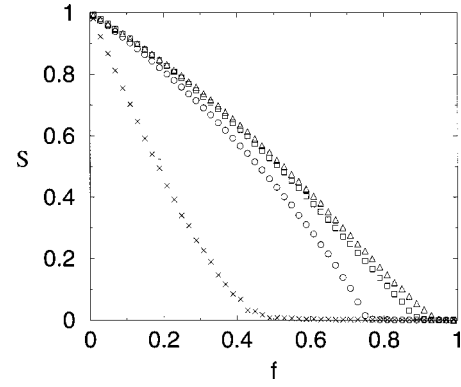


FIG. 33. The fraction of nodes in a giant cluster S as a function of the fraction of randomly removed nodes for scale-free random networks with $\gamma=3.5$ (\times) and $\gamma=2.5$ ($\circ, \square, \triangle$). In the latter case three different system sizes were used, with corresponding largest degree values K : \circ , $K=25$; \square , $K=100$; \triangle , $K=400$. The different curves illustrate that the fragmentation transition exists only for finite networks, while $f_c \rightarrow 1$ as $N \rightarrow \infty$. After Cohen *et al.* (2000).

Since in the original graph $k_0 = pN$ and $k_0^2 = (pN)^2 + pN$ (see Sec. III.C), Eq. (135) implies that $f_c = 1 - 1/(pN)$. If in the original system $\langle k_0^2 \rangle / \langle k_0 \rangle = 2$, meaning that $pN = \langle k \rangle = 1$ (the familiar condition for the appearance of the giant cluster in a random graph), the above equation indicates that $f_c = 0$, i.e., any amount of node removal leads to the network's fragmentation. The higher the original degree $\langle k_0 \rangle$ of the network, the larger the damage it can survive without breaking apart.

The critical probability is rather different for a scale-free networks. If the degree distribution follows a power law

$$P(k) = ck^{-\gamma}, \quad k = m, m+1, \dots, K, \quad (136)$$

where m and $K \approx mN^{1/\gamma-1}$ are the smallest and the largest degree values, respectively, using a continuum approximation valid in the limit $K \gg m \gg 1$, we obtain

$$\frac{\langle k_0^2 \rangle}{\langle k_0 \rangle} \rightarrow \frac{|2-\gamma|}{|3-\gamma|} \times \begin{cases} m & \text{if } \gamma > 3 \\ m^{\gamma-2} K^{3-\gamma} & \text{if } 2 < \gamma < 3 \\ K & \text{if } 1 < \gamma < 2. \end{cases} \quad (137)$$

We can see that for $\gamma > 3$ the ratio is finite and there is a transition at

$$f_c = 1 - \frac{1}{\frac{\gamma-2}{\gamma-3} m - 1} \quad (138)$$

(see Fig. 33).

However, for $\gamma < 3$ Eq. (137) indicates that the ratio diverges with K . For example, in the case $2 < \gamma < 3$,

$$f_c = 1 - \frac{1}{\frac{\gamma-2}{3-\gamma} m^{\gamma-2} K^{3-\gamma} - 1} \quad (139)$$

(Fig. 33), and thus $f_c \rightarrow 1$ when $N \rightarrow \infty$, true also for $\gamma < 2$. This result implies that infinite systems with $\gamma < 3$ do not break down under random failures, as a spanning

cluster exists for arbitrarily large f . In finite systems a transition is observed, although the transition threshold is very high. This result is in agreement with the numerical results discussed in the previous subsection (Albert, Jeong, and Barabási, 2000) indicating a delayed and very small peak in the ℓ curve for the failure of the Barabási-Albert model (having $\gamma=3$).

Callaway *et al.* (2000) investigate percolation on generalized random networks, considering that the occupation probability of nodes is coupled to the node degree. The authors use the method of generating functions discussed in Sec. V.B and generalize it to include the probability of occupancy of a certain node. The generating function for the degree distribution, corresponding to Eq. (46) in Sec. V.B, becomes

$$F_0(x) = \sum_{k=0}^{\infty} P(k) q_k x^k, \quad (140)$$

where q_k stands for the probability that a node with degree k is present. The overall fraction of nodes that are present in the network is $q = F_0(1)$, which is also equal to $1-f$ where f is the fraction of nodes missing from the system. This formulation includes the random occupancy (or conversely, random failure) case as the special case of uniform occupation probability $q_k = q$.

The authors consider random networks with a truncated power-law degree distribution

$$P(k) = \begin{cases} 0 & \text{for } k=0 \\ Ck^{-\gamma}e^{-k/\kappa} & \text{for } k \geq 1. \end{cases} \quad (141)$$

The exponential cutoff of this distribution has the role of regularizing the calculations in the same way as the largest degree K in the study of Cohen *et al.* (2000).

In the case of uniform occupation probability q corresponding to the random breakdown of a fraction $f=1-q$ of the nodes, the critical occupation probability follows

$$q_c = 1 - f_c = \frac{1}{\frac{Li_{\gamma-2}(e^{-1/\kappa})}{Li_{\gamma-1}(e^{-1/\kappa})} - 1}. \quad (142)$$

Here $Li_n(x)$ is the n th polylogarithm of x , defined as $Li_n(x) = \sum_{k=1}^{\infty} x^k/k^n$. This expression is similar to Eqs. (138) and (139) derived by Cohen *et al.* (2000). In the case of infinite network size we can take $\kappa \rightarrow \infty$, and the expression for the critical occupation probability becomes

$$q_c = \frac{1}{\frac{\zeta(\gamma-2)}{\zeta(\gamma-1)} - 1}, \quad (143)$$

where $\zeta(x)$ is the Riemann ζ function defined in the region $x > 1$; thus this expression is valid only for $\gamma > 3$. Since $\zeta(x) \rightarrow \infty$ as $x \rightarrow 1$, q_c becomes zero as γ approaches 3, indicating that for infinite scale-free networks even infinitesimal occupation probabilities can ensure the presence of an infinite cluster.

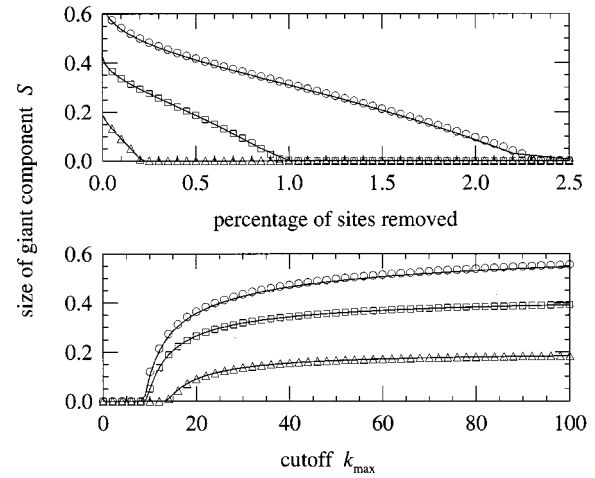


FIG. 34. Fraction of nodes in a spanning cluster in a scale-free random network with all nodes of degree greater than k_{\max} unoccupied: \circ , $\gamma=2.4$; \square , $\gamma=2.7$; \triangle , $\gamma=3.0$. The solid lines are the analytical prediction. Upper frame: as a function of f . Lower frame: as a function of the cutoff k_{\max} . After Callaway *et al.* (2000).

C. Attack tolerance: Analytical results

In the general framework of Callaway *et al.* (2000), an intentional attack targeting the nodes with the highest degree is equivalent to setting

$$q_k = \theta(k_{\max} - k) = \begin{cases} 1 & \text{if } k \leq k_{\max} \\ 0 & \text{if } k > k_{\max}. \end{cases} \quad (144)$$

This way only the nodes with degree $k \leq k_{\max}$ are occupied, which is equivalent to removing all nodes with $k > k_{\max}$. The number of removed nodes can be increased by lowering the value of k_{\max} . Callaway *et al.* (2000) calculate the fraction of nodes in the largest cluster S as a function of f and k_{\max} (Fig. 34). This figure is in agreement with the results of Albert, Jeong, and Barabási (2000) indicating that scale-free networks become fragmented after a small fraction f_c of highly connected nodes is removed. It also indicates that a small percentage of the most highly connected nodes can contain nodes with surprisingly low degree, agreeing also with the finding of Broder *et al.* (2000) that the World Wide Web is resilient to the removal of all nodes with degree higher than 5.

The theoretical framework of Cohen *et al.* (2000) can also be extended to the case of intentional attack on a scale-free network with degree distribution (136) (Cohen *et al.*, 2001). Under attack two things happen: (a) the cutoff degree K is reduced to a new value $\tilde{K} < K$, and (b) the degree distribution of the remaining nodes is changed. The new cutoff can be estimated from the relation

$$\sum_{k=\tilde{K}}^K P(k) = \sum_{k=\tilde{K}}^{\infty} P(k) - \frac{1}{N} = f, \quad (145)$$

which for large N implies

$$\tilde{K} = m f^{1/(1-\gamma)}. \quad (146)$$

The removal of a fraction f of the most connected nodes results in a random removal of a fraction \tilde{f} of edges from the remaining nodes. The probability that an edge leads to a deleted node equals the fraction of edges belonging to deleted nodes,

$$\tilde{f} = \frac{\sum_{k=\tilde{K}}^K kP(k)}{\langle k_0 \rangle} = f^{(2-\gamma)/(1-\gamma)}, \quad (147)$$

for $\gamma > 2$. We can see that in the limit $\gamma \rightarrow 2$ any nonzero f will lead to $\tilde{f} \rightarrow 1$ and thus to the breakdown of the whole network. Even in a finite network, where the upper cutoff of Eq. (145) is $K \simeq N$, in the limit $\gamma = 2$, $\tilde{f} = \ln(Nf/m)$, thus very small f values can lead to the destruction of a large fraction of the edges.

