УДК 621.394, 621.395; 004.056, 004.512.2, 004.588

Ю.Н. Минаев, д-р техн. наук
Национальный авиационный университет
(Украина, 03057, Киев, пр-т космонавта Комарова, 1,
тел. (044) 2495454, e-mail: minaev@rambler.r),
О.Ю. Филимонова, Ю.И.Минаева, кандидаты техн. наук
Киевский национальный университет строительства и архитектуры
(Украина, 03037, Киев, Воздухофлотский пр-т, 31,
тел. (044) 2486427, e-mail: filimonova @nm.ru,
тел. (044) 2425462, e-mail: jumin @big-mir.net)

Идентификация аномалий трафика компьютерных систем на основе методики структурирования многомерного трафика

Идентификация аномальных состояний трафика компьютерной системы при его представлении в виде многомерного временного ряда, структурированного покомпонентно и попакетно, дает возможность получить структурные характеристики потока данных, используемые в дальнейшем для идентификации. Показана возможность применения *p*-адических моделей для анализа трафика при использовании в качестве динамического паттерна состояния предшествующего потока данных. Приведены примеры, свидетельствующие об эффективности предложенной методики.

Ідентифікація аномальних станів трафіка компьютерної системи при його представленні у вигляді багатовимірного часового ряду, структурованого покомпонентно та попакетно, дає можливість отримати структурні характеристики потоку даних, які використовуються в подальшому для ідентифікації. Показано можливість застосування *p*-адичних моделей для аналізу трафіка при використанні в якості динамічного паттерна стану попереднього потоку даних. Наведено приклади, які свідчать про ефективність запропонованої методики.

Ключевые слова: многомерный трафик, интеллектуальный анализ данных, пакет, компонент трафика, p-адическая модель, бинарное дерево, фрактальная размерность, фрактальное число.

Проблема безопасного функционирования компьютерных систем (КС) в настоящее время многократно усложнилась. Попытки внедрения в чужие сети стали настолько изощренными, что все методы идентификации таких вторжений, основанные на выделении признаков, т.е. величин компонент трафика, практически исчерпали себя. Можно привести множество приме-

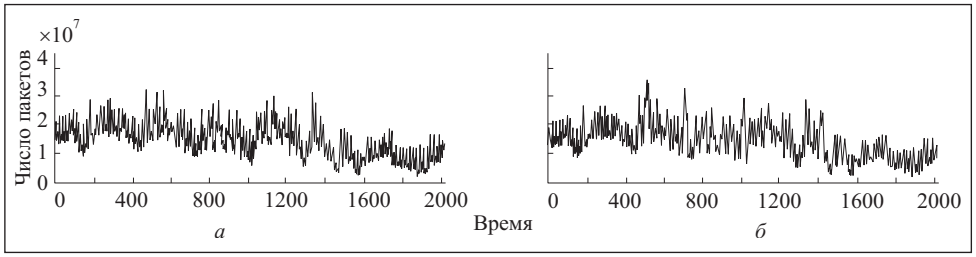


Рис. 1. Графики сетевого трафика при наличии атаки (*а*) и без нее (*б*) [1]

ров, когда аномальный и нормальный режимы трафика имеют одинаковые параметры признаков. Рассмотрим пример из работы [1].

По мнению авторов работы [1], обнаружить разницу в двух приведенных на рис. 1 графиках, практически невозможно. Следовательно, невозможно обнаружить аномальные компоненты в сетевом трафике при такой постановке задачи, т.е. во временном интервале. Современные схемы организации атак и других противозаконных действий весьма сложны. Поэтому при попытке идентификации аномалий, например посредством распознавания образов, требуется обнаруживать сложные комбинации сетевых транзакций и сопутствующих им факторов, вследствие чего обучающее множество становится сверхбольшим, так как его приходится формировать по комбинациям двух, трех и более транзакций. Таким образом, исходная задача становится практически неразрешимой.

Современное состояние исследований. В настоящее время при решении проблемы безопасности КС основное внимание уделяется многомерному трафику КС, так как учет всех (или некоторого числа) компонент и их взаимосвязей позволяет идентифицировать аномалию, а масштабируемое пространство — определить необходимые параметры. Современный подход к решению обобщенной задачи идентификации аномалий наиболее полно изложен в работе [2]. Считается, что традиционные методы интеллектуального анализа данных (ИАД) не приспособлены для решения подобных задач, так как требуют наличия большого количества обучающего материала в виде положительных и отрицательных примеров. Трудность состоит в отсутствии достаточного объема обучающих данных по аномалиям трафика КС (включая компьютерные атаки, несанкционированные действия и др.). Тем не менее, современные тенденции в данной области исследований связаны именно с возможностями ИАД.

