ОБЪЕДИНЕННЫЙ ИНСТИТУТ ЯДЕРНЫХ ИССЛЕДОВАНИЙ

11-2009-46

На правах рукописи
УДК 51-7:[004.6:004.732]

ИВАНОВ
Валерий Викторович

СТАТИСТИЧЕСКАЯ МОДЕЛЬ ИНФОРМАЦИОННОГО ТРАФИКА

Специальность: 05.13.18 — математическое моделирование, численные методы и комплексы программ

Автореферат диссертации на соискание ученой степени кандидата физико-математических наук

Дубна 2009

Работа выполнена в *Лаборатории информационных технологий Объединенного института ядерных исследований.*

Научный руководитель: *кандидат физико-математических наук Зрелов Петр Валентинович*

Официальные оппоненты: *доктор физико-математических наук, профессор Крянев Александр Витальевич, МИФИ, г. Москва*

доктор технических наук Гостев Иван Михайлович, МИЭМ, г. Москва

Ведущая организация: *Российский университет дружбы народов, г. Москва*

Защита состоится «22» мая 2009 г. в 14[00] часов на заседании диссертационного совета *Д 720.001.04* при *Лаборатории информационных технологий ОИЯИ, 141980, г. Дубна Московской обл., ул. Жолио-Кюри 6.*

С диссертацией можно ознакомиться в библиотеке *Объединенного института ядерных исследований.*

Автореферат разослан «10» 04 2009 г.

Ученый секретарь
диссертационного совета,
кандидат физико-математических наук *Иванченко З.М.*

Общая характеристика работы

В настоящей работе развиты новые математические модели и методы для исследования характерных особенностей информационных потоков (трафика) в компьютерных сетях.

Актуальность работы. В условиях глобального информационного общества быстрый, надежный и безопасный обмен данными между локальными и глобальными компьютерными сетями представляет собой проблему высочайшего приоритета. Исследования сетевого трафика показали, что он представляет собой сложный динамический процесс, характеризующийся, в частности, распределениями с тяжелыми хвостами, длинно-масштабными корреляциями, мультифрактальностью и т.д. [1]-[5]. Трудности, с которыми столкнулись исследователи, привели их к выводу о том, что сетевой трафик нельзя адекватно описать в рамках существующих моделей [6, 7], а традиционные математические методы малопригодны для анализа временных рядов, отвечающих информационным потокам [8, 9]. В то же время, функционирование компьютерных сетей ключевым образом зависит от их технической и программной поддержки, в том числе с учетом моделей, построенных на основе выявленных закономерностей и отражающих основные особенности сетевого трафика.

В этой связи, важной задачей для скоростных телекоммуникационных систем и компьютерных сетей является разработка моделей трафика, которые бы реалистично отражали основные его особенности, а также математических методов, адекватных анализируемым случайным процессам. Такие методы и модели могут помочь в разработке методов и средств, нацеленных на повышение качества обслуживания компьютерных сетей, обеспечение эффективного контроля и управления информационными потоками, защиту сетей от несанкционированных вторжений и т.д.

Цель диссертационной работы. Разработка новых моделей и методов для изучения характерных особенностей информационного трафика и их применение в решении конкретных задач.

Научная новизна:

1. Получена оценка размерности вложения динамического процесса информационного трафика.

2. На основе нейронной сети построена модель информационного трафика, с помощью которой удалось воспроизвести статистическое распределение его потока, а также подтвердить оценку размерности вложения соответствующего процесса.

3. Получено статистическое распределение информационного потока, с

высокой точностью отвечающее логнормальному закону распределения.

4. На основе подхода "Гусеница", критерия знаков, χ^2- и ω^2-критериев разработана процедура разбиения всего набора главных компонент на ведущие (ответственные за формирование логнормального распределения) и остаточные, носящие характер высокочастотного шума.

5. Развиты новые методы определения моментов смены состояния анализируемого динамического процесса.

Практическая ценность:

- Развитые в работе математические методы позволили исследовать характерные особенности сетевого трафика и получить новые результаты о соответствующем динамическом процессе.

- На основе искусственной нейронной сети (ИНС) разработана модель трафика, позволившая оценить размерность вложения соответствующего процесса и воспроизвести статистическое распределение потока информации.

- Построена статистическая модель информационного трафика, которая может служить основой для разработки новых методов и средств для более эффективного контроля и управления информационными потоками и защиты компьютерных сетей от несанкционированного доступа.

- Разработаны новые методы детектирования точек смены состояния анализируемого процесса, позволяющие вести эффективный контроль сетевого трафика.

Результаты и положения, выносимые на защиту:

1. С помощью методов нелинейного анализа получены оценки интервала корреляции и размерности вложения для динамического процесса, ответственного за формирование сетевого трафика.

2. Непараметрическая модель сетевого трафика, построенная на основе искусственной нейронной сети (ИНС) прямого распространения, которая воспроизвела статистические особенности информационного трафика, а также подтвердила оценку размерности вложения трафика, полученную с помощью метода главных компонент.

3. Статистический закон распределения информационных потоков для агрегированных измерений трафика, аппроксимируемый с высокой точностью функцией логнормального распределения.

4. Метод разбиения всего набора главных компонент разложения временного ряда измерений трафика на основные (ответственные за формирование логнормального распределения) и остаточные, которые носят характер высокочастотного шума, наложенного на основной процесс.

5. Новый алгоритм пороговой вейвлет-фильтрации исходных измерений трафика для исключения высокочастотной (шумовой) составляющей трафика, что позволило описать основную составляющую трафика минимальным (2-3) набором основных компонент.

6. Новые методы детектирования моментов смены состояния анализируемого временного ряда, в основу которых положена дискриминация по принципу "свой-чужой".

Апробация работы. Основные положения и результаты работы докладывались и обсуждались на научных семинарах ЛИТ, кафедры прикладной математики Московского инженерно-физического института, Российского университета дружбы народов и на различных международных конференциях, в том числе ([A7] [A14]):

- VIII Int. Workshop on "Advanced Computing and Analysis Techniques in Physics Research" - ACAT'2002, 24-28 June, 2002, Moscow, RUSSIA.

- 5-й Международный конгресс по математическому моделированию, г. Дубна, Россия, 30 сентября - 6 октября, 2002.

- I-st Int. Conf. on "Mathematics and Informatics for Industry", MII, 14-16 April 2003, Thessaloniki, Greece.

- VII world multiconference on "Systemics, Cybernetics and Informatics", SCI 2003, Focus Symposium on "Quantum Physics and Communication", Dubna, Russia, 30 July - 2 August, 2003.

- Летняя школа DAAD "Трафик и эконофизика", г. Дубна, Россия, 28 июля - 17 августа, 2003.

- XIX International Symposium on Nuclear Electronics & Computing, NEC'2003, September 15-20, 2003, Varna, Bulgaria.

- Международная конференция "Распределённые вычисления и Грид-технологии в науке и образовании", г. Дубна, Россия, 29 июня - 2 июля, 2004.

Публикации. В основу диссертации положены 20 работ, которые опубликованы как в реферируемых журналах:

- Физика элементарных частиц и атомного ядра (ЭЧАЯ) [A1],
- Письма в ЭЧАЯ [A2],
- Physica D [A3],
- Nuclear Instruments & Methods in Physics Research [A4],
- Physica A [A5],
- Discrete Dynamics in Nature & Society [A6]

и материалах международных конференций ([A7] [A14]), так и в виде препринтов и сообщений ОИЯИ ([A15] [A20]).

Личный вклад автора. Вклад автора является определяющим.

Структура и объем диссертации. Диссертация содержит введение, обзор литературы, 6 глав, заключение, список литературы (163 ссылки) и имеет объем 150 страниц.

Содержание работы

Во Введении обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследований, показана практическая значимость полученных результатов, представлены выносимые на защиту научные положения.

В первой главе рассмотрена специализированная система сбора, анализа и управления трафиком (ССАУ "Трафик") [11]. С ее помощью проводились детальные измерения сетевого трафика на входном шлюзе локальной сети университета "Дубна" [10]. Блок-схема ССАУ "Трафик" представлена на рис. 1.

Эта система позволяет в реальном времени контролировать параметры трафика, записывать регистрируемую информацию в базу данных и обеспечивает наглядную визуализацию результатов анализа трафика.