Since for random node deletion the probability of an edge's leading to a deleted node equals the fraction of deleted nodes, Cohen *et al.* (2001) argue that the network after undergoing an attack is equivalent to a scale-free network with cutoff \tilde{K} that has undergone random removal of a fraction \tilde{f} of its nodes. Replacing f with \tilde{f} and K with \tilde{K} in Eq. (135), we obtain the following equation for \tilde{K} :

$$\left(\frac{\tilde{K}}{m}\right)^{2-\gamma} - 2 = \frac{2-\gamma}{3-\gamma} m \left[\left(\frac{\tilde{K}}{m}\right)^{3-\gamma} - 1 \right]. \quad (148)$$

This equation can be solved numerically to obtain \tilde{K} as a function of m and γ , and $f_c(m, \gamma)$ can then be determined from Eq. (146). The results indicate that a breakdown phase transition exists for $\gamma > 2$, and f_c is very small for all γ values, on the order of a few percent. An interesting feature of the $f_c(\gamma)$ curve is that it has a maximum around $\gamma = 2.25$. It is not surprising that smaller γ values lead to increased vulnerability to attacks due to the special role the highly connected nodes play in connecting the system. However, Cohen *et al.* (2001) argue that the cause of the increased susceptibility of high γ networks is that for these even the original network is formed by several independent clusters, and the size of the largest cluster decreases with increasing γ . Indeed, the results of Aiello, Chung, and Lu (2000; see also Sec. V) indicate that for $2 < \gamma < 3.478$ the original network contains an infinite cluster and several smaller clusters of size at most $\ln N$, and for $\gamma > 3.478$ the original network has no infinite cluster.

D. The robustness of real networks

Systematic studies of the error and attack tolerance of real networks are available for three systems highly relevant to science and technology.

1. Communication networks

The error and attack tolerance of the Internet and the World Wide Web was investigated by Albert, Jeong, and Barabási (2000). Of the two networks, the Internet's robustness has more practical significance, as about 0.3%

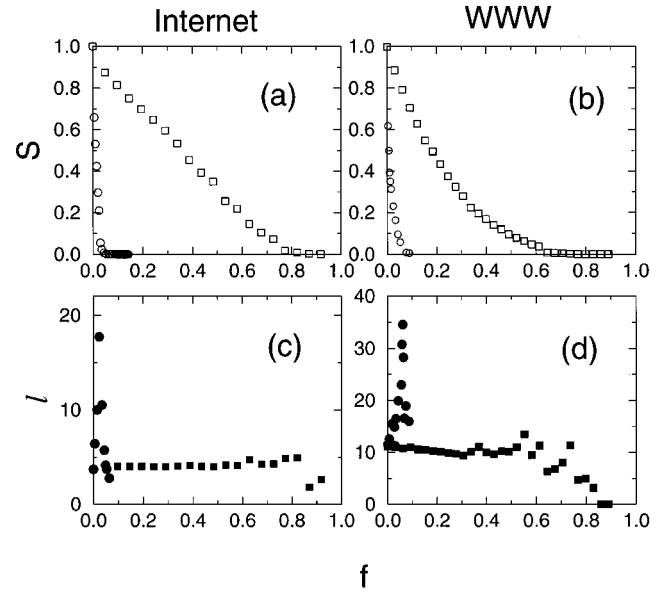


FIG. 35. The relative size S (a),(b) and average path length L (c),(d) of the largest cluster in two communication networks when a fraction f of the nodes are removed: (a),(c) Internet at the domain level, $N=6209$, $\langle k \rangle=3.93$; (b),(d) subset of the World Wide Web (WWW) with $N=325\,729$ and $\langle k \rangle=4.59$. \square , random node removal; \circ , preferential removal of the most connected nodes. After Albert, Jeong, and Barabási (2000).

of the routers regularly malfunction (random errors), and the Internet is occasionally subject to hacker attacks targeting some of the most connected nodes. The results, based on the latest map of the Internet topology at the interdomain (autonomous system) level, indicate that the average path length on the Internet is unaffected by the random removal of as many as 60% of the nodes, while if the most connected nodes are eliminated (attack), L peaks at a very small f [Fig. 35(a)]. Similarly, the large connected cluster persists for high rates of random node removal, but if nodes are removed in the attack mode, the size of the fragments that break off increases rapidly, the critical point appearing at a very small threshold, $f_c^I \simeq 0.03$ [Fig. 35(c)].

The World Wide Web study was limited to a subset of the web containing 325 729 nodes, the sample investigated in Albert, Jeong, and Barabási (1999). As the World Wide Web is directed, not all nodes can be reached from all nodes, even for the starting network. To resolve this problem, only distances between nodes that had a path between them were included in the average distance between nodes. Second, directed networks cannot be separated into clusters unambiguously: two nodes can be seen as part of the same cluster when starting from a certain node, yet they appear to be in separate clusters when starting from another. Hence the number of independent clusters was ambiguous, but the largest cluster could still be determined. Third, when simulating an attack on the World Wide Web, the nodes with the highest number of outgoing edges were removed, since k_{out} can be readily obtained by looking at a web document, while k_{in} can only be determined from

a map of the whole web. Despite these methodological differences, the response of the World Wide Web is similar to that of the undirected networks: after a slight initial increase, ℓ remains constant in the case of random failures [Fig. 35(b)], while it increases for attacks. The network survives as a large cluster under high rates of failure, but under attack the system abruptly falls apart at $f_c^v = 0.067$ [Fig. 35(d)].

2. Cellular networks

Cellular networks can be subject to random errors as a result of mutations or protein misfolding, as well as harsh external conditions eliminating essential metabolites. Jeong *et al.* (2000) studied the responses of the metabolic networks of several organisms to random and preferential node removal. Removing up to 8% of the substrates, they found that the average path length did not increase when nodes were removed randomly, but it increased rapidly upon removal of the most-connected nodes, attaining a 500% increase with the removal of only 8% of the nodes. Similar results have been obtained for the protein network of yeast as well (Jeong, Mason, *et al.*, 2001; see also Vogelstein, Lane, and Levine 2000).

3. Ecological networks

As a result of human actions or environmental changes, species are deleted from food webs, an issue of major concern for ecology and environmental science. Solé and Montoya (2001) studied the response of the food webs discussed in Sec. II to the removal of species (nodes; Montoya and Solé, 2000). The authors measured the relative size S of the largest cluster, the average size $\langle s \rangle$ of the rest of the species clusters, and the fraction of species becoming isolated due to the removal of other species on whom their survival depended (secondary extinctions). The results indicate that random species removal causes the fraction of species contained in the largest cluster to decrease linearly. At the same time the values of $\langle s \rangle$ remain 0 or 1, and the secondary extinction rates remain very low (smaller than 0.1) even when a high fraction of the nodes is removed. The estimate of Eq. (135) for the critical fraction at which the network fragments gives f_c^{fail} values around 0.95 for all networks, indicating that these networks are error tolerant. However, when the most connected (keystone) species are successively removed, S decays quickly and becomes zero at $f_c^{attack} \approx 0.2$, while $\langle s \rangle$ peaks. The secondary extinctions increase dramatically, reaching 1 at relatively low values of f ($f \approx 0.16$ for the Silwood Park web).

The results presented in this section offer a simple but compelling picture: scale-free networks display a high degree of robustness against random errors, coupled with a susceptibility to attacks. This double feature is the result of the heterogeneity of the network topology, encoded by the power-law degree distribution. While we focused on two measures only, S and ℓ , it is likely that most network measures will show distinct behavior for scale-free and random networks.

The types of disturbances we considered were static, that is, the removal of a node affected other nodes only in the topological sense. However, in many networks there is a dynamical aspect to error tolerance: the removal of a node could affect the functionality of other nodes as well. For example, the removal of a highly connected router on the Internet will redirect traffic to other routers that may not have the capacity to handle the increased traffic, creating an effective denial of service. Thus in many systems errors lead to cascading failures, affecting a large fraction of the network. While little is known about such events, Watts (2000) has recently shown that the network topology makes a big difference under cascading failures as well. He investigated a binary model in which the state of a node changed from off to on if a threshold fraction of its neighbors were on. In this model the probability that a perturbation in an initially all-off state will spread to the entire network can be connected to the existence of a giant cluster of vulnerable nodes. Using the method of generating functions, Watts (2000) showed that scale-free random graphs are much less vulnerable to random perturbations than are Erdős-Rényi random graphs with the same average degree.

It is often assumed that the robustness of many complex systems is rooted in their redundancy, which for networks represents the existence of many alternative paths that can preserve communication between nodes even if some nodes are absent. We are not aware of any research that would attempt to address this issue in quantitative terms, uncovering the degree to which redundancy plays a role.

X. OUTLOOK

The field of complex networks is rapidly evolving. While the potential for new and important discoveries is high, the field has attained a degree of coherence that made a review necessary and appropriate. The fact that the obtained results have reached a critical mass is best illustrated by the amount of work that had to be omitted from this review for lack of space. Being forced to make a choice, we focused on the mechanisms and models that describe network topology. In the following we briefly discuss some results that could not be covered in this approach but that are important for the field. In many ways work in these areas is as important as the work covered so far.

A. Dynamical processes on networks

Most networks offer support for various dynamical processes, and often the topology plays a crucial role in determining the system's dynamical features. The range of possible dynamical processes is wide. Watts (1999) studied the impact of clustering on several processes, including games, cooperation, the Prisoner's Dilemma, cellular automata, and synchronization (see also Lago-Fernández *et al.*, 2001). Wang and Chen (2001) have shown that the inhomogeneous scale-free topology plays

an important role in determining synchronization in a complex network, but search and random walks in complex networks is also a much investigated topic (Huberman *et al.*, 1998; Kleinberg, 2000; Adamic *et al.*, 2001; Bilke and Peterson, 2001; Burda *et al.*, 2001; Walsh, 2001). Modeling dynamics on a fixed topology is legitimate when the time scales describing the network topology and the dynamical process superposed to the network differ widely. A good example is Internet traffic, whose modeling requires time resolutions from milliseconds up to a day (Crovella and Bestavros, 1997; Willinger *et al.*, 1997; Solé and Valverde, 2001), compared with the months required for significant topological changes. Similarly, within a cell the concentrations of different chemicals change much faster than the cellular network topology (Savageau, 1998; Schilling and Palsen, 1998; Elowitz and Leibler, 2000; Gardner *et al.*, 2000), which is shaped by evolution over many generations.