В работе [2] показано, что число обучающих случаев значительно меньше общего числа случаев. Поэтому возникла проблема обнаружения редких, или дисбалансных, закономерностей, что, естественно, потребо-

вало развития новых технологий ИАД, в частности технологии, названной обнаружением взаимосвязей (OB-Link Discovery), с помощью которой исследуется возможность обнаружения взаимосвязей между несопоставимыми, редкими и дисбалансными явлениями среди большого числа потоков и источников данных.

Эта технология воплощена в разработках CISCO [3]. Важным ее элементом являются отчеты о движении потоков трафика, в которых происходит постоянное сопоставление сетевого трафика с базовыми нормами сети. Такие отчеты можно получать, используя технологию Cisco Net-Flow, согласно которой сбор информации о потоках трафика и выявление угрозы атак рассматривается как обнаружение аномалий.

Cisco Net-Flow классифицирует пакеты по потокам, причем каждый пакет имеет семь уникальных характеристик: входной интерфейс; тип протокола IP; байт типа сервиса (ToS); исходный и целевой IP-адреса; номера исходного и целевого портов. Эти характеристики дают достаточно информации для создания базового профиля нормальных закономерностей поведения трафика. Составляя детальный отчет о потоках трафика, Cisco Net-Flow позволяет пользователям IT выявлять отклонения от типовых закономерностей поведения трафика, т.е. ранние признаки потенциальных атак (DDoS) [4].

В работе [5] рассмотрены аварийный сетевой трафик и обнаруживающий метод, основанный на понятии системный прототип, а также предложен алгоритм обнаружения, в котором использованы изменения в образцах трафика, появляющиеся в течение атаки при собирании пакетов, принадлежащих идентичному потоку. Этот алгоритм позволяет обнаружить даже мутант-атаки, использующие новый номер порта или измененную нагрузку, пока сигнатуры системы не способны обнаружить эти типы атак.

Обнаружение аномалии через заголовок пакета предполагает детектирование части заголовка, т.е. проверки величин, хранящихся в области заголовка пакета, что позволяет классифицировать атаку величинами области заголовка. При этом поток определяется как множество пакетов с кортежем, состоящим из пяти элементнов: исходный IP адрес; расположение IP адреса; исходный порт; порт расположения; номер протокола.

Многомерный трафик КС основан на использовании теории тензоров [6]. В работе [7] предложено использовать тензорные декомпозиции как средство моделирования и сигнального процессинга для решения проблемы обработки массивов. Например, если принятый сигнал — трехпорядковый тензор, то это означает, что каждый принятый сигнал должен быть связан с данной системой координат в трехмерном пространстве, т.е.

моделирование следует выполнить с помощью трех индексов, каждый из которых характеризует систематизированную вариацию принятого сигнала.

Пространственная размерность соответствует числу приемных антенн, а временная размерность — числу символов, обрабатываемых в приемнике. Дополнительная (третья) размерность относится к частному типу обработки, выполняемой в передатчике (источнике) или приемнике, или и в том и в другом одновременно. Например, для систем CDMA трехразмерность диверсификации есть расширение пространства, возникающего при расширении кодов в передатчике.

В работе [8] предложен метод динамического тензорного анализа (ДТА), обеспечивающий компактный анализ для высокопорядковых и высокоразмерных данных. Метод ДТА позволяет проверить динамическое поведение трафика, идентифицировать аномалии как потенциальное вторжение или атаку, а также определять корреляцию между различными источниками, приемниками и портами. Динамический и потоковый тензоранализы позволяют разложить исходные тензоры высокой размерности в тензоры «сердцевины» и «матрицы проекций» (для каждого направления) значительно меньшей размерности.