Измерения сетевого трафика выполняются с помощью сетевого адаптера в режиме открытого драйвера в целях создания условий для приема и анализа передаваемых по сети пакетов.

Драйвер открытого режима записывает принятые пакеты в буфер предварительного захвата и выставляет флаг приема пакета, после чего активизируется модуль приема пакета и производится анализ поля типа пакета для выделения из общего потока лишь пакетов стека TCP/IP. После идентификации возможно отделение заголовка пакета и уничтожение блока данных, а также запись заголовка в базу данных SQL - сервера. Наряду с данными

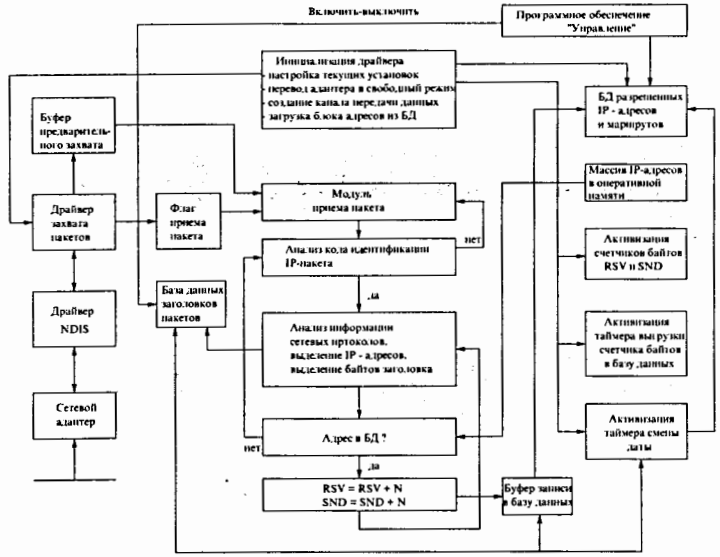

Рис. 1. Блок-схема системы сбора данных

о переданном объеме информации в запись включается также время приема пакета, измеренное с точностью до микросекунды.

Следует отметить, что в локальной сети университета "Дубна" используются два протокола: протокол NetBEUI применяется для внутренних обменов в локальной сети, а протокол TCP/IP используется для внешних обменов.

Наш анализ показал, что вклад NetBEUI-трафика в выполненные нами измерения составил в среднем 1-6 пакетов в секунду в течение рабочего дня. Это ничтожно мало по сравнению с объемом TCP/IP трафика. В связи с этим, мы можем пренебречь влиянием трафика NetBEUI на TCP/IP трафик.

Данные информационного трафика, анализ которых приводится в главах 2-5, отвечают примерно 20 часам измерений. Часть этих измерений, агрегированных с разными размерами окна агрегации, представлена на рис. 2. В главе 6 использовались два других набора измерений, зарегистрированных на входном шлюзе компьютерной сети университета "Дубна" (детали см. в главе 6).

Во второй главе представлены результаты применения методов нелинейного анализа к временным рядам, соответствующим измерениям сетевого трафика. С их помощью получены оценки временной задержки и размерности вложения. На основе искусственной нейронной сети (ИНС) построена динамическая модель трафика, которая позволила:

- получить оценку размерности вложения,

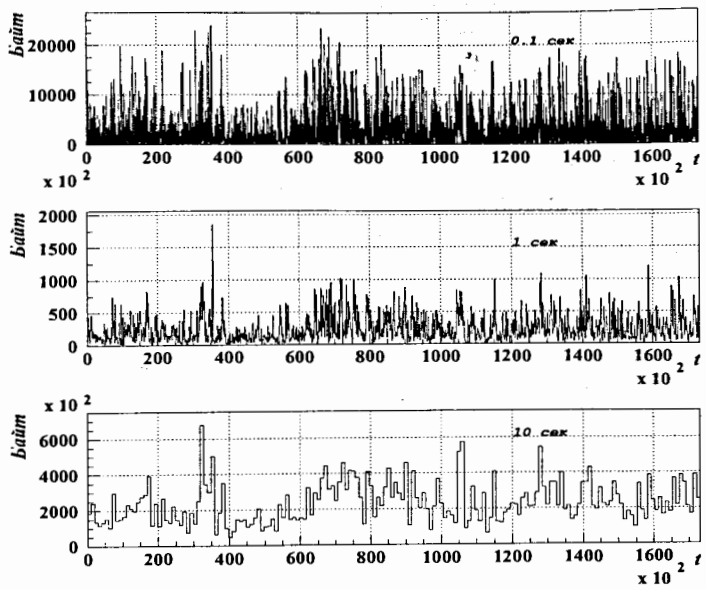

Рис. 2. Измерения информационного трафика, агрегированные с разными окнами агрегации: 0.1 сек, 1 сек, 10 сек

- воспроизвести форму распределения потока информации.

При нелинейном анализе временных рядов сигнал $\{x_i\}$ представляется в виде одномерной проекции динамической системы, действующей в пространстве векторов $\vec{y}_i$ большей размерности:

$$\vec{y}_i = (x_i, x_{i+\tau}, ..., x_{i+(m-1)\tau}),$$

здесь m - размерность вложения, а τ - временная задержка.

В качестве оценки τ взята величина первого пересечения автокорреляционной функции с нулем. Зависимость этой величины от размера окна агрегации представлена на рис. 3.

Для уровней агрегации от 0.1 сек до 10 сек величина τ находится в области: $\tau \sim 10$ сек. Измерения, отстоящие друг от друга на величину временного интервала τ, могут рассматриваться как линейно независимые.

Последовательность некоррелированных измерений может быть рассмотрена как m-мерный вектор, отвечающий искомой динамической системе. Оценка размерности была выполнена с помощью алгоритма Грассбергера-

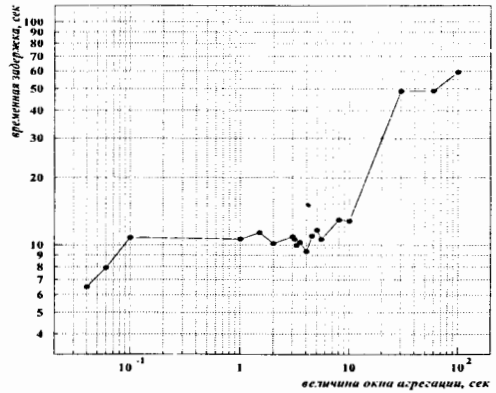

Рис. 3. Зависимость τ от размера окна агрегации

Прокаччиа. Корреляционный интеграл, отвечающий этому алгоритму, вычислялся по формуле

$$C_n^m(r) = \frac{2}{N(N-1)} \sum_{i \neq j} \Theta(r - |\vec{y}_i - \vec{y}_j|),$$

где $|\vec{y}_i - \vec{y}_j| = \max\left\{|x_i - x_j|, ...|x_{i+(m-1)\tau} - x_{j+(m-1)\tau}|\right\}$. Величина $C_n^m(r)$ определяет вероятность того, что расстояние между случайно выбранной парой векторов будет не больше r. Если, начиная с некоторого m, зависимость между логарифмами $C_2(r)$ и r становится линейной

$$\log C_2(r) \approx \beta \log r + \gamma\,,$$

минимальная величина размерности вложения d_E может быть оценена с помощью соотношения

$$\beta < d_E < m.$$

Для реконструкции динамической системы, соответствующей измерениям сетевого трафика, использовалась нейронная сеть прямого распространения. Основное преимущество нейронной сети заключается в том, что она не требует априорной информации, что особенно важно в нашем случае, не только из-за того, что динамическая система трафика сложна, но также из-за отсутствия информации о вкладе отдельных компонент в динамику системы.

Входной слой сети содержал число нейронов, равное величине размерности вложения, два скрытых слоя с переменным числом нейронов и один

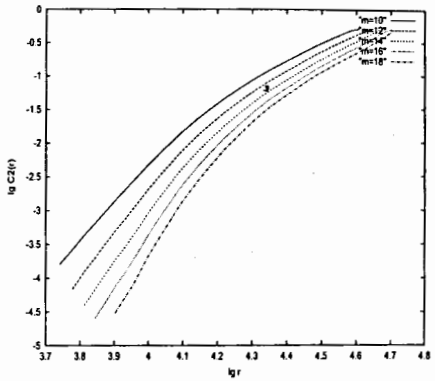

Рис. 4. Корреляционные интегралы $C_2^m(\tau)$ измерений сетевого трафика, агрегированных с окном 1 сек для $\tau = 10$ сек и m =12, 14, 16, 18

выходной нейрон. Выходной нейрон ИНС выдавал предсказываемую величину.