The network structure plays a crucial role in determining the spread of ideas, innovations, or computer viruses (Coleman, Menzel, and Katz, 1957; Valente, 1995). In this light, spreading and diffusion has been studied on several types of networks regular (Kauffman, 1993; Keeling, 1999), random (Solomonoff and Rapoport, 1951; Rapoport, 1957; Weigt and Hartmann, 2001), small-world (Moukarzel, 1999; Newman and Watts, 1999a, 1999b; Moore and Newman, 2000a, 2000b; Newman, Moore, and Watts, 2000; Kuperman and Abramson, 2001), and scale-free (Johansen and Sornette, 2000; Bilke and Peterson, 2001; Tadić, 2001b; Watts, 2000). A particularly surprising result was offered recently by Pastor-Satorras and Vespignani (2001a, 2001b), who studied the effect of network topology on the spread of disease. They showed that while for random networks a local infection spreads to the whole network only if the spreading rate is larger than a critical value λ_c , for scale-free networks any spreading rate leads to the infection of the whole network. That is, for scale-free networks the critical spreading rate reduces to zero, a highly unexpected result that goes against volumes of particles written on this topic.

When the time scales governing the dynamics on the network are comparable to that characterizing the network assembly, the dynamical processes can influence the topological evolution. This appears to be the case in various biological models inspired by the evolution of communities or the emergence of the cellular topology (Slanina and Kotrla, 1999, 2000; Bornholdt and Sneppen, 2000; Hornquist, 2001; Jain and Krishna, 2001; Lässig *et al.*, 2001). In the current models these systems are often not allowed to “grow,” but they exist in a stationary state that gives room for diverse network topologies (Slanina and Kotrla, 1999, 2000). Interestingly, these models do not lead to scale-free networks in the stationary state, although it is known that cellular networks are scale free (Jeong *et al.*, 2000; Wagner and Fell, 2000; Jeong, Mason, *et al.*, 2001). Thus it is an open challenge to design evolutionary models that, based on selection

or optimization mechanisms, could produce topologies similar to those seen in the real world.

In general, when it comes to understanding the dynamics of networks, as well as the coupling between the dynamics and network assembly, we are only at the beginning of a promising journey (Strogatz, 2001). So far we lack simple organizing principles that would match the coherence and universality characterizing network topology. Due to the importance of the problem and the rapid advances we have witnessed in describing network topology, we foresee it as being a rapidly growing area.

B. Directed networks

Many important networks, including the World Wide Web or metabolic networks, have directed edges. In directed networks, however, not all nodes can be reached from a given node. This leads to a fragmented cluster structure in which the clusters are not unique, but depend on the starting point of the inquiry. Beyond some general aspects, little is known about such directed networks, but important insights could emerge in the near future. A promising step in this direction is the empirical study of the cluster structure of the World Wide Web (Broder *et al.*, 2000), finding that the web can be partitioned into several qualitatively different domains. The results indicate that 28% of the nodes are part of the strongly connected component, in which any pair of nodes is connected by paths in either direction. Another 23% of the nodes can be reached from the strongly connected component but cannot connect to it in the other direction, while a roughly equal fraction of the nodes have paths leading to the strongly connected component but cannot be reached from it. As several groups have pointed out, this structure is not specific to the World Wide Web but is common to all directed networks, ranging from cell metabolism to citation networks (Newman, Strogatz, and Watts, 2000; Dorogovtsev, Mendes, and Samukhin, 2001b).

Most network models (including small-world and evolving networks) ignore the network’s directedness. However, as the World Wide Web measurements have shown, incoming and outgoing edges could follow different scaling laws. In this respect, the Barabási-Albert model (Barabási and Albert, 1999) explains only the incoming degree distribution, as, due to its construction, each node has exactly m outgoing edges; thus the outgoing degree distribution is a delta function (Sec. VII.A). While several models have recently investigated directed evolving networks, obtaining a power law for both outgoing and incoming edges (Krapivsky, Rodgers, and Redner, 2001; Tadić, 2001a, 2001b). The generic features of such complex directed models could hold further surprises.

C. Weighted networks, optimization, allometric scaling

Many real networks are weighted networks, in contrast with the binary networks investigated so far, in which the edge weights can have only two values 0 and 1

(absent or present). Indeed, in social networks it is often important to assign a strength to each acquaintance edge, indicating how well the two individuals know each other (Newman, 2001b, 2001c). Similarly, cellular networks are characterized by reaction rates, and the edges on the Internet by bandwidth. What are the mechanisms that determine these weights? Do they obey nontrivial scaling behavior? To what degree are they determined by the network topology? Most answers to these questions come from two directions: theoretical biology and ecology, concerned with issues related to allometric scaling, and random resistor networks (Derrida and Vannimenus, 1982; Duxbury, Beale, and Moukarzel, 1995), a topic much studied in statistical mechanics. *Allometric scaling* describes the transport of material through the underlying network characterizing various biological systems. Most of these systems have a branching, tree-like topology. The combination of the tree topology with the desire to minimize the cost of transportation leads to nontrivial scaling in the weights of the edges (West *et al.*, 1997; Enquist *et al.*, 1998, 1999).

In a more general context, Banavar and collaborators have shown that when the aim is to minimize the cost of transportation, the optimal network topology can vary widely, ranging from treelike structures to spirals or loop-dominated highly interconnected networks (Banavar *et al.*, 1999, 2000). Beyond giving systematic methods and principles to predict the topology of transportation networks, these studies raise some important questions that need to be addressed in the future. For example, to what degree is the network topology shaped by global optimization, or the local processes seen in scale-free networks? There are fundamental differences between transportation and evolving networks. In transportation models the network topology is determined by a global optimization process, in which edges are positioned to minimize, over the whole network, some predefined quantity, such as cost or energy of transportation. In contrast, for evolving networks such global optimization is absent, as the decision about where to link is delegated to the node level. However, this decision is not entirely local in scale-free networks either, as the node has information about the degree of all nodes in the network, from which it chooses one following Eq. (78), the normalization factor making the system fully coupled. The interplay between such local and global optimization processes is far from being fully understood (Carlson and Doyle, 1999, 2000; Doyle and Carlson, 2000).

While edge weights are well understood for trees and some much-studied physical networks, ranging from river networks (Banavar *et al.*, 1997; Rodríguez-Iturbe and Rinaldo, 1997) to random resistor networks (Derrida and Vannimenus, 1982; Duxbury, Beale, and Moukarzel, 1995), little work has been done on these problems in the case of small-world or scale-free networks. Recently Yook *et al.* (2001) have investigated an evolving network model in which the weights were added dynamically, resulting in unexpected scaling behavior. Newman (2001b) has also assigned weights to

characterize the collaboration strength between scientists. These studies make an important point, however: despite the practical relevance and potential phenomenological richness, the understanding of weighted networks is still in its infancy.

D. Internet and World Wide Web

A few real networks, with high technological or intellectual importance, have received special attention. In these studies the goal is to develop models that go beyond the basic growth mechanisms and incorporate the specific and often unique details of a given system. Along these lines much attention has focused on developing realistic World Wide Web models that explain everything from the average path length to incoming and outgoing degree distribution (Adamic and Huberman, 1999; Flake *et al.*, 2000; Krapivsky, Rodgers, and Redner, 2001; Tadić, 2001a). Many studies focus on the identification of web communities as well, representing clusters of nodes that are highly connected to each other (Gibson *et al.*, 1998; Adamic and Adar, 2000; Flake *et al.*, 2000; Pennock *et al.*, 2000).

There is a race in computer science to create good Internet topology generators (Paxson and Floyd, 1997; Comellas *et al.*, 2000). New Internet protocols are tested on model networks before their implementation, and protocol optimization is sensitive to the underlying network topology (Labovitz *et al.*, 2000). Prompted by the discovery that the Internet is a scale-free network, all topology generators are being reviewed and redesigned. These studies have resulted in careful investigations into what processes could contribute to the correct topology, reaffirming that growth and preferential attachment are necessary conditions for realistic Internet models (Medina *et al.*, 2000; Palmer and Steffan, 2000; Jeong, Nédá, and Barabási, 2001; Pastor-Satorras *et al.*, 2001; Yook, Jeong, and Barabási, 2001b). In addition, an interesting link has recently been found (Caldarelli *et al.*, 2000) to river networks, a much-studied topic in statistical mechanics (see Banavar *et al.*, 1999; Dodds and Rothman 2000, 2001a, 2001b, 2001c).

E. General questions

The high interest in scale-free networks might give the impression that all complex networks in nature have power-law degree distributions. As we discussed in Sec. II, that is far from being the case. It is true that several complex networks of high interest for the scientific community, such as the World Wide Web, cellular networks, the Internet, some social networks, and the citation network, are scale free. However, others, such as the power grid or the neural network of *C. elegans*, appear to be exponential. Does that mean that they are random? Far from it. These systems are best described as evolving networks. As we have seen in many examples in Sec. VIII, evolving networks can develop both power-law and exponential degree distributions. While the power-law regime appears to be robust, sublinear preferential

attachment, aging effects, and growth constraints lead to crossovers to exponential decay. Thus, while evolving networks are rather successful at describing a wide range of systems, the functional form of $P(k)$ cannot be guessed until the microscopic details of the network evolution are fully understood. If all processes shaping the topology of a certain network are properly incorporated, the resulting $P(k)$ often has a rather complex form, described by a combination of power laws and exponentials.