В работе [9] описан метод мониторинга сети реального времени в терминах цифровых времязависимых функций параметров протокола. На основании комплексной теории систем трафик описан как траектория в мультимерном пространстве время—параметр с 10—12 измерениями. Для описания информационных потоков в сети введено понятие биржи пакетов между компьютерами, структура пакетов и их размеры изменяются в зависимости от процесса. Каждый пакет состоит из заголовка и инкапсулированных данных. Поскольку часть данных не влияет на распространение пакета через сеть, предложено рассматривать только информацию, включенную в заголовок, состоящий из инкапсулированных протоколов, связанных с различными слоями коммуникаций — от слоя связи до прикладного слоя. Информация, содержащаяся в заголовках, управляет трафиком всей сети. Для того чтобы извлечь эту информацию, используют tcpdump-утилиты [10]. Полученная информация используется для анализа сетевого трафика, определения сигнатуры аномального сетевого поведения и обнаружения возможных вторжений.

Важное отличие предлагаемого метода от традиционно используемых — представление информации, содержащейся в заголовках, в виде хорошо определенных функций [19]. Для того чтобы читать двоичные tcpdump-файлы и представлять все параметры протокола в соответствии с изменяющимися во времени функциями, разработано программное обеспечение.

Постановка задачи. Кластерный анализ (КА) как часть ИАД широко применяется при решении достаточно большого класса задач, например, в

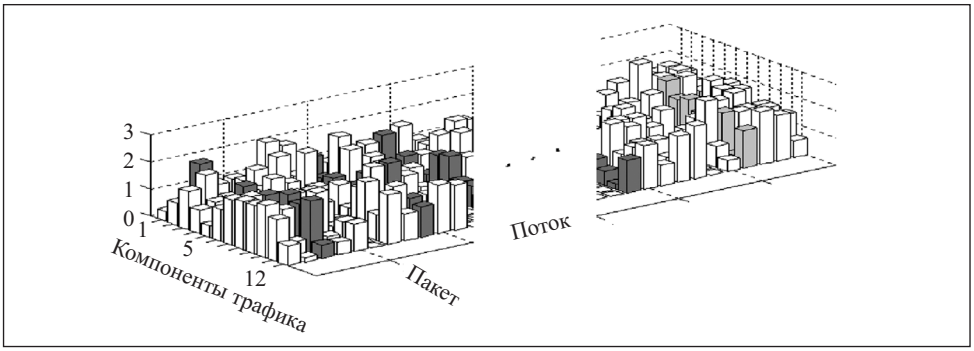


Рис. 2. Расчетная модель трафика (темным цветом выделены аномальные пакеты)

биоинформатике. Однако в задачах анализа трафика КС КА используется недостаточно [11]. Существующие системы мониторинга трафика, как правило, учитывают 3—9 или максимум 12 его параметров. Обычно выполняется мониторинг каждого параметра трафика во времени на основании множества измерений $\{x_i^{(j)}(t)\}$, где i — номер компоненты, j — номер измерения, с последующей статистической обработкой. Вычисляется интервал возможных колебаний для нормального и аномального состояний трафика и разрабатываются алгоритмы оценки вероятности попаданий реальной величины в тот или иной интервал, на основании чего делается вывод о характере трафика (рис. 2).

Недостаток существующих методов заключается в том, что, как правило, ограничиваются покомпонетным анализом трафика. Однако результаты практических исследований свидетельствуют о недостаточности такого подхода, так как поведение интегрального трафика (имеющего 3—12 компонент) отличается от поведения однокомпонентного трафика. Кроме того, важное значение с точки зрения получения новых знаний имеет анализ множества пакетов, поступивших за единицу времени, в частности анализ их структуры. Если вектор трафика в момент t^0 обозначить через $\mathbf{x}^0$, множества векторов нормального и аномального трафиков — через $\mathbf{x}^N$ и $\mathbf{x}^A$ и полагать, что их компонентами есть интервалы, $\mathbf{x}^N = \{{}^N\mathbf{I}_i^x\}$, $\mathbf{x}^A = \{{}^A\mathbf{I}_i^x\}$, $i = 1, n, \mathbf{I}_i^x = [x_i^{\min}, x_i^{\max}]$, $i = 1, n$, то задачу идентификации аномальных состояний трафика КС можно рассматривать как задачу определения близости векторов $\{\mathbf{x}^0 \in \mathbf{X}^N\}$ и $\{\mathbf{x}^0 \in \mathbf{X}^A\}$, или задачу оценки принадлежности к заданной области пространства.