Для обучения сети использовались данные, агрегированные с окном 1 сек. Эти данные были предварительно отнормированы к интервалу [-1,1]. Для образования входного вектора были взяты следующие параметры: $\tau = 10$ сек и $d_E = 15 \div 20$. На рис. 5 приведены распределения потока (нормированного на интервал $[-1, 1]$) для реальных измерений трафика (рис. 5а) и ряда, сгенерированного с помощью ИНС (рис. 5б).

На рис. 6а представлен временной ряд измерений, агрегированных с окном 1 сек, а также результаты моделирования трафика с помощью ИНС. На рис. 6б представлено распределение абсолютных величин весов между выходным нейроном и нейронами второго скрытого слоя. Представленная зависимость показывает, что размерность динамической системы близка к 12, поскольку вклад остальных весов близок нулю. Распределение имеет тот же вид, что и распределение, полученное с помощью другого метода — метода главных компонент (см. рис. 11).

В третьей главе исследуется влияние агрегации на формирование статистического распределения потока информации.

На рис. 7а представлено распределение размеров пакетов оригинальных измерений трафика, в то время как на рис. 7б приведено распределение сетевого трафика, агрегированного с окном 100 мсек.

Видно, что для малых значений величины окна агрегации распределение информационного потока не носит выраженного характера. Однако, при приближении размера окна агрегации к 1 сек (см. рис. 8а), распределение приобретает устойчивую форму, которая не меняет характера при дальней-

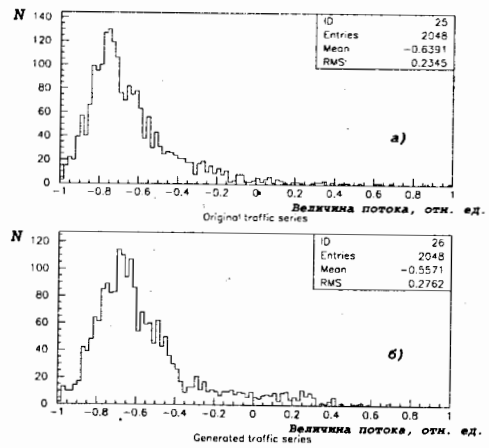

Рис. 5. Распределение потока (нормированного к интервалу $[-1, 1]$) для: а) исходных данных и б) данных, сгенерированных обученной ИНС

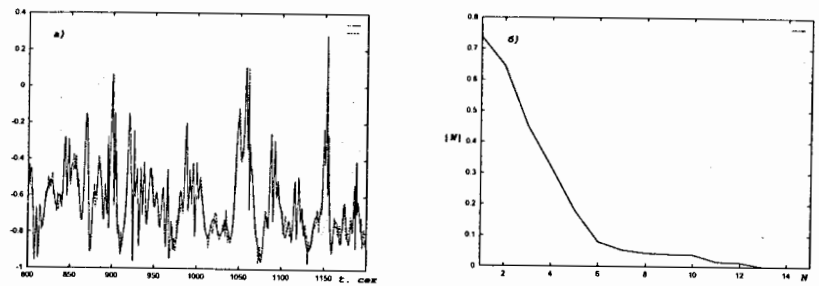

Рис. 6. а) Предсказание временного ряда сетевого трафика (нормированного к интервалу $[-1, 1]$) с помощью обученной ИНС, б) распределение абсолютных величин весов между выходным нейроном и нейронами второго скрытого слоя

шем увеличении окна агрегации: см., например, рис. 8б, соответствующий агрегации с окном 10 сек. Аппроксимирующие кривые, представленные на рисунках 8, отвечают функции логнормального распределения

$$f(x) = \frac{A}{\sqrt{2\pi}\sigma}\frac{1}{x}\exp\left[-\frac{1}{2\sigma^2}(\ln x - \mu)^2\right], \qquad (1)$$

где x - переменная, σ- и μ- параметры логнормального распределения, A- нормировочный множитель.

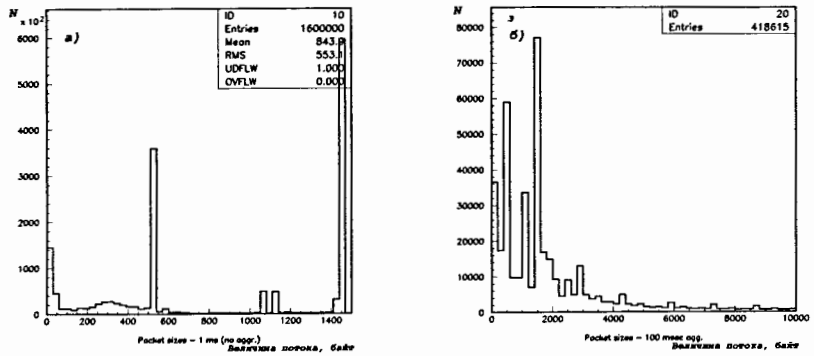

Рис. 7. а) Распределение размеров пакетов для исходных данных, б) распределение потока информации для данных, агрегированных с окном 100 мсек

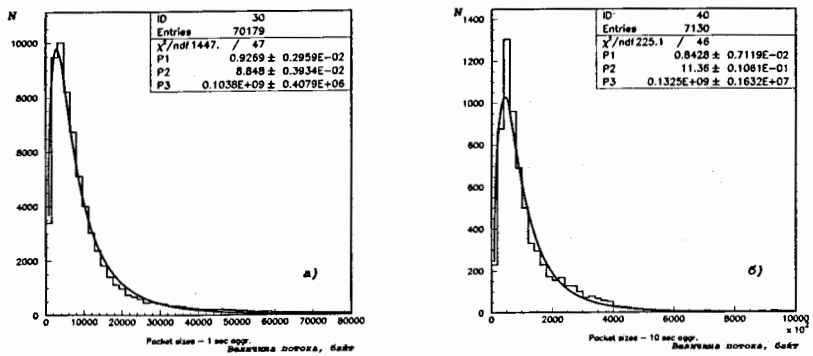

Рис. 8. Распределение потока информации (аппроксимирующая кривая - функция (1)): а) для данных, агрегированных с окном 1 сек, б) для данных, агрегированных с окном 10 сек

Следует заметить, что распределения, приведенные на рисунках 8, включают в себя весь набор данных, что соответствует приблизительно 20 часам непрерывных измерений. В то же время, поведение трафика, также как и соответствующее ему статистическое распределение, меняется в зависимости от того, когда делались эти измерения – в течение рабочего дня или в ночное время (смотри также главу 6). В частности, если рассматривать только дневную часть измерений трафика, то соответствующее распределение информационного потока с высокой точностью согласуется с гипотезой (1) см. рис. 9.

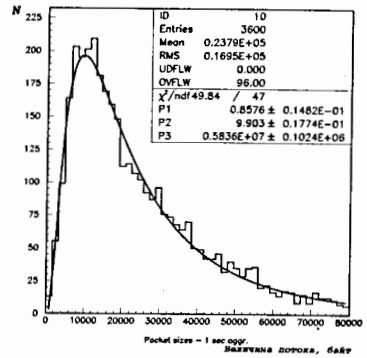

Рис. 9. Распределение сетевого трафика, агрегированного с окном 1 сек, для фрагмента дневных измерений: аппроксимация функцией (1)

С целью выявления особенностей сетевого трафика, влияющих на формирование логнормального закона, была построена модель, в которой методом Монте-Карло генерировались две случайные величины: размер регистрируемого пакета P_s и временной интервал T_{int}, разделяющий последовательно приходящие пакеты. Для моделирования величины P_s использовалось эмпирическое распределение для исходных измерений трафика (рис. 7а), а для величины T_{int} использовались как соответствующее эмпирическое распределение, так и его приближение экспоненциальным распределением (что соответствует пуассоновской модели). При моделировании случайные величины P_s и T_{int} полагались независимыми, что было предварительно установлено на основании анализа имеющихся данных.

Сгенерированный с помощью этой модели ряд подвергался процедуре агрегации на различных уровнях, а полученные статистические распределения аппроксимировались логнормальным распределением. Результаты аппроксимации (рис. 10а и б) в целом подтверждают предположение о независимости временных интервалов между соседними пакетами от величины этих пакетов. При этом пуассоновская модель была отброшена, поскольку она не воспроизводит распределение реальных данных.