In critical phenomena we are accustomed to unique scaling exponents that characterize complex systems. Indeed, the critical exponents are uniquely determined by robust factors, such as the dimension of the space or conservation laws (Stanley, 1971; Ma, 1976; Hohenberg and Halperin, 1977). The most studied exponents in terms of evolving networks are the dynamic exponent β and the degree exponent γ . While the former characterizes the network dynamics, the latter is a measure of the network topology. The inseparability of the topology and dynamics of evolving networks is shown by the fact that these exponents are related by the scaling relation (86) (Dorogovtsev, Mendes, and Samukhin, 2000a), underlying the fact that a network's assembly uniquely determines its topology. However, in no case are these exponents unique. They can be tuned continuously by such parameters as the frequency of internal edges, rewiring rates, initial node attractiveness, and so on. While it is difficult to search for universality in the value of the exponents, this does not imply that the exponents are not uniquely defined. Indeed, if all processes contributing to the network assembly and evolution are known, the exponents can be calculated exactly. But they do not assume the discrete values we are accustomed to in critical phenomena.

Some real networks have an underlying bipartite structure (Sec. V.D). For example, the actor network can be represented as a graph consisting of two types of nodes: actors and movies, the edges always connecting two nodes of different types. These networks can be described as generalized random graphs (Newman, Strogatz, and Watts, 2001). It is important to note, however, that both subsets of these bipartite graphs are growing in time. While it has not yet been attempted, the theoretical methods developed for evolving networks can be generalized for bipartite networks as well, leading to coupled continuum equations. We expect that extending these methods, whenever appropriate, would lead to a much more realistic description of several real systems.

The classical thinking on complex networks, rooted in percolation and random-graph theory (see Aldous, 1999), is that they appear as a result of a percolation process in which isolated nodes eventually join a giant cluster as the number of edges increases between them. Thus a much-studied question concerns the threshold at which the giant cluster appears. With a few exceptions (Callaway *et al.*, 2001), evolving networks do not follow this percolation picture, since they are connected from

their construction. Naturally, if node or edge removal is allowed, percolation-type questions do emerge (Sec. IX).

F. Conclusions

The shift that we have experienced in the past three years in our understanding of networks was swift and unexpected. We have learned through empirical studies, models, and analytic approaches that real networks are far from being random, but display generic organizing principles shared by rather different systems. These advances have created a prolific branch of statistical mechanics, followed with equal interest by sociologists, biologists, and computer scientists. Our goal here was to summarize, in a coherent fashion, what is known so far. Yet we believe that these results are only the tip of the iceberg. We have uncovered some generic topological and dynamical principles, but the answers to the open questions could hide new concepts and ideas that might turn out to be just as exciting as those we have encountered so far. The future could bring new tools as well, as the recent importation of ideas from field theory (Burda *et al.*, 2001) and quantum statistics (Bianconi, 2000a, 2001; Bianconi and Barabási, 2001b; Zizzi, 2001) indicates. Consequently this article is intended to be as much a review as a catalyst for further advances. We hope that the latter aspect will dominate.

ACKNOWLEDGMENTS

We wish to thank István Albert, Alain Barrat, Ginestra Bianconi, Duncan Callaway, Reuven Cohen, Ramesh Govindan, José Mendes, Christian Moukarzel, Zoltán Nédá, Mark Newman, Steven Strogatz, Andrew Tomkins, Duncan Watts, and Altavista for allowing us to reproduce their figures. We are grateful to Luis N. Amaral, Alain Barrat, Duncan Callaway, Reuven Cohen, Imre Derényi, Sergei Dorogovtsev, Illés Farkas, Shlomo Havlin, Miroslav Kotrla, Paul Krapivsky, José Mendes, Christian Moukarzel, Mark Newman, Sidney Redner, Frantisek Slanina, Ricard Solé, Steven Strogatz, Bosiljka Tadić, Tamás Vicsek, and Duncan Watts for reading our manuscript and providing helpful suggestions. We have benefited from discussions with István Albert, Ginestra Bianconi, Zoltán Dezső, Hawoong Jeong, Zoltán Nédá, Erzsébet Ravasz, and Soon-Hyung Yook. This research was supported by NSF-DMR-9710998 and NSF-PHYS-9988674.

REFERENCES

- Abello, J., P. M. Pardalos, and M. G. C. Resende, 1999, in *External Memory Algorithms*, edited by J. Abello and J. Vitter, DIMACS Series in Discrete Mathematics Theoretical Computer Science (American Mathematical Society), p. 119.
- Adamic, L. A., 1999, *Proceedings of the Third European Conference, ECDL'99* (Springer-Verlag, Berlin), p. 443.
- Adamic, L. A., and E. Adar, 2000, preprint, www.hpl.hp.com/shl/papers/web10/index.html

- Adamic, L. A., and B. A. Huberman, 1999, *Nature* (London) **401**, 131.
- Adamic, L. A., and B. A. Huberman, 2000, *Science* **287**, 2115.
- Adamic, L. A., R. M. Lukose, A. R. Puniyani, and B. A. Huberman, 2001, *Phys. Rev. E* **64**, 046135.
- Aiello, W., F. Chung, and L. Lu, 2000, *Proceedings of the 32nd ACM Symposium on the Theory of Computing* (ACM, New York), p. 171.
- Albert, R., and A.-L. Barabási, 2000, *Phys. Rev. Lett.* **85**, 5234.
- Albert, R., H. Jeong, and A.-L. Barabási, 1999, *Nature* (London) **401**, 130.
- Albert, R., H. Jeong, and A.-L. Barabási, 2000, *Nature* (London) **406**, 378; 2001, **409**, 542(E).
- Aldous, D., 1999, *Bernoulli* **5**, 3.
- Amaral, L. A. N., M. Barthélemy, and P. Gopikrishnan, 1999, unpublished.
- Amaral, L. A. N., A. Scala, M. Barthélemy, and H. E. Stanley, 2000, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11 149.
- Ambjorn, J., B. Durhuus, and T. Jonsson, 1990, *Phys. Lett. B* **244**, 403.
- Argollo de Menezes, M., C. F. Moukarzel, and T. J. P. Penna, 2000, *Europhys. Lett.* **50**, 574.
- Banavar, J. R., F. Colaiori, A. Flammini, A. Giacometti, A. Maritan, and A. Rinaldo, 1997, *Phys. Rev. Lett.* **78**, 4522.
- Banavar, J. R., F. Colaiori, A. Flammini, A. Maritan, and A. Rinaldo, 2000, *Phys. Rev. Lett.* **84**, 4745.
- Banavar, J. R., A. Maritan, and A. Rinaldo, 1999, *Nature* (London) **399**, 130.
- Barabási, A.-L., and R. Albert, 1999, *Science* **286**, 509.
- Barabási, A.-L., R. Albert, and H. Jeong, 1999, *Physica A* **272**, 173.
- Barabási, A.-L., H. Jeong, E. Ravasz, Z. Nédá, A. Schubert, and T. Vicsek, 2001, preprint cond-mat/0104162.
- Barrat, A., 1999, e-print cond-mat/9903323.
- Barrat, A., and M. Weigt, 2000, *Eur. Phys. J. B* **13**, 547.
- Barthélemy, M., and L. A. N. Amaral, 1999, *Phys. Rev. Lett.* **82**, 3180; **82**, 5180(E).
- ben Avraham, D., and S. Havlin, 2000, *Diffusion and Reactions in Fractals and Disordered Systems* (Cambridge University Press, Cambridge/New York).
- Bianconi, G., 2000a, *Int. J. Mod. Phys. B* **14**, 3356.
- Bianconi, G., 2000b, unpublished.
- Bianconi, G., 2001, *Int. J. Mod. Phys. B* **15**, 313.
- Bianconi, G., and A.-L. Barabási, 2001a, *Europhys. Lett.* **54**, 436.
- Bianconi, G., and A.-L. Barabási, 2001b, *Phys. Rev. Lett.* **86**, 5632.
- Bilke, S., and C. Peterson, 2001, *Phys. Rev. E* **64**, 036106.
- Bollobás, B., 1981, *Discrete Math.* **33**, 1.
- Bollobás, B., 1984, *Trans. Am. Math. Soc.* **286**, 257.
- Bollobás, B., 1985, *Random Graphs* (Academic, London).
- Bollobás, B., and O. Riordan, 2001, "The diameter of a scale-free random graph," preprint.
- Bornholdt, S., and H. Ebel, 2001, *Phys. Rev. E* **64**, 035104(R).
- Bornholdt, S., and K. Sneppen, 2000, *Proc. R. Soc. London, Ser. B* **267**, 2281.
- Broder, A., R. Kumar, F. Maghoul, P. Raghavan, S. Rajalopagan, R. Stata, A. Tomkins, and J. Wiener, 2000, *Comput. Netw.* **33**, 309.
- Bunde, A., and S. Havlin, 1994, Eds., *Fractals in Science* (Springer, Berlin).
- Bunde, A., and S. Havlin, 1996, Eds., *Fractals and Disordered Systems* (Springer, Berlin).
- Burda, Z., J. D. Correia, and A. Krzywicki, 2001, *Phys. Rev. E* **64**, 046118.
- Burton, R. M., and M. Keane, 1989, *Commun. Math. Phys.* **121**, 501.
- Caldarelli, G., R. Marchetti, and L. Pietronero, 2000, *Europhys. Lett.* **52**, 386.
- Callaway, D. S., J. E. Hopcroft, J. M. Kleinberg, M. E. J. Newman, and S. H. Strogatz, 2001, *Phys. Rev. E* **64**, 041902.
- Callaway, D. S., M. E. J. Newman, S. H. Strogatz, and D. J. Watts, 2000, *Phys. Rev. Lett.* **85**, 5468.
- Camacho, J., R. Guimerà, and L. A. N. Amaral, 2001a, preprint cond-mat/0102127.
- Camacho, J., R. Guimerà, and L. A. N. Amaral, 2001b, preprint cond-mat/0103114.
- Carlson, J. M., and J. Doyle, 1999, *Phys. Rev. E* **60**, 1412.
- Carlson, J. M., and J. Doyle, 2000, *Phys. Rev. Lett.* **84**, 2529.
- Christensen, K., R. Donangelo, B. Koiler, and K. Sneppen, 1998, *Phys. Rev. Lett.* **81**, 2380.
- Chung, F., and L. Lu, 2001, *Adv. Appl. Math.* **26**, 257.
- Cohen, J. E., 1988, *Discrete Appl. Math.* **19**, 113.
- Cohen, R., K. Erez, D. ben-Avraham, and S. Havlin, 2000, *Phys. Rev. Lett.* **85**, 4626.
- Cohen, R., K. Erez, D. ben-Avraham, and S. Havlin, 2001, *Phys. Rev. Lett.* **86**, 3682.
- Coleman, J. S., H. Menzel, and E. Katz, 1957, *Sociometry* **20**, 253.
- Comellas, F., J. Ozon, and J. G. Peters, 2000, *Inf. Process. Lett.* **76**, 83.
- Crisanti, A., G. Paladin, and A. Vulpiani, 1993, *Products of Random Matrices in Statistical Physics* (Springer, Berlin).
- Crovella, M. E., and A. Bestavros, 1997, *IEEE/ACM Trans. Netw.* **5**, 835.
- Derrida, B., and J. Vannimenus, 1982, *J. Phys. A* **15**, L557.
- Dodds, P. S., and D. H. Rothman, 2000, *Annu. Rev. Earth Planet Sci.* **28**, 571.
- Dodds, P. S., and D. H. Rothman, 2001a, *Phys. Rev. E* **63**, 016115.
- Dodds, P. S., and D. H. Rothman, 2001b, *Phys. Rev. E* **63**, 016116.
- Dodds, P. S., and D. H. Rothman, 2001c, *Phys. Rev. E* **63**, 016117.
- Dorogovtsev, S. N., and J. F. F. Mendes, 2000a, *Europhys. Lett.* **50**, 1.
- Dorogovtsev, S. N., and J. F. F. Mendes, 2000b, *Phys. Rev. E* **62**, 1842.
- Dorogovtsev, S. N., and J. F. F. Mendes, 2000c, *Europhys. Lett.* **52**, 33.
- Dorogovtsev, S. N., and J. F. F. Mendes, 2001a, *Phys. Rev. E* **63**, 025101.
- Dorogovtsev, S. N., and J. F. F. Mendes, 2001b, *Phys. Rev. E* **63**, 056125.
- Dorogovtsev, S. N., and J. F. F. Mendes, 2001c, *Proc. R. Soc. London Ser. B* **268**, 2603.
- Dorogovtsev, S. N., J. F. F. Mendes, and A. N. Samukhin, 2000a, *Phys. Rev. Lett.* **85**, 4633.
- Dorogovtsev, S. N., J. F. F. Mendes, and A. N. Samukhin, 2000b, preprint cond-mat/0009090.
- Dorogovtsev, S. N., J. F. F. Mendes, and A. N. Samukhin, 2001c, preprint cond-mat/0011077.
- Dorogovtsev, S. N., J. F. F. Mendes, and A. N. Samukhin, 2001a, *Phys. Rev. E* **63**, 056125.
- Dorogovtsev, S. N., J. F. F. Mendes, and A. N. Samukhin, 2001b, *Phys. Rev. E* **64**, R025101.