Алгоритм решения задачи. Предлагаемый новый метод идентификации аномальных состояний заключается в следующем. Для интервала времени $T^{(p)} = [t_1, t_2, ..., t_f]^{(p)}$, в течении которого выполняется монито-

ринг КС, формируется матрица. Ее строками являются пакеты потока данных, а столбцами — компоненты трафика для значений $t_j \in T^{(p)}$:

$$\begin{pmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \vdots \\ \mathbf{x}^{(f)} \end{pmatrix} = (\mathbf{x}^{(1)}, \ \mathbf{x}^{(2)}, \ \cdots, \ \mathbf{x}^{(f)}) = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ & & \ddots & \\ x_1^{(f)} & x_2^{(f)} & \cdots & x_n^{(f)} \end{pmatrix}.$$

Для полученной матрицы методом иерархического кластерного анализа (ИКА) строится бинарное дерево (БДр), или дендрограмма, и вычисляются характеристики структуры БДр, полученного на основе матрицы близости (расстояний),

$$D = \begin{pmatrix} 0\, d(\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \cdots d(\mathbf{x}^{(1)}, \mathbf{x}^{(f)}) \\ d(\mathbf{x}^{(2)}, \mathbf{x}^{(1)})\, 0 \cdots d(\mathbf{x}^{(2)}, \mathbf{x}^{(f)}) \\ \vdots \\ d(\mathbf{x}^{(f)}, \mathbf{x}^{(1)})\, d(\mathbf{x}^{(f)}, \mathbf{x}^{(2)}) \cdots 0 \end{pmatrix} \to S,$$

при покомпонентной (по столбцам) или попакетной (по строкам) кластеризации. Характеристиками рассматриваемой структуры являются многоуровневое расположение кластеров, кластеры каждого уровня и их объединения (собственно структура), а также фрактальное число и фрактальная размерность закодированного БДр.

Далее формируется новая матрица для нового интервала времени $T^{(p+1)} = [t_1, t_2, ..., t_f]^{(p+1)}$, в течении которого выполняется мониторинг КС $(\mathbf{x}^{(2)}, \mathbf{x}^{(3)}, ..., \mathbf{x}^{(f)})$. Для полученной матрицы вычисляется новое множество характеристик структуры БДр $S^{(p+1)}$. Если характеристики $S^{(p)}$ и $S^{(p+1)}$ близки, осуществляется нормальный трафик, в противном случае возникает аномалия (рис. 3). Таким образом, тест-множество трафика, если оно изначально было нормальным, пополняется и система, по существу, становится саморегулируемой.

У п р о щ е н н ы й а л г о р и т м.

1. В начальный момент времени формируем множества значений, представляющие собой:

однокомпонентный трафик — $X^{(i)}(t) = \{x_j^{(i)}\}$, где i — номер компоненты, i = 1 ÷ 5 (трафик CISCO) или i = 1 ÷ 9 (общее число компонент трафика) ; j— номер сканирования, j = 1, 2,… .

обобщенный многомерный трафик, в котором одновременно учтены все компоненты.