Четвертая глава посвящена сингулярно-спектральному анализу (ССА) измерений сетевого трафика с помощью подхода "Гусеница" [13, 14].

Анализируется временной ряд, отвечающий произвольной функции $f(t)$, определенной на равномерной сетке:

$$x_i = f[t_i] = f[(i-1)\Delta t], \quad i = 1, 2, \ldots, K, \tag{2}$$

где Δt – временной интервал (в нашем случае $\Delta t = 1$).

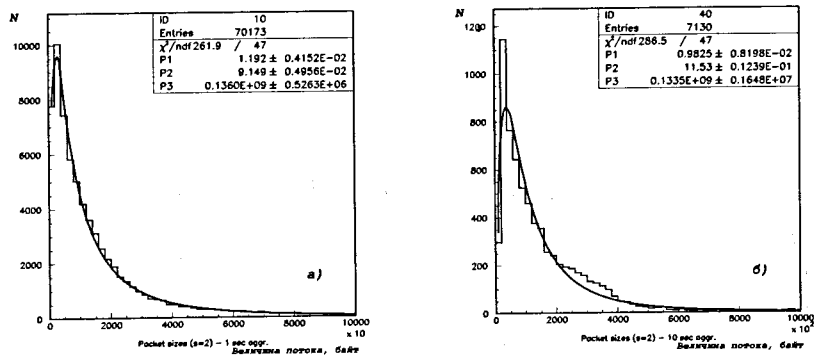

Рис. 10. Распределение потока информации для модельных данных: а) агрегированных с окном 1 сек, б) агрегированных с окном 10 сек

Стандартная схема "Гусеницы"-ССА состоит из четырех этапов:

1. преобразование одномерного ряда к многомерному виду,
2. сингулярное разложение выборочной ковариационной матрицы,
3. анализ этого разложения с помощью метода главных компонент и отбор ведущих компонент,
4. реконструкция одномерного ряда на основе отобранных компонент.

Преобразование ряда (2) к многомерному виду подразумевает его представление в матричной форме:

$$X = (x_{ij})_{i,j=1}^{k,L} = \begin{pmatrix} x_1 & x_2 & x_3 & \dots & x_L \\ x_2 & x_3 & x_4 & \dots & x_{L+1} \\ x_3 & x_4 & x_5 & \dots & x_{L+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_k & x_{k+1} & x_{k+2} & \dots & x_K \end{pmatrix},$$

где $L < K$ называется длиной "гусеницы", а $k = K - L + 1$.

Затем находятся собственные значения λ_i, $i = 1, 2, \dots, L$ и собственные вектора $\vec{V}_i$, $i = 1, 2, \dots, L$ ковариационной матрицы $C = \frac{1}{k} X X^T$. Матрица собственных векторов V используется для перехода к главным компонентам

$$Y = V^T X = (Y_1, Y_2, \dots, Y_L), \qquad (3)$$

где Y_i $(i = 1, 2, \dots, L)$ столбцы матрицы, состоящие из k элементов.

Равенство $\sum_{i=1}^{L} \frac{\lambda_i}{L} = \sum_{i=1}^{L} \alpha_i = 1$ позволяет оценить вклад α_i i-ой компоненты в анализируемый ряд.

На рис. 11 показан вклад α_i главных компонент в разложение исходного ряда трафика (в порядке убывания) при длине "гусеницы" $C_L = 12$ и 20. На основе этой информации можно оценить число ведущих компонент, определяющих характерное поведение трафика.

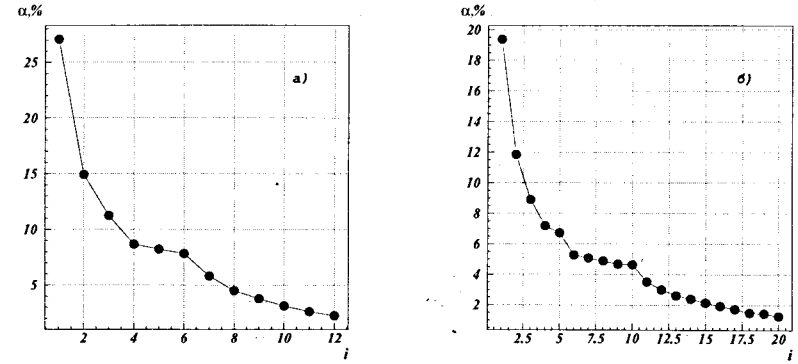

Рис. 11. Вклад α (в процентах) главных компонент в разложение исходного ряда трафика при длине "гусеницы" а) $C_L = 12$ и б) $C_L = 20$

Учитывая результаты, полученные в предыдущей главе, следует ожидать, что распределение информационного потока, восстановленного на основании ведущих компонент, должно описываться логнормальным законом. На рис. 12 представлены результаты аппроксимации распределений информационного потока, соответствующие различному числу $N = 1, 2, ..., C_L$ ведущих компонент при длине "гусеницы" $C_L = 20$, функцией (1). Здесь χ^2 это значение критерия χ^2, а ν число степеней свободы.

Прямые, параллельные оси абсцисс, показывают уровни значимости вероятность 10% соответствует верхней прямой $\chi^2/\nu = 1.247$, а вероятность 89.5% нижней прямой $\chi^2/\nu = 0.732$ при проверке нулевой гипотезы с числом степеней свободы $\nu = 47$. Из этой зависимости видно, что уже при $N = 3$ достигается достаточно высокий уровень соответствия статистического распределения гипотезе (1). В области больших N наблюдается рост χ^2, который можно объяснить влиянием остаточных компонент, носящих характер случайного шума.

Для оценки числа компонент, которые можно отбросить без влияния на основную составляющую трафика, весь набор главных компонент был разбит

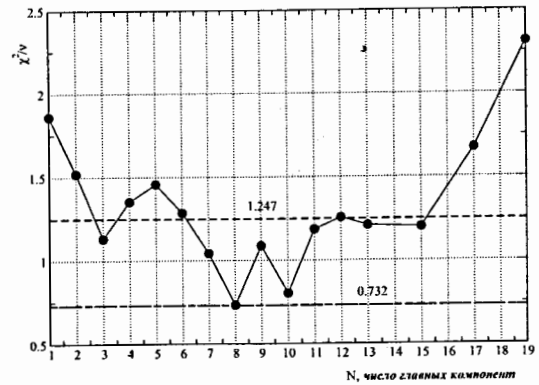

Рис. 12. Зависимость χ^2/ν от числа главных компонент

на две части:

1. ведущие компоненты, формирующие основную составляющую трафика, характеризующуюся логнормальным распределением,

2. остаточные компоненты, отвечающие части трафика с характеристиками случайного шума.

Для отбора остаточных компонент использовался "момент" нарушения симметрии распределения величин временного ряда, восстановленного на основании указанных компонент. Для проверки гипотезы о симметрии распределения был взят критерий знаков:

$$\mu = \sum_{i=1}^{n} \Theta(X_i), \tag{3}$$

где $X_1, \ldots, X_n$ - измерения трафика, n - объем выборки, а Θ - функция Хевисайда:

$$\Theta(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0. \end{cases}$$

Для правильной гипотезы и больших n распределение величин μ имеет вид:

$$P\{\mu \leq m \mid n, p\} \approx \Phi\left(\frac{m - np + 0.5}{\sqrt{np(1-p)}}\right),$$

где Φ - функция распределения нормального распределения (в нашем случае $p = 0.5$ и $n = 2048$).

На рис. 13 представлена зависимость значений величины μ от числа остаточных компонент (для $C_L = 12$ и 20). Видно, что число остаточных компо-

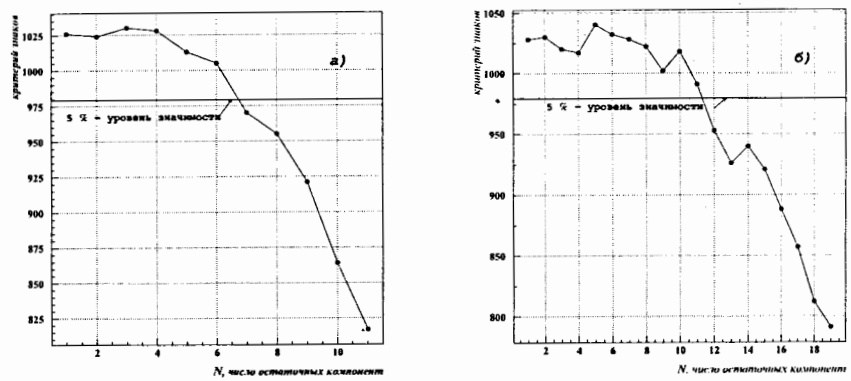

Рис. 13. Зависимость величин μ критерия знаков от числа остаточных компонент для а) $C_L = 12$ и б) $C_L = 20$

нент, отвечающих 5% уровню значимости, сооставляет 6 для $C_L = 12$ и 11 для $C_L = 20$.