- Doyle, J., and J. M. Carlson, 2000, *Phys. Rev. Lett.* **84**, 5656.
- Durrett, R. T., 1985, *Z. Wahrscheinlichkeitstheor. Verwandte Geb.* **69**, 421.
- Duxbury, P. M., P. D. Beale, and C. Moukarzel, 1995, *Phys. Rev. B* **51**, 3476.
- Elowitz, M. B., and S. Leibler, 2000, *Nature (London)* **403**, 335.
- Enquist, B. J., J. H. Brown, and G. B. West, 1998, *Nature (London)* **395**, 163.
- Enquist, B. J., G. B. West, E. L. Charnov, and J. H. Brown, 1999, *Nature (London)* **401**, 907.
- Erdős, P., and A. Rényi, 1959, *Publ. Math. (Debrecen)* **6**, 290.
- Erdős, P., and A. Rényi, 1960, *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17.
- Erdős, P., and A. Rényi, 1961, *Bull. Inst. Int. Stat.* **38**, 343.
- Ergün, G., and G. J. Rodgers, 2001, preprint cond-mat/0103423.
- Faloutsos, M., P. Faloutsos, and C. Faloutsos, 1999, *Comput. Commun. Rev.* **29**, 251.
- Farkas, I. J., I. Derényi, A.-L. Barabási, and T. Vicsek, 2001, *Phys. Rev. E* **64**, 026704.
- Fell, D. A., and A. Wagner, 2000, *Nat. Biotechnol.* **18**, 1121.
- Ferrer i Cancho, R., and R. V. Solé, 2001, Santa Fe Institute working paper 01-03-016.
- Flake, G., S. Lawrence, and C. L. Giles, 2000, *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, MA, August 2000 (ACM, New York), p. 150.
- Gardner, T. S., C. R. Cantor, and J. J. Collins, 2000, *Nature (London)* **403**, 520.
- Gibson, D., J. Kleinberg, and P. Raghavan, 1998, *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, New York, 1998 (ACM, New York), p. 225.
- Gleiss, P. M., P. F. Stadler, A. Wagner, and D. Fell, 2001, *Adv. Complex Syst.* **4**, 207.
- Goh, K.-I., B. Kahng, and D. Kim, 2001, *Phys. Rev. E* **64**, 051903.
- Govindan, R., and H. Tangmunarunkit, 2000, in *Proceedings of IEEE INFOCOM 2000*, Tel Aviv, Israel (IEEE, Piscataway, N.J.), Vol. 3, p. 1371.
- Granovetter, M. S., 1973, *Am. J. Sociol.* **78**, 1360.
- Grimmett, G., 1999, *Percolation* (Springer, Berlin).
- Guhr, T., A. Müller-Groeling, and H. A. Weidenmüller, 1998, *Phys. Rep.* **299**, 189.
- Guimerà, R., A. Arenas, and A. Díaz-Guilera, 2001, *Physica A* **299**, 247.
- Hammersley, J. M., 1957, *Ann. Math. Stat.* **28**, 790.
- Hara, T., and G. Slade, 1990, *Commun. Math. Phys.* **128**, 333.
- Hartwell, L. H., J. J. Hopfield, S. Leibler, and A. W. Murray, 1999, *Nature (London)* **402**, C47.
- Havlin, S., and R. Nossal, 1984, *J. Phys. A* **17**, L427.
- Hohenberg, P. C., and B. I. Halperin, 1977, *Rev. Mod. Phys.* **49**, 435.
- Hornquist, M., 2001, preprint nlin.AO/0104016.
- Hubermann, B. A., P. L. T. Pioroli, J. E. Pitkow, and R. M. Lukose, 1998, *Science* **280**, 95.
- Jain, S., and S. Krishna, 2001, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 543.
- Jeong, H., S. P. Mason, Z. N. Oltvai, and A.-L. Barabási, 2001, *Nature (London)* **411**, 41.
- Jeong, H., Z. Nédá, and A.-L. Barabási, 2001, preprint cond-mat/0104131.
- Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, 2000, *Nature (London)* **407**, 651.
- Jespersen, S., and A. Blumen, 2000, *Phys. Rev. E* **62**, 6270.
- Johansen, A., and D. Sornette, 2000, *Physica A* **276**, 338.
- Karonski, M., and A. Ruciński, 1997, in *The Mathematics of Paul Erdős*, edited by R. L. Graham and J. Nešetřil (Springer, Berlin).
- Kasturirangan, R., 1999, e-print cond-mat/9904419.
- Kauffman, S. A., 1993, *The Origins of Order* (Oxford University, New York).
- Kauffman, S. A., 1995, *At Home in the Universe: The Search for Laws of Self-Organization and Complexity* (Oxford University, New York).
- Keeling, M. J., 1999, *Proc. R. Soc. London, Ser. B* **266**, 859.
- Kleinberg, J. M., 2000, *Nature (London)* **406**, 845.
- Kleinberg, J. M., R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, 1999, *Proceedings of the 5th Annual International Conference, COCOON'99*, Tokyo, July 1999 (Springer-Verlag, Berlin), p. 1.
- Kochen, M., 1989, Ed., *The Small World* (Ablex, Norwood, NJ).
- Kolchin, V. F., 1986, *Theor. Probab. Appl.* **31**, 439.
- Krapivsky, P. L., and S. Redner, 2001, *Phys. Rev. E* **63**, 066123.
- Krapivsky, P. L., S. Redner, and F. Leyvraz, 2000, *Phys. Rev. Lett.* **85**, 4629.
- Krapivsky, P. L., G. J. Rodgers, and S. Redner, 2001, *Phys. Rev. Lett.* **86**, 5401.
- Kulkarni, R. V., E. Almaas, and D. Stroud, 2000, *Phys. Rev. E* **61**, 4268.
- Kullmann, L., and J. Kertész, 2001, *Phys. Rev. E* **63**, 051112.
- Kumar, R., P. Raghavan, S. Rajalopagan, D. Sivakumar, A. S. Tomkins, and E. Upfal, 2000a, *Proceedings of the 19th Symposium on Principles of Database Systems*, p. 1.
- Kumar, R., P. Raghavan, S. Rajalopagan, D. Sivakumar, A. S. Tomkins, and E. Upfal, 2000b, *Proceedings of the 41st IEEE Symposium on Foundations of Computer Science* (IEEE Computing Society, Los Alamitos, Calif.), p. 57.
- Kumar, R., P. Raghavan, S. Rajalopagan, and A. Tomkins, 1999, *Proceedings of the 9th ACM Symposium on Principles of Database Systems*, p. 1.
- Kuperman, M., and G. Abramson, 2001, *Phys. Rev. Lett.* **86**, 2909.
- Labovitz, C., R. Wattenhofer, S. Venkatachar y, and A. Ahuja, 2000, Technical Report MSR-TR-2000-74, Microsoft Research.
- Lago-Fernández, L. F., R. Huerta, F. Corbacho, and J. A. Sigüenza, 2000, *Phys. Rev. Lett.* **84**, 2758.
- Lässig, M., U. Bastolla, S. C. Manrubia, and A. Valleriani, 2001, *Phys. Rev. Lett.* **86**, 4418.
- Lawrence, S., and C. L. Giles, 1998, *Science* **280**, 98.
- Lawrence, S., and C. L. Giles, 1999, *Nature (London)* **400**, 107.
- Leath, P. L., 1976, *Phys. Rev. B* **14**, 5064.
- Liljeros, F., C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aberg, 2001, *Nature (London)* **411**, 907.
- Lotka, A. J., 1926, *J. Wash. Acad. Sci.* **16**, 317.
- Luczak, T., 1990, *Random Struct. Algorithms* **1**, 287.
- Ma, S. K., 1976, *Modern Theory of Critical Phenomena* (Benjamin/Cummings, Reading).
- Mandelbrot, B. B., 1982, *The Fractal Geometry of Nature* (Freeman, New York).
- Margulis, G. A., 1974, *Probl. Inf. Transm.* **10**, 101.
- Medina, A., I. Matta, and J. Byers, 2000, *Comput. Commun. Rev.* **30**, 18.
- Mehta, M. L., 1991, *Random Matrices*, 2nd ed. (Academic, New York).