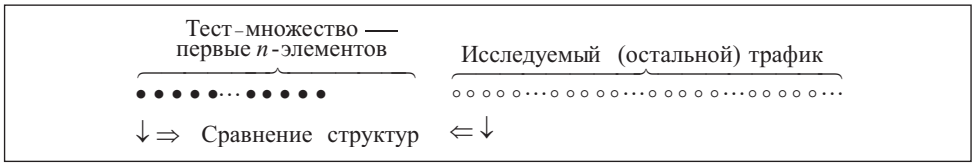


Рис. 3. Схема поддержки принятия решения о появлении аномалии в текущем трафике

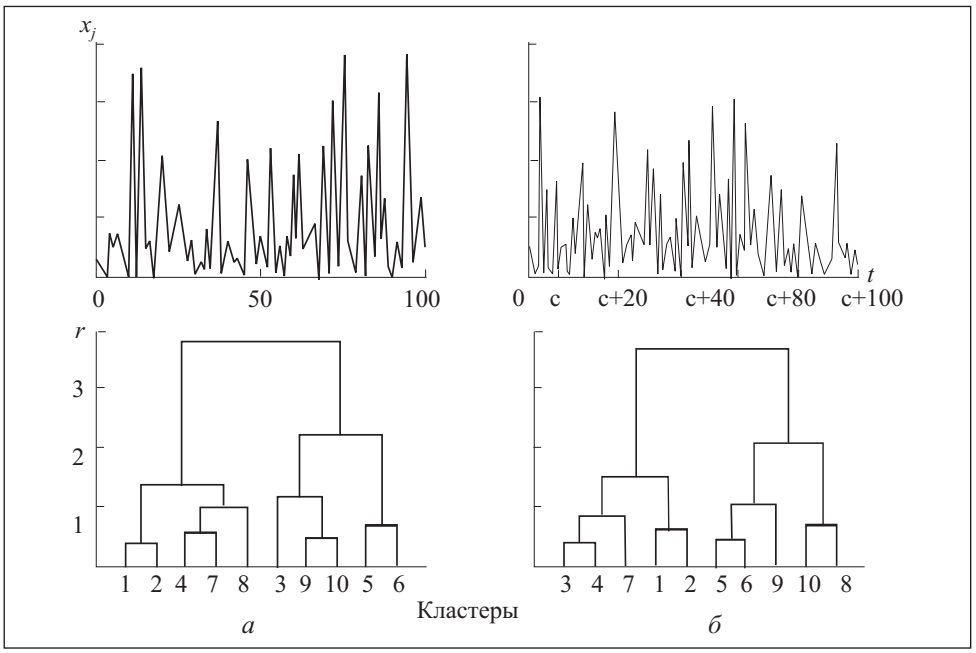


Рис. 4. Начальный нормальный трафик $X^1(t) = \{x^1(t)_{t=1}^{100}\}$ и его БДр (*а*) и альтернативный трафик и его БДр (*б*)

В качестве элементов этих множеств могут быть выбраны соответствующие элементы базы данных, однозначно интерпретируемые как нормальный трафик, или его начальные *f* значений, полученные сканированием КС и также интерпретируемые как нормальный трафик.

2. Выбираем число кластеров *k*, которые будут определять иерархическую структуру выбранного множества значений трафика. При этом желательно иметь техническую возможность реализовать кластеризацию типа один объект на один кластер, а БДр должно позволять удобную визуализацию (рис. 4).

Заметим, что в обоих случаях число кластеров в БДр одинаково, вследствие чего достигается идентичность сравниваемых объектов.

Структурированный трафик КС. Кластер представляет собой часть данных, которая отличается от остальных данных наличием или отсутствием некоторой однородности элементов. В простейшем случае это — похожесть элементов, в идеальном случае — совпадение значений основных переменных или иного рода близость, представляемая, в частности, геометрической близостью соответствующих объектов [12].

Согласно ИКА при наличии множества объектов $X = \{x_j\}_{j=1}^{n}$, характеризуемых признаками $\{x^{(i)}\}_{i=1}^{m}$, $x_j = \{x_j^{(i)}\}_{i=1}^{m}$, вычисляются попарные расстояния $d(x_j, x_f)$ в той или ной метрике и их матрица расстояний $D = (d(x_j, x_f))$. Объекты собираются в кластеры по критерию максимальной или минимальной удаленности друг от друга или по другому принципу [13].

Возможности КА для принятия решений относительно аномальности трафика определяются прежде всего тем, что он позволяет использовать многообразную, обозримую и формализованную систему правил. Поддержка принятия решений определяется возможностью решения таких задач:

структуризация данных в различных форматах и метриках;

анализ состава и основных компонент данных;

выявление групп объектов, к которым применимы одинаковые критерии;

выявление и анализ структуры взаимодействия основных подсистем (в данном случае компонент трафика или пакетов).

Одним из мощных средств анализа трафика, до сих пор практически не исследовванным, является визуализация структуры системы — основное средство поддержки и принятия решений, стимулирующее интуицию. При анализе трафика важное значение имеет интуитивное представление об объекте. После того как основные кластеры установлены, их взаимодействие и развитие значительно легче визуализировать, чем в исходной массе данных. На стадии идентификации превалируют рекурсивные методы восприятия, характеризуемые глубокой обратной связью глаз—мозг, что связано с использованием структурного представления изображения. Эта процедура визуализации эффективна при определении степени аномальности трафика.

Трафик КС имеет существенные различия в величинах отдельных компонент. При его анализе, а особенно при визуализации, эффективным является преобразование евклидовых расстояний в матрицы сходства с помощью так называемых кернел-функций [14]. Свойства этого преобразования позволяют объединить многие алгоритмы, до недавнего времени рассматривавшиеся изолированно (в трафике КС—компоненты, имеющие

различную природу: IP-адреса, номера портов, число передаваемых байт, качественные оценки, преобразование переменных, критерии КА и др.)

Важное значение для анализа трафика КС имеет правильный выбор суммарного числа кластеров, на который влияют число компонент трафика, учитывающее системы мониторинга, и число отсчетов отдельных компонент трафика. Во всех случаях желательно иметь возможность построения иерархической структуры по принципу один кластер на один объект. Заметим, что в существующих системах математического моделирования число кластеров ограничено (30—40).