Для дополнительной проверки этих результатов использовался критерий симметрии на основе статистики ω_n^2 [15]. Данный критерий проверяет симметрию относительно $x = 0$ функции распределения $F(x)$ измерений $X_1, \ldots, X_n$, т.е. нулевую гипотезу H_0: $F(x) = 1 - F(x)$. Соответствующая статистика ω_n^2 имеет вид:

$$\omega_n^2 = n \int_{-\infty}^{\infty} [F_n(x) + F_n(-x) - 1]^2 \, dF_n(x), \tag{4}$$

где $F_n(x)$ - эмпирическая функция распределения. Для расчетов статистики (4) удобнее пользоваться формулой:

$$\omega_n^2 = \sum_{j=1}^{n} \left[F_n(-X_{(j)}) - \frac{n - j + 1}{n}\right]^2,$$

где $X_{(1)} \leq \ldots \leq X_{(n)}$ вариационный ряд, построенный на основе измерений.

На рисунке 14 представлена зависимость ω_n^2 от числа остаточных компонент для $C_L = 12$ и 20.

Число остаточных компонент, отвечающих 5% - уровню значимости критерия, составляет 6 для $C_L = 12$ и 11 для $C_L = 20$, что полностью совпадает

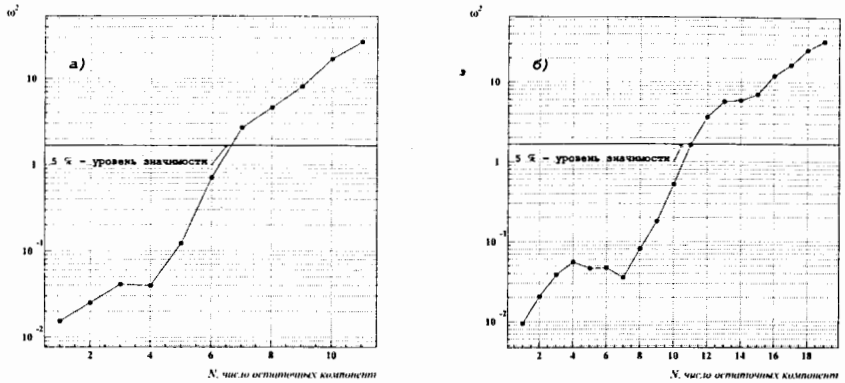

Рис. 14. Зависимость величины ω_n^2 от числа остаточных компонент для а) $C_L = 12$ и б) $C_L = 20$

с результатом, полученным с помощью критерия знаков (рис. 13). Это также согласуется с результатом, полученным с помощью критерия χ^2 (рис. 12).

Таким образом, статистический анализ измерений трафика, основанный на совместном использовании χ^2- и ω^2-критериев, позволил разбить набор главных компонент на две группы. Первая группа включает ведущие компоненты, ответственные за формирование основной составляющей трафика. Вторая группа, состоящая из остаточных компонент, может быть интерпретирована как шум. Детальный анализ пограничной области между этими двумя группами может дать дополнительную информацию о структуре трафика и упростить понимание его динамики.

В пятой главе сетевой трафик исследуется методами спектрального и вейвлет-анализа.

Для оценки числа вейвлет-коэффициентов, ответственных за формирование высокочастотной (шумовой) составляющей трафика, использовался критерий симметрии на основе ω_n^2. На рис. 15а представлена зависимость ω_n^2 от числа отброшенных наименьших вейвлет-коэффициентов M. Она имеет минимум при $M = 768$. Распределение восстановленного информационного потока при $M = 768$, представленное на рис. 15б, аппроксимируется логнормальной функцией (1) с хорошим уровнем значимости. Из рис. 15а видно, что максимальное число коэффициентов, которые можно отбросить, не превышая 5% - го уровня значимости, равно $M = 1408$. Это составляет $\approx 70\%$ от общего числа коэффициентов ($n = 2048$).

Для дополнительной проверки данного результата анализировалось поведение автокорреляционной функции [16]

$$C(\tau) = \frac{\sum_{i=1}^{K}(y_{i+\tau} - \bar{y})(y_i - \bar{y})}{\sum_{i=1}^{K}(y_i - \bar{y})^2}, \qquad \bar{y} = \frac{1}{K}\sum_{i=1}^{K} y_i, \tag{5}$$

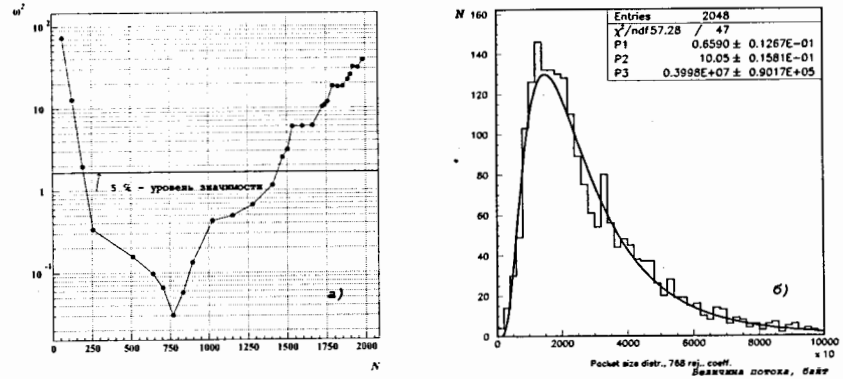

Рис. 15. а) Зависимость величины ω_n^2 от числа отброшенных вейвлет-коэффициентов, б) распределение потока информации, отвечающее отобранным коэффициентам после отбрасывания $M = 768$ наименьших коэффициентов

как для шумовой, так и регулярной составляющих сетевого трафика. Логично предположить, что элементы временного ряда, соответствующего шумовой составляющей, должны быть некоррелированными.

На рис. 16а представлены автокорреляционные функции для шумовой составляющей, отвечающие разному числу отброшенных коэффициентов M. Видно, что при $M \leq 1408$ отбрасываемая составляющая может рассматриваться как шум. На рис. 16б приведены автокорреляционные функции регулярной составляющей для разного числа отбрасываемых коэффициентов. Видно, что исключение не более, чем 1408 наименьших коэффициентов, практически не влияет на форму автокорреляционной функции.

К отфильтрованным данным была снова применена процедура обработки на основе подхода "Гусеница", описанная в главе 4. На рис. 17 представлен вклад α_i (в процентах) главных компонент для данных трафика после исключения 1408 наименьших коэффициентов. Видно, что вклад остаточных компонент значительно уменьшился по сравнению с результатами для исходных измерений, в тоже время вклад ведущих компонент заметно вырос (ср. рис. 11).