- Milgram, S., 1967, *Psychol. Today* **1**, 60.
- Molloy, M., and B. Reed, 1995, *Random Struct. Algorithms* **6**, 161.
- Molloy, M., and B. Reed, 1998, *Combinatorics, Probab. Comput.* **7**, 295.
- Monasson, R., 2000, *Eur. Phys. J. B* **12**, 555.
- Montoya, J. M., and R. V. Solé, 2000, preprint cond-mat/0011195.
- Moore, C., and M. E. J. Newman, 2000a, *Phys. Rev. E* **61**, 5678.
- Moore, C., and M. E. J. Newman, 2000b, *Phys. Rev. E* **62**, 7059.
- Moore, E. F., and C. E. Shannon, 1956a, *J. Franklin Inst.* **262**, 201.
- Moore, E. F., and C. E. Shannon, 1956b, *J. Franklin Inst.* **262**, 281.
- Morra, S., M. Barthélémy, S. H. Stanley, and L. A. N. Amaral, 2001, "Information filtering and the growth of scale-free networks: Power law with exponential tail," preprint.
- Moukarzel, C., 1999, *Phys. Rev. E* **60**, R6263.
- Newman, M. E. J., 2000, *J. Stat. Phys.* **101**, 819.
- Newman, M. E. J., 2001a, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 404.
- Newman, M. E. J., 2001b, *Phys. Rev. E* **64**, 016131.
- Newman, M. E. J., 2001c, *Phys. Rev. E* **64**, 016132.
- Newman, M. E. J., 2001d, *Phys. Rev. E* **64**, 025102(R).
- Newman, M. E. J., 2001e, unpublished.
- Newman, M. E. J., C. Moore, and D. J. Watts, 2000, *Phys. Rev. Lett.* **84**, 3201.
- Newman, M. E. J., and D. J. Watts, 1999a, *Phys. Lett. A* **263**, 341.
- Newman, M. E. J., and D. J. Watts, 1999b, *Phys. Rev. E* **60**, 7332.
- Newman, M. E. J., S. H. Strogatz, and D. J. Watts, 2001, *Phys. Rev. E* **64**, 026118.
- Palmer, C. R., and J. G. Steffan, 2000, in *Proceedings of IEEE Globecom'00*, San Francisco, Calif. (IEEE, Piscataway, N.J.), Vol. 1, p. 434.
- Pandit, S. A., and R. A. Amritkar, 1999, *Phys. Rev. E* **60**, R1119.
- Pareto, V., 1897, *Cours d'Economie Politique*, Vol. 2 (Université de Lausanne, Lausanne).
- Pastor-Satorras, R., A. Vázquez and A. Vespignani, 2001, preprint cond-mat/0105161.
- Pastor-Satorras, R., and A. Vespignani, 2001a, *Phys. Rev. Lett.* **86**, 3200.
- Pastor-Satorras, R., and A. Vespignani, 2001b, preprint cond-mat/0102028.
- Paxson, V., and S. Floyd, 1997, *Proceedings of the 1997 Winter Simulation Conference*, San Diego (Conference Board of Directors, San Diego), p. 1037.
- Pennock, D. M., C. L. Giles, G. W. Flake, S. Lawrence, and E. Glover, 2000, NEC Research Institute Technical Report 2000-164.
- Pimm, S. L., 1991, *The Balance of Nature* (University of Chicago, Chicago).
- Rapoport, A., 1957, *Bull. Math. Biophys.* **19**, 257.
- Redner, S., 1998, *Eur. Phys. J. B* **4**, 131.
- Rodríguez-Iturbe, I., and A. Rinaldo, 1997, *Fractal River Basins* (Cambridge University, Cambridge/New York).
- Savageau, M. A., 1998, *BioSystems* **47**, 9.
- Scala, A., L. A. N. Amaral, and M. Barthélémy, 2000, *Europhys. Lett.* **55**, 594.
- Schilling, C. H., and B. O. Palsson, 1998, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4193.
- Simon, H. A., 1955, *Biometrika* **42**, 425.
- Slanina, F., and M. Kotrla, 1999, *Phys. Rev. Lett.* **83**, 5587.
- Slanina, F., and M. Kotrla, 2000, *Phys. Rev. E* **62**, 6170.
- Solé, R. V., and J. M. Montoya, 2001, *Proc. R. Soc. London Ser. B* **268**, 2039.
- Solé, R. V., and S. Valverde, 2001, *Physica A* **289**, 595.
- Solomonoff, R., and A. Rapoport, 1951, *Bull. Math. Biophys.* **13**, 107.
- Stanley, H. E., 1971, *Introduction to Phase Transitions and Critical Phenomena* (Oxford University, New York).
- Stauffer, D., and A. Aharony, 1992, *Introduction to Percolation Theory* (Taylor & Francis, London).
- Steyvers, M., and J. B. Tenenbaum, 2001, preprint, www-psych.stanford.edu/~jbt/
- Strogatz, S. H., 2001, *Nature (London)* **410**, 268.
- Tadić, B., 2001a, *Physica A* **293**, 273.
- Tadić, B., 2001b, preprint cond-mat/0104029.
- Valente, T., 1995, *Network Models of the Diffusion of Innovations* (Hampton, Cresskill, NJ).
- Vázquez, A., 2000, preprint cond-mat/0006132.
- Vázquez, A., 2001, preprint cond-mat/0105031.
- Vogelstein, B., D. Lane, and A. J. Levine, 2000, *Nature (London)* **408**, 307.
- Wagner, A., and D. Fell, 2000, Technical Report No. 00-07-041, Santa Fe Institute.
- Walsh, T., 2001, in *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, August 2001, Seattle.
- Wang, X. F., and G. Chen, 2001, preprint cond-mat/0105014.
- Wasserman, S., and K. Faust, 1994, *Social Network Analysis: Methods and Applications* (Cambridge University, Cambridge).
- Watts, D. J., 1999, *Small Worlds: The Dynamics of Networks between Order and Randomness* (Princeton University, Princeton, NJ).
- Watts, D. J., 2000, Santa Fe Working Paper No. 00-12-062.
- Watts, D. J., and S. H. Strogatz, 1998, *Nature (London)* **393**, 440.
- Weigt, M., and A. K. Hartmann, 2001, *Phys. Rev. Lett.* **86**, 1658.
- West, G. B., J. H. Brown, and B. J. Enquist, 1997, *Science* **276**, 122.
- Wigner, E. P., 1955, *Ann. Math.* **62**, 548.
- Wigner, E. P., 1957, *Ann. Math.* **65**, 203.
- Wigner, E. P., 1958, *Ann. Math.* **67**, 325.
- Wilf, H. S., 1990, *Generating Functionology* (Academic, Boston).
- Williams, R. J., N. D. Martinez, E. L. Berlow, J. A. Dunne, and A.-L. Barabási, 2000, Santa Fe Institute Working Paper 01-07-036.
- Willinger, W., M. S. Taqqu, R. Sherman, and D. V. Wilson, 1997, *IEEE/ACM Trans. Netw.* **5**, 71.
- Yook, S., H. Jeong, and A.-L. Barabási, 2001a, preprint cond-mat/0107417.
- Yook, S., H. Jeong, and A.-L. Barabási, 2001b, unpublished.
- Yook, S., H. Jeong, A.-L. Barabási, and Y. Tu, 2001, *Phys. Rev. Lett.* **86**, 5835.
- Zipf, G. K., 1949, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Addison-Wesley, Cambridge, MA).
- Zizzi, P. A., 2001, preprint gr-qc/0103002.

Statistical Theory of Asteroid Escape Rates

Charles Jaffé,^{1,2,3} Shane D. Ross,^{1,2} Martin W. Lo,^{1,2} Jerrold Marsden,¹ David Farrelly,⁴ and T. Uzer^{1,5}

¹*Control and Dynamical Systems Division 107-81, California Institute of Technology, Pasadena, California 91125*

²*Navigation and Flight Mechanics, Jet Propulsion Laboratory, Pasadena, California 91109-8099*

³*Department of Chemistry, West Virginia University, Morgantown, West Virginia 26506-6045*

⁴*Department of Chemistry, Utah State University, Logan, Utah 84322-0300*

⁵*Center for Nonlinear Sciences and School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332-0430*

(Received 1 February 2002; published 12 June 2002)

Transition states in phase space are identified and shown to regulate the rate of escape of asteroids temporarily captured in circumplanetary orbits. The transition states, similar to those occurring in chemical reaction dynamics, are then used to develop a statistical semianalytical theory for the rate of escape of asteroids temporarily captured by Mars. Theory and numerical simulations are found to agree to better than 1%. These calculations suggest that further development of transition state theory in celestial mechanics, as an alternative to large-scale numerical simulations, will be a fruitful approach to mass transport calculations.