Недостаток методов ИКА состоит в том, что один и тот же вид трафика может иметь различные структуры для различных методов кластеризации, тем не менее, они однозначно идентифицируют факт появления аномалии. Однако не исключены случаи неоднозначной интерпретации полученных моделей (структур). В настоящее время актуальна проблема кластерного консенсуса, так как ситуация, когда число доступных программ кластеризации, реализующих различные методы, формирующие на одних и тех же данных различные кластерные структуры, не может считаться нормальной в прикладном плане.

Не решив вопроса о выборе финальной структуры, наилучшим образом представляющей все возможные структуры, сложно говорить о возможности однозначного определения аномалий трафика. Кластерный консенсус — это агрегирование множества кластерных структур в единую структуру, наилучшим образом их представляющую.

При идентификации аномалий трафика в рассматриваемом случае ставится следующая задача. Даны два множества структур: $S^{(1)} = \{S_j^{(1)}\}, j = 1, n$, и $S^{(2)} = \{S_j^{(2)}\}, j = 1, m$. Каждая из структур $S_j^{(1)} \in S^{(1)}$ и $S_j^{(2)} \in S^{(2)}$ представляет аномальный и нормальный трафик. Идентификация трафика реализована в виде следующего правила:

если структура трафика $S^{(*)}$, полученная, например, методом наиболее удаленного соседа (НУС), близка к структуре $S_j^{(1)} \in S^{(1)}$, $j = 1, n$, полученной тем же методом, т.е. $abs[P(S_j^{(1)} - S^{(*)})] < \delta_S$ $(\exists j)$, где P — параметр, характеризующий структуру интегрального трафика; δ_S — некоторый порог, то трафик аномальный;

если структура трафика $S^{(*)}$, полученная методом НУС, близка к структуре $S_j^{(2)} \in S^{(2)}$, $j = 1, n$, полученной тем же методом, т.е. $abs[P(S_j^{(2)} - P(S^{(*)})] < \delta_S$ $(\exists j)$, где P — некоторый параметр, характеризующий структуру интегрального трафика, δ_S — некоторый порог, то трафик нормальный.

Кластерный консенсус при такой постановке задачи сводится к поиску гипотетической структуры S^0, для которой $abs(P(S_j^{(2)}) - P(S^{(0)})) \rightarrow$

$\to \min(\forall j)$ в случае нормального трафика и $abs(P(S_j^{(1)}) - P(S^{(0)})) \to$ $\to \min(\forall j)$ в случае аномального трафика. Аналогично формулируется задача и для случая идентификации аномального трафика по одной или нескольким его компонентам.

Идентификация аномального трафика существенно зависит от корректной интерпретации результатов КА. Следует заметить, что в настоящее время недостаточно внимания уделяется методам интерпретации. В работе [12] показано, что в общем случае возможны пять уровней интерпретации кластеров. Структурный анализ трафика должен в общем случае позволить определить признаки (компоненты трафика), вносящие наибольший вклад в изменение структуры.

Интерпретация кластеров на уровне p-адических моделей. Множество X с заданной в нем метрикой d называется метрическим пространством (МП). Одно и то же множество X может содержать различные структуры МП (X, d). Обычное расстояние между рациональными числами определяется по формуле $d(r_1, r_2) = |r_1 - r_2|$, однако его можно определить по так называемой p-адической норме [15].

Известно, что любое рациональное число r однозначно записывается в виде дроби $r = p^k(m/n)$, где $k \in Z$; m/n — несократимая дробь, числитель и знаменатель которой взаимно просты с величиной p; p^{-k} — p-адическая норма числа r, обозначаемая как $\|r\|_p$. В [15] показано, что расстояние, определяемое из выражения $|r_1 - r_2|_p$, обладает свойствами обычного расстояния, т.е. удовлетворяет, в частности, аксиоме треугольника $d_p(r_1, r_2) +$ $+ d_p(r_2, r_3) \ge d_p(r_1, r_3)$.

Однако есть принципиальные отличия вновь определенного расстояния от обычного, а именно наличие у p-адического расстояния так называемых неархимедовых свойств (ультраметрики). В частности, в отличие от обычного расстояния в p-адическом целые числа образуют ограниченное множество диаметра единица. Если к этому множеству применить процедуру пополнения