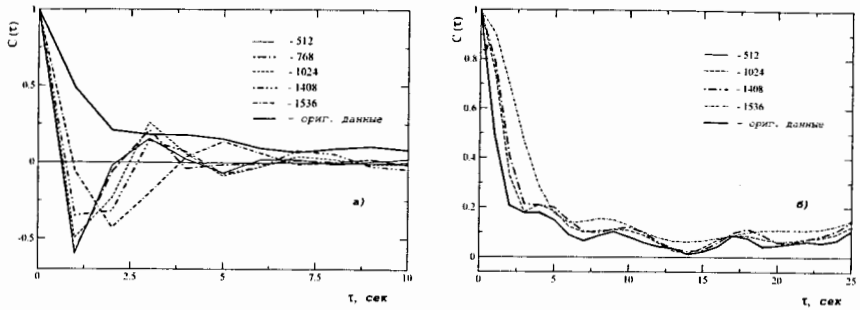

Рис. 16. Автокорреляционные функции $C(\tau)$ для а) шумовой и б) регулярной составляющих, соответствующих различному числу отброшенных коэффициентов

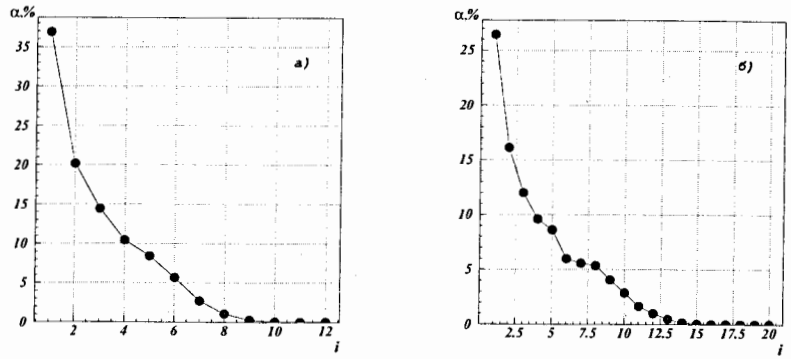

Рис. 17. Вклад α (в процентах) главных компонент в разложение отфильтрованного ряда трафика при длине "гусеницы" а) $C_L = 12$ и б) $C_L = 20$

На рис. 18 представлены результаты аппроксимации распределения информационного потока для отфильтрованных данных (для разных $N = 1, 2, ..., C_L$ при $C_L = 20$) функцией (1). Видно, что уже три ведущие компоненты формируют распределение, которое наилучшим образом согласуется с гипотезой (1).

Для оценки числа остаточных компонент, которые можно исключить из измерений трафика без заметного влияния на основную составляющую, вновь использовался критерий симметрии на основе ω_n^2. На рис. 19 приведена за-

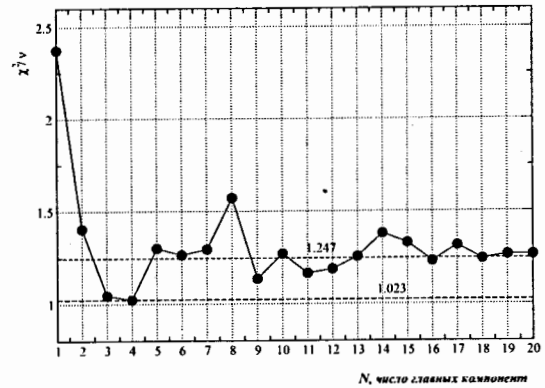

Рис. 18. Зависимость величины χ^2/ν от числа N главных компонент

висимость ω_n^2 от числа остаточных компонент для отфильтрованных данных при длине "гусеницы" $C_L = 20$. Величина ω_n^2 превышает граничное значение,

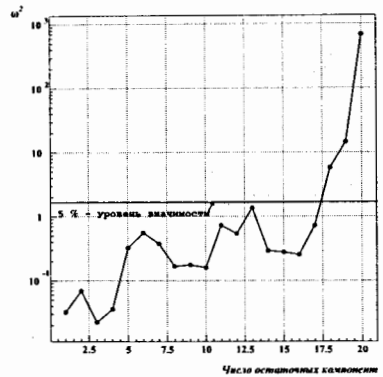

Рис. 19. Зависимость величины ω_n^2 от числа остаточных компонент для отфильтрованных данных при длине "гусеницы" $C_L = 20$

соответствующее 5% - уровню значимости, при числе остаточных компонент более 17 (для исходных данных - больше 10). Этот результат показывает, что после вейвлет-фильтрации только 3 компоненты формируют основную составляющую трафика, а 17 остаточных компонент могут быть исключены как шумовые (из общего числа $C_L = 20$). Это находится в согласии с результатом, полученным на основе критерия χ^2 (рис. 18). Также показано, что ряд, восстановленный на основе этих трех компонент, сохраняет основные спектральные характеристики исходного ряда измерений информационного

трафика. Это позволяет предположить, что преобразования, произведенные над исходным рядом, не нарушают основных свойств трафика.

В шестой главе развиты новые методы для определения моментов смены состояния анализируемого временного ряда. В их основу положена гипотеза о том, что в установившемся режиме при определенном уровне агрегации распределение информационного потока отвечает логнормальному закону, а изменение состояния системы, связанное, например, с увеличением активности пользователей, приводит либо к подобному режиму, но с другими параметрами логнормального распределения, либо к переходному режиму, распределение информационного потока в котором не отвечает логнормальному закону (например, при перегрузках, или в случае сетевых атак).

На рис. 20а представлены временные ряды для разных уровней агрегации 0.1, 1 и 10 сек для данных, полученных на входном шлюзе сети университета "Дубна" (32 часа измерений).

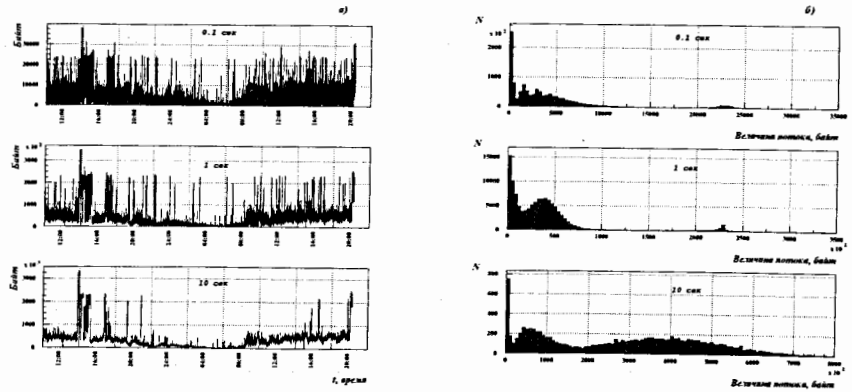

Рис. 20. а) Временные ряды агрегированного (при 0.1, 1 и 10 сек) потока информации (включая дневное и ночное время), б) распределения агрегированного информационного потока при 0.1, 1 и 10 сек

Данные демонстрируют достаточно нестабильный характер, что, в свою очередь, подразумевает определенные изменения свойств внутренней динамики системы. Интервалы, соответствующие различным режимам функционирования системы, могут рассматриваться на временной оси как состояния "до" и "после" соответствующих моментов смены режима. В предыдущих главах было показано, что для достаточно коротких временных интервалов в условиях отсутствия перегрузок, сетевых атак и т.д. (т.е. критических режимов), распределение информационных потоков соответствует логнормальному распределению.

Очевидно, что для продолжительных интервалов параметры распределения не могут оставаться неизменными (самый простой пример - это падение сетевой активности в ночное время). Естественно предположить, что в этом случае временной ряд измерений можно разбить на определенное количество интервалов, каждый из которых соответствует своему режиму функционирования.

Статистические распределения для соответствующих временных рядов представлены на рис. 20б. В отличие от ранее рассмотренных примеров, увеличение уровня агрегации не приводит к формированию распределения, отвечающего единственному логнормальному распределению. Вместо этого наблюдается распределение, представляющее собой сумму разных распределений. Составляющие его распределения, в соответствии с нашей гипотезой, должны отвечать различным режимам функционирования сети. Для того, чтобы исследовать эти режимы, необходимо решить задачу их разделения.

В работе развит новый подход для определения моментов смены состояния системы на основе принципов иммунокомпьютинга [18]-[20]. Его можно рассматривать как расширение алгоритма "отбора от противного". Предлагаемый алгоритм можно сформулировать следующим образом:

- определяется набор выборок заданного объема n, формируемых из последовательных величин временного ряда;
- строится вектор признаковых параметров, характеризующих анализируемый процесс;
- определяется множество "свой" в виде набора векторов, отвечающих основному режиму процесса;
- задается правило соответствия, определяющее отличие "своего" вектора от "чужого".

Анализ данных показал, что использование двумерных векторов позволяет свести задачу к поиску и классификации кластеров на плоскости.

Для анализа использовались два типа векторов и, соответственно, два метода разделения кластеров, применяемых последовательно. В первом методе (методе временных задержек) использовалось двумерное распределение на фазовой плоскости для "задержанных" координат $(x_i, x_{i+\tau})$ с последующей классификацией различных аттракторов, как "нормальных" (своих) или "аномальных" (чужих).

Во втором методе использовались две величины - среднее и дисперсия скользящей выборки объема $n = 10$, определяющих вектор (μ_j, σ_j). Также, как и в первом методе, использовалось разделение кластеров на плоскости (μ, σ) на "свои" и "чужие".