DOI: 10.1103/PhysRevLett.89.011101

PACS numbers: 96.35.-j, 05.45.-a, 82.20.Db, 95.10.Ce

While large-scale chaos exists in the Solar System, it is sufficiently weak that the motions of most of the planets appear quite regular, at least on relatively short time scales [1]. In contrast, smaller bodies such as asteroids and comets, through their interactions with the planets and the Sun, can exhibit strongly chaotic motion. Nevertheless, the ability to predict the behavior of *populations* of these small but numerous objects is essential for understanding such problems as the evolution of both short- and long-range comets originating in the Kuiper Belt and the Oort Cloud, respectively [2], the dynamics of near-Earth asteroids [4,5], zodiacal and circumplanetary dust dynamics [3], and the gravitationally assisted transport of spacecraft [6]. The discovery of possible traces of a living organism in a Martian meteorite found in Antarctica [7] has stimulated investigations into the feasibility of viable microbes transporting through space [8] and illustrates the fundamental importance of understanding mass transport in the Solar System to theories of the origin of life. In this Letter we study the problem of computing average rates of asteroid escape from Mars because of its bearing on understanding the feasibility of transport of viable microorganisms between the two planets.

On its face it might seem that once individual orbits are known the problem is solved. However, deeper insights can be obtained by computing *rates* because these allow models of the evolution of populations of particles to be constructed. In principle, the computation of rates of mass transport can be accomplished by large numerical simulations in which the orbits of vast numbers of test particles are propagated in time including as many interactions as desirable [2]. However, such calculations are computationally demanding and it may be difficult to extract from them information about key dynamical mechanisms. They do have the considerable advantage, however, that a variety of nongravitational effects can easily be included, even if these destroy the Hamiltonian nature of

the problem. In this Letter a complementary approach is developed that can be used provided that the problem is of autonomous Hamiltonian form. We first identify transition states that regulate mass transport through bottlenecks in phase space [9]. Knowledge of the transition state allows us to compute rates using a Rice-Ramsperger-Kassel-Marcus (RRKM)-like approach similar to that developed in chemical dynamics [10]. Molecular RRKM theory is an essentially statistical approach in which it is assumed that all molecular configurations within the initial microcanonical ensemble are equally likely to react. In other words, the rate of intramolecular energy redistribution is rapid compared to the rate of reaction. As a consequence, the reaction rate can be expressed as the ratio of the flux across the transition state divided by the total volume of *phase* space associated with the reactants. We provide the first application of this statistical approach in celestial mechanics and obtain a very high level of agreement with numerical simulations in which individual orbits are integrated numerically. Our results suggest that the validity of RRKM theory might even be better in celestial mechanics than in chemical dynamics.

First of all we demonstrate the existence of transition states in phase space which serve as bottlenecks to transport in the restricted three body problem (RTBP) which describes a rather wide variety of interesting phenomena in the Solar System [6,11,12]. One such example, which demonstrates the equivalence to a chemical reaction, is provided in Fig. 1: The left panel in Fig. 1 illustrates the dynamics shown by Jupiter-family comets such as *Oterma* and *Gehrels 3* which shuttle back and forth between rather complicated heliocentric orbits outside the orbit of Jupiter (*exterior* region) and orbits inside (*interior* region) [6]. During these transitions the comets are frequently captured temporarily by Jupiter. Because the interior orbits are typically close to a 3:2 resonance (three revolutions around the Sun during two periods of Jupiter) while those in the

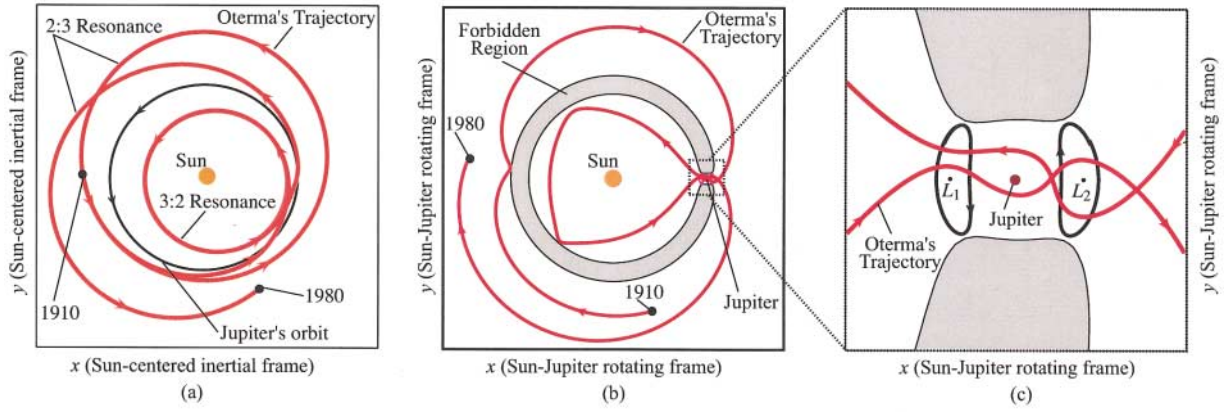


FIG. 1 (color). Left panel: plot of a resonance switching orbit of comet Oterma in heliocentric coordinates. Middle panel: same as left panel in a coordinate system rotating with Jupiter. Right panel: expanded view of the bottleneck region.

exterior regions are near the 2:3 resonance the transitions shown in Fig. 1 are known as “resonance transitions.” In noninertial coordinates, in a frame locked to the rotation of Jupiter, the picture is considerably clearer and the existence of the bottleneck region is readily apparent in the middle panel of Fig. 1. In fact, a dynamical invariant (the Jacobi integral—defined below) divides coordinate space into interior “reactant”) and exterior (“product”) regions which communicate through the narrow bottleneck containing Jupiter and two of the five Lagrange equilibrium points, L_1 and L_2 . The passage of the comet through this region is regulated by phase space structures in the vicinity of the two saddle points L_1 and L_2 —see Fig. 1. Indeed—and this is the key point of this Letter—the periodic orbits around the two saddles, which correspond to the *transition states* [13], control mass transport through the bottleneck. The identification of these structures allows us to compute average rates accurately without the need for large numerical simulations.

The RTBP Hamiltonian is

$$H = E = \frac{1}{2} (p_x^2 + p_y^2) - (xp_y - yp_x) - \frac{1-\mu}{r_1} - \frac{\mu}{r_2} - \frac{1}{2} \mu(1-\mu). \quad (1)$$

Here E is the energy, $r_1 = \sqrt{(x+\mu)^2 + y^2}$, $r_2 = \sqrt{(x-1+\mu)^2 + y^2}$, and the masses for the Sun and the planet are $m_S = 1-\mu$ and $m_P = \mu$. The coordinate system rotates about the center of mass with the Sun and the planet lying on the x axis at $(-\mu, 0)$ and $(1-\mu, 0)$, respectively. The position of the third body (the asteroid) relative to the positions of the Sun and the planet is given as (x, y) . The Jacobi integral is

$$C(x, y, \dot{x}, \dot{y}) = -(\dot{x}^2 + \dot{y}^2) + 2\Omega(x, y) = -2E(x, y, \dot{x}, \dot{y}), \quad (2)$$

where

$$\Omega(x, y) = \frac{1}{2} (x^2 + y^2) + \frac{1-\mu}{r_1} + \frac{\mu}{r_2} + \frac{1}{2} \mu(1-\mu). \quad (3)$$

Unlike most chemical reactions, this problem contains two saddle points with resonance transitions for Jupiter-family comets taking place when the Jacobi integral, C , is in the range $C_3 < C < C_2$ with the subscripts referring to the values of C at the similarly numbered Lagrange points. Then, the third body moves from the interior to the exterior of the orbit of Jupiter by passing through the narrow bottleneck in the vicinity of the planet. The presence of *two* equally important transition states makes this problem more intricate than is commonly found in chemical reaction dynamics where the flow is usually regulated by a single dominant transition state. By considering a different range of values of the Jacobi constant in the RTBP, however, we are able to find a direct analogy to the most well studied class of chemical reaction—unimolecular reactions. This occurs when $C_2 < C < C_1$ and describes an asteroid (or any planetoid or particle) that is already orbiting a planet and which can escape inwards (towards the Sun) but not outwards, i.e., a particle trapped in the region near the planet can escape only through the bottleneck at L_1 shown in Fig. 1. We now specialize to asteroid escape from Mars orbits.

The general procedure for identifying the transition state(s) in two or more degrees of freedom using Hamiltonian normal form theory [14] is described in detail in Ref. [9]. Briefly, a sequence of local, nonlinear transformations of the coordinates is developed that transforms the Hamiltonian into the so-called *normal form*. In the Mars asteroid problem the linear part of the normal form in the neighborhood of the equilibrium point L_1 at $(k, 0)$ is

$$H_l = \frac{1}{2} \{ (p_x + y)^2 + (p_y - x)^2 - ax^2 + by^2 \}, \quad (4)$$

where $a = 2\rho + 1$ and $b = \rho - 1$ and where

$$\rho = \mu|k - 1 + \mu|^{-3} + (1 - \mu)|k + \mu|^{-3}. \quad (5)$$

The L_1 equilibrium point is a saddle center [14]. Interestingly, in appropriate coordinates, this Hamiltonian is formally identical to that obtained in the problem of the ionization of a hydrogen atom in crossed electric and magnetic fields [9,15,16]. The transport of the electron from the interior region to the exterior region was shown to be regulated by a phase-space transition state [13,17] associated with the periodic orbit centered on the equilibrium point at the Stark saddle—the atomic equivalent of L_1 . This normal form expansion is valid in the neighborhood of the equilibrium point. Once the structure of phase space has been established locally, it can then be continued numerically outside of the local region. In the energy domain, this continuation is valid up to the next bifurcation of phase-space structure.

We consider a situation in which an asteroid (or other body) starts out in a circumplanetary orbit around Mars—perhaps the body arrived in orbit after it was ejected from the Martian surface following an impact. However they arrived in orbit, we are here only interested in the *rate* of escape of such bodies. Two sets of calculations were performed. The first set consists of direct Monte Carlo simulations of survival probabilities; 107 000 particles with randomly selected initial conditions were started at 200 Mars radii from Mars and integrated until they escaped the planet by crossing the transition state and entered an orbit around the Sun. This ensemble is similar to a microcanonical ensemble in a chemical reaction in that the only restrictions on the initial conditions are due to energy constraints. The results of the simulations were binned and from this the rate was calculated directly. In the second set of computations (i) the transition state was identified, as described above and (ii) RRKM theory was used to determine the rate [10].

In RRKM theory the reaction rate, k , can be expressed as the ratio of the flux across the transition state divided by the volume of the portion of phase space associated with the reactants. Gray *et al.* have expressed this ratio as [18]

$$k = \frac{1}{N_A} \int dR dP d\mathbf{r} d\mathbf{p} \delta(E - H) \times \delta(R - R^\ddagger) \theta(P) \left(\frac{P}{\mu} \right). \quad (6)$$

Here the integral is over all of phase space. The variables R, P are the reaction coordinate and its conjugate momentum, while \mathbf{r}, \mathbf{p} refer to the bath coordinates (i.e., coordinates other than the reaction coordinate) and their conjugate momenta, and R^\ddagger is the coordinate of the transition state. The first delta function restricts us to the energy shell and the second delta function restricts us to the phase-space transition state within that energy shell. This surface can be divided into two hemispheres characterized by the direction of the flow across the transition state [9]. The

Heaviside step function $\theta(P)$ selects the hemisphere with positive flux (from reactants to products). Thus, the integral in the numerator is the flux across the phase-space transition state. N_A is essentially a normalization factor and in the simplest formulations is taken to be equal to the volume of the reactant region of the energy shell. The actual computations are done in Levi-Civita regularizing coordinates as described in Ref. [13] with the rate being thought of as the ratio of the flux that actually crosses the transition state to the flux attempting the transition, $\Phi_{\text{react}}/\Phi_{\text{total}}$.

The computations can be simplified by noting that, in two degree-of-freedom systems, the projection of an unstable periodic orbit into configuration space defines a surface that separates reactants and products. It is, therefore, called a “periodic orbit dividing surface” or PODS because, in phase space, the stable and unstable manifolds of this orbit partition the energy shell [9]. To compute Φ_{total} the PODS on the Mars side that has *maximum* flux is located. This PODS is then used to define the Poincaré surface of section shown in Fig. 2. The chaotic flux, equal to the area of the chaotic sea, is evaluated by finding the “last” invariant tori bounding this sea. There are three of these, one for the outer shore, and two for the shores of the two inner islands. The sum and difference of the action integrals over these tori provide an excellent approximation to the area of the chaotic sea. The flux across the transition state, Φ_{react} , is obtained by integration of the action integral over the PODS that has *minimum* flux over it.

Figure 3 shows the results of the two calculations for the *initial* rates, that is, after 40 orbital periods defined by the mean motion. After a brief initial period during

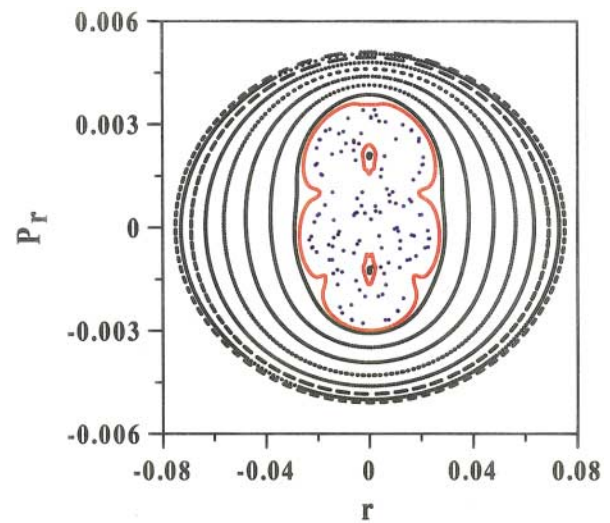


FIG. 2 (color). A composite surface of section (SOS) obtained by integration of 14 different trajectories. The SOS is defined using the PODS of maximum flux. The black points lie on invariant tori, while the blue points correspond to a single orbit that escapes after 228 periods. The red points correspond to the three trajectories which are confined to the last invariant tori defining the shores of the chaotic sea.

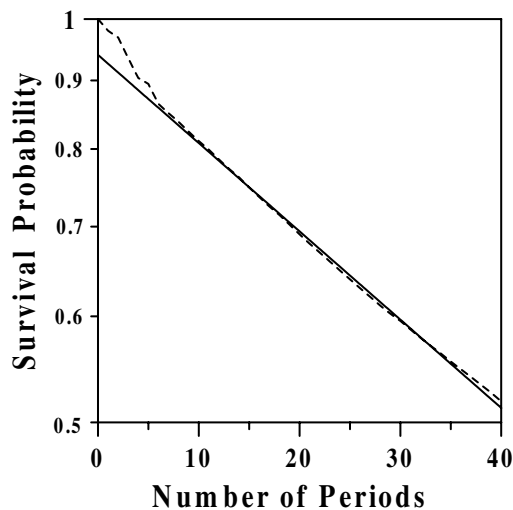


FIG. 3. Survival probability obtained from Monte Carlo (dashed line) and transition state theory (solid line) as described in the text. In the Monte Carlo calculations 107 000 orbits were integrated using initial conditions chosen randomly from a one-dimensional manifold defined by a given value of the Jacobi integral ($C = -3.000\,936$). After a short initial equilibration period the largest error settles down to on the order of 1% and error bars would not be visible on this plot. Scaled units are used.

which transients (due to details of the particular selection of initial conditions) die out, transition state theory settles down to agree to better than 1% with the numerical simulations. Non-RRKM behavior is the exception rather than the rule in chemistry, and our results suggest that in problems in the Solar System, the rate of the analogous mixing might also be fast compared to the rate of transition. Our calculations, which provide support for this assertion, will be tested further by extending our methods to higher dimensions [9]. In the asteroid escape problem studied here, the dynamics was confined to the plane so as to allow the simplest illustration of the method. However, the phase-space transition state theory is most powerful for multidimensional degree-of-freedom systems for which simulations become more difficult and insight into the dynamical mechanisms is harder to extract [9]. Examples include the evolution of long-range comets [2] and circumplanetary dust escape from nonequatorial “halo” orbits [19]. While we have not considered nongravitational

forces our methods allow the inclusion, for example, of interactions of charged dust grains with planetary magnetic fields or the effect of solar radiation pressure.

This work was partly supported by the U.S. National Science Foundation, by the American Chemical Society (PRF), by the West Virginia–NASA Space Grant Program, and by NASA-ASEE (C. J.) This work was carried out in part at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

-
- [1] J. Laskar, *Nature (London)* **338**, 237 (1989).
 - [2] P. Weigert and S. Tremaine, *Icarus* **137**, 84 (1998).
 - [3] P. Michel, F. Migliorini, A. Morbidelli, and V. Zappalà, *Icarus* **145**, 332 (2000).
 - [4] P. Farinella *et al.*, *Nature (London)* **371**, 314 (1994).
 - [5] M. Horányi, *Annu. Rev. Astron. Astrophys.* **34**, 383 (1996).
 - [6] W. S. Koon, M. W. Lo, J. E. Marsden, and S. D. Ross, *Chaos* **10**, 427 (2000).
 - [7] D. S. MacKay, E. K. Gibson, K. L. Thomas-Keprta, H. Vali, C. S. Romanek, S. J. Clemett, X. D. F. Chillier, C. R. Maechling, and R. N. Zare, *Science* **243**, 924 (1996).
 - [8] C. Mileikowsky *et al.*, *Icarus* **145**, 391 (2000).
 - [9] S. Wiggins, L. Wiesenfeld, and C. Jaffé, and T. Uzer, *Phys. Rev. Lett.* **86**, 5478 (2001).
 - [10] For reviews, see, e.g., J. C. Keck, *Adv. Chem. Phys.* **13**, 85 (1967); D. G. Truhlar, *J. Phys. Chem.* **100**, 12 271 (1996).
 - [11] R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Addison-Wesley, New York, 1978), 2nd ed.
 - [12] K. E. Meyer and R. Hall, *Hamiltonian Mechanics and the N-body Problem* (Springer-Verlag, New York, 1992).
 - [13] C. Jaffé, D. Farrelly, and T. Uzer, *Phys. Rev. Lett.* **84**, 610 (2000); *Phys. Rev. A* **60**, 3833 (1999).
 - [14] V. I. Arnol'd, V. V. Kozlov, and A. I. Neishtadt, *Mathematical Aspects of Classical and Celestial Mechanics* (Springer-Verlag, New York, 1988).
 - [15] J. von Milczewski, D. Farrelly, and T. Uzer, *Phys. Rev. Lett.* **78**, 2349 (1997).
 - [16] T. Uzer and D. Farrelly, *Phys. Rev. A* **52**, R2501 (1995).
 - [17] S. Wiggins, *Normally Hyperbolic Invariant Manifolds in Dynamical Systems* (Springer-Verlag, New York, 1994).
 - [18] S. K. Gray, S. A. Rice, and M. J. Davis, *J. Phys. Chem.* **90**, 3470 (1986).
 - [19] J. E. Howard, H. R. Dullin, and M. Horányi, *Phys. Rev. Lett.* **84**, 3244 (2000).