Оба метода имеют в своей основе утверждение, что изменения в динамике должны вести к специфичным изменениям в статистическом распределении

величин временного ряда. Оба метода имеют в качестве "целевой функции" принадлежность последовательно разделяемых классов (состояний временного ряда) к классу логнормальных распределений.

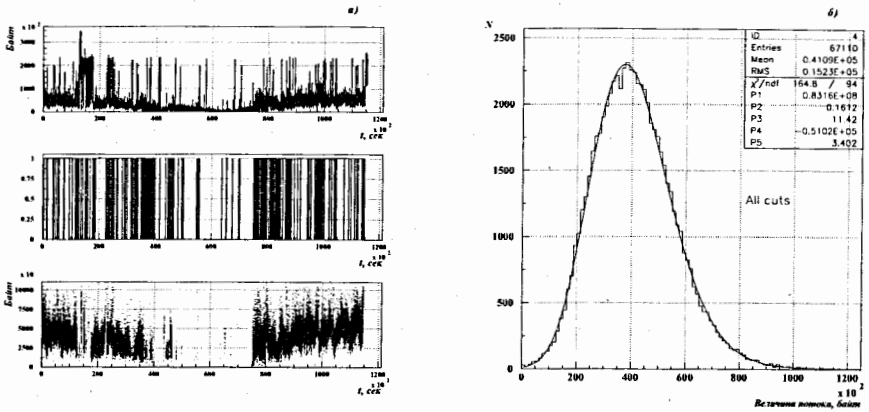

Рис. 21. а) Временной ряд агрегированного на уровне 1 сек трафика (вверху). Моменты структурных изменений (средняя диаграмма). "Отфильтрованный" трафик - дневное состояние (внизу), б) статистическое распределение для дневного трафика и его аппроксимация логнормальной функцией

На рис. 21а вверху приведен временной ряд трафика (32 часа измерений), агрегированный на уровне 1 сек. "Отфильтрованный" трафик - основное (дневное) состояние общей длительностью 18.5 часов (58% от общего времени измерений) показан на нижнем рис. 21а. Моменты структурных изменений представлены на диаграмме на среднем рис. 21а. В средней части этой диаграммы отчетливо выделяется интервал, где количество структурных изменений не так велико. Этот интервал соответствует ночному времени. На рис. 21б представлено распределение сетевого трафика, находящегося в основном (дневном) состоянии. На этом же рисунке представлен результат аппроксимации полученного эмпирического распределения функцией (1).

На рис. 22а представлены временные ряды, отвечающие ночному режиму, для разных уровней агрегации (0.1, 1 и 10 сек). На рис. 22б представлено распределение сетевого трафика, агрегированного на уровне 10 сек, с наложенной фитирующей кривой логнормального распределения.

На рис. 23а приведен пример хакерской атаки на один из компьютеров локальной сети университета "Дубна". Для локализации атаки использовался метод на основе классификации векторов (μ, σ), соответствующих "скользящей" выборке объема $n = 20$ при уровне агрегации 1 сек. На диаграмме рассеяния (рис. 23б) хорошо видно, что кластер, характеризующий основное

состояние системы, расположен в области средних значений μ, в то время как кластер, отвечающий состоянию системы в период атаки, расположен на диаграмме слева в области малых μ и легко отделяется от других кластеров (состояний). Векторы (μ, σ), попадающие в эту область, характеризуются как "чужие". На нижнем рис. 23а представлен исследуемый ряд после исключения из него фрагмента атаки. Следует отметить, что в рассматриваемом

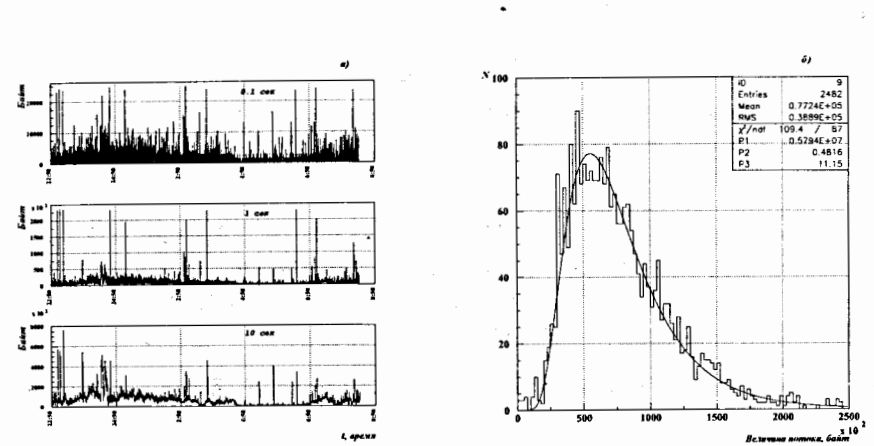

Рис. 22. а) Агрегированный при 0.1, 1 и 10 сек (сверху вниз) трафик (ночное время), б) распределение агрегированного на уровне 10 сек "ночного" трафика и его аппроксимация логнормальной функцией

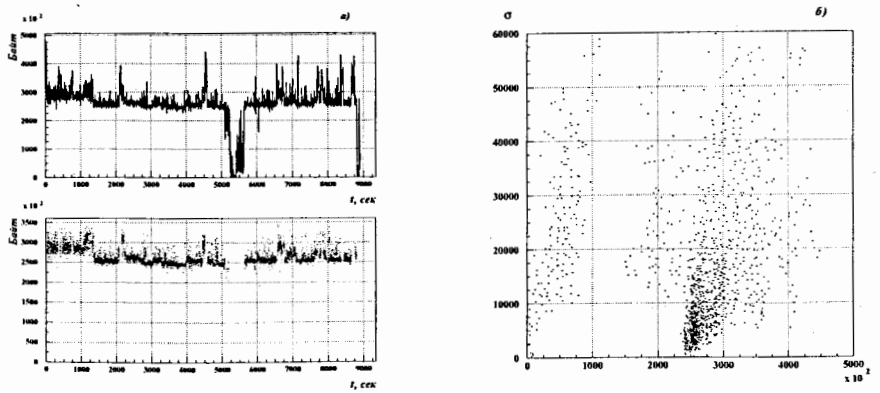

Рис. 23. а) Сверху-вниз: измерения трафика, агрегированные с окном 1 сек; временной ряд трафика после исключения из него участка с хакерской атакой, б) диаграмма рассеяния двух величин – среднего значения и дисперсии скользящей выборки объема $n = 20$

примере простая дискриминация по величине потока также позволяет разделить "нормальную" и "аномальную" моды в динамике трафика, однако в более сложном случае этого может оказаться недостаточно.

В Заключении сформулированы основные результаты диссертации, дается краткое описание работ, положенных в ее основу.

Список публикаций

[A1] Антониу Я., Иванов В.В., Иванов Валерий В., Зрелов П.В.: *Статистическая модель информационного трафика*, "Физика элементарных частиц и атомного ядра" (ЭЧАЯ). 2004. Т.35. Вып.4. С.984-1019 (на англ. яз.).

[A2] Антониу Я., Иванов В.В., Иванов Валерий В., Зрелов П.В.: *Анализ главных компонент измерений информационного трафика: подход "Caterpillar"-SSA*, "Письма в ЭЧАЯ", 2004, Т.1, №4 (121). С.87-95 (на англ. яз.).

[A3] I. Antoniou, V.V. Ivanov, Valery V. Ivanov, and P.V. Zrelov: *On the Log-Normal Distribution of Network Traffic*, Physica D 167 (2002) 72-85.

[A4] I. Antoniou, V.V. Ivanov, Valery V. Ivanov and P.V. Zrelov: *On a Statistical Model of Network Traffic*, "Nuclear Instruments & Methods in Physics Research", A 502 (2003) 768-771.

[A5] I. Antoniou, V.V. Ivanov, Valery V. Ivanov and P.V. Zrelov: *Wavelet Filtering of Network Traffic Measurements*, Physica A 324 (2003) 733-753.

[A6] I. Antoniou, Victor V. Ivanov, Valery V. Ivanov, Yu.L. Kalinovsky and P.V. Zrelov: *On a Kinetic Model of the Internet Traffic*, "Discrete Dynamics in Nature & Society", 2004:1 (2004) 19-34.

[A7] P. Zrelov, I. Antoniou, V. Ivanov, Valery Ivanov: *Principal Component Analysis of Network Traffic: the "Caterpillar"-SSA Approach*, VIII Int. Workshop on "Advanced Computing and Analysis Techniques in Physics Research" - ACAT'2002, 24-28 June, 2002, Moscow, RUSSIA, Book of abstracts, p. 176.

[A8] I. Antoniou, V. Ivanov, Valery Ivanov and P. Zrelov: On a Statistical Model of Network Traffic, VIII Int. Workshop on *"Advanced Computing and Analysis Techniques in Physics Research"* - ACAT'2002, 24-28 June, 2002, Moscow, RUSSIA, Book of abstracts, p. 177.

[A9] I. Antoniou, V.V. Ivanov, Valery V. Ivanov and P.V. Zrelov: *Wavelet Filtering of Network Traffic Measurements*, V Int. Congress on Mathematical Modeling, September 30-October 6, 2002, Book of abstracts, Vol. I, p. 137, Dubna, Moscow region, Russia, 2002.

[A10] I. Antoniou, V.V. Ivanov, Valery V. Ivanov and P.V. Zrelov: *Statistical Model of Network Traffic*, V Int. Congress on Mathematical Modeling, September 30-October 6, 2002, Book of abstracts, Vol. I, p. 138, Dubna, Moscow region, Russia, 2002.

[A11] I. Antoniou, V.V. Ivanov, Valery V. Ivanov and P.V. Zrelov: *Statistical Model of Network Traffic*, XIX International Symposium on Nuclear Electronics & Computing, NEC'2003, September 15-20, 2003, Varna, Bulgaria, Book of abstracts, Dubna, 2003, p. 14.

[A12] V.V. Ivanov, Valery V. Ivanov, Yu.A. Kryukov and P.V. Zrelov: *Detection of abrupt changes in network traffic dynamics*, In: Int. Conf. "Distributed computing and Grid-technologies in science and education", Dubna, June 29 - July 2, 2004, Book of abstracts, p. 86.

[A13] V.V. Ivanov, Valery V. Ivanov, Yu.L. Kalinovsky and P.V. Zrelov: *Statistical and kinetic models of Internet traffic flows*, In: Int. Conf. "Distributed computing and Grid-technologies in science and education", Dubna, June 29 - July 2, 2004, Book of abstracts, p. 87.

[A14] I. Antoniou, V.V. Ivanov, Valery V. Ivanov and P.V. Zrelov: *Principal Component Analysis of Network Traffic*, In: Proc. of I-st Int. Conf. on "Mathematics and Informatics for Industry", MII 2003, 14-16 April 2003, Thessaloniki, Greece, pp. 170-181.

[A15] I. Antoniou, V.V. Ivanov, Valery V. Ivanov and P.V. Zrelov: *Statistical Model of Network Traffic*, JINR Communication, E11-2002-222, JINR, Dubna, RUSSIA, 2002, 38 pp.

[A16] I. Antoniou, V.V. Ivanov, Valery V. Ivanov and P.V. Zrelov: *Wavelet Filtering of Network Traffic Measurements*, JINR Communication, E11-2002-223, JINR, Dubna, RUSSIA, 2002, 22 pp.

[A17] I. Antoniou, V.V. Ivanov, Valery V. Ivanov and P.V. Zrelov: *Principal Component Analysis of Network Traffic Measurements*, Preprint JINR, E11-2003-148, JINR, Dubna, RUSSIA, 2003, 15 pp.

[A18] Victor V. Ivanov, Valery V. Ivanov, Yu.L. Kalinovsky, P.V. Zrelov, I. Antoniou: *Statistical and Kinetic Models of Network Traffic*, In: "Annual

report 2003. Laboratory of Information Technologies". Ed. by Gh. Adam, V.V. Ivanov and T.A. Strizh, JINR, Dubna, 2004, pp. 28-31.

[A19] Я. Антониоу, П.В. Зрелов, В.В. Иванов, Валерий В. Иванов, Ю.Л. Калиновский: *Статистическая и кинетическая модели сетевого трафика*, Новости ОИЯИ, 3/2004, стр. 32-35.

[A20] V.V. Ivanov, Valery V. Ivanov, Yu.A. Kryukov and P.V. Zrelov: *Detection of Abrupt Changes in Network Traffic Dynamics*, In: "Annual report 2004-2005 years. Laboratory of Information Technologies". Ed. by Gh. Adam, V.V. Ivanov and T.A. Strizh, JINR, 2005-179, Dubna, 2005, pp. 66-72.

Литература

[1] W Leland, M.Taqqu, W. Willinger, and D.Wilson: *On the Self-Similar Nature of Ethernet Traffic (Extended Version)*, IEEE/ACM Transactions on Networking, 2(1), pp. 1-15, February 1994.

[2] M.T. Lucas, D.E. Wrege, B.J. Dempsey, and A.C. Weaver: *Statistical Characterization of Wide-Area Self-Similar Network Traffic*, University of Virginia Technical Report CS97-04, October 9, 1996.

[3] M.E. Crovella and A. Bestavros: *Self-Similarity in World Web Traffic: Evidence and Possible Causes*, IEEE/ACM Transactions on Networking, Vol.5, No. 6, pp. 835-846, December 1997.

[4] Vishal Misra and Wei-Bo Gong: *A Hierarchical Model for Teletraffic*, Department of Electrical and Computer Engineering, University of Massachusetts, Amherst MA 01003, 1998.

[5] Jon M. Peha: *Protocals Can Make Traffic Appear Self-Similar*, In: Proc. of the 1997 *IEEA/ACM/SCS* Communication Networks and Distributed Systems Modeling and Simulation Conference.

[6] A. Erramilli, P. Pruthi and W. Willinger: *Resent Developments in Fractal Traffic Modelling*, In: Proc. Inter. Teletraffic Seminar, St. Petersburg, 26 June 2 July, 1995.

[7] D.L. Jagerman, B. Melamed, and W. Willinger: *Stochastic Modeling of Traffic Processes*, Technical Report, 1996.

[8] S.M. Kay: *Modern Spectral Estimation: Theory and Applications.* Prentice Hall, New Jersey, 1988.

[9] M.B. Priestley: *Non-linear and Non-stationary Time Series Analysis.* Academic Press, 1988.

[10] Международный Университет природы, общества и человека "Дубна": http://www.uni-dubna.ru.

[11] П.М. Васильев, В.В. Иванов, В.В. Корешков, Ю.А. Крюков, С.И. Купцов: *Система сбора, анализа и управления сетевым трафиком фрагмента сети ОИЯИ на примере подсети Университета "Дубна"*, Сообщение ОИЯИ, Д11-2001-266, Дубна, 2001.

[12] Колмогоров А.Н.: *О логарифмически нормальном законе распределения размеров частиц при дроблении*, Доклады АН СССР. 1941. Т.31. С.99-101.

[13] Данилов Д.Л., Жиглявский А.А., редакторы: *Главные компоненты временных рядов: метод "Гусеница"*. Изд-во СПбГУ, 1997.

[14] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky: *Analysis of time series structure: SSA and related techniques*, Chapman & Hall/CRC, 2001.

[15] Г.В. Мартынов: *Критерии омега-квадрат*, Москва, "Наука", 1978.

[16] Henry D.I. Abarbanel: *Analysis of Observed Chaotic Data*, 1996 Springer-Verlag New York, Inc.

[17] E.L. Crow, K. Shimizu (eds.): *Lognormal Distributions. Theory and Applications*, Marcel Dekker, Inc., New York, 1988.

[18] D. Dasgupta: *"An Overview of Artificial Immune Systems and Their Applications*, In: Artificial Immune Systems and Their Applications, Springer-Verlag Berlin Heidelberg 1999, 3-21, 1999.

[19] D. Dasgupta and Nii Attoh-Okine: *Immunity-Based Systems: A Survey*, In: Proc. of the IEEE Int. Conf. on Systems, Man, and Cybernetics, Orlando, October 12-15, 1997.

[20] S. Forrest, A.S. Perelson, L. Allen, and R. Cherukuri: *Self-Nonself Discrimination in a Computer*. In: Proc. of IEEE Symposium on Research in Security and Privacy, pp. 202-212, Oakland, CA, 16-18 May 1994.

Получено 31 марта 2009 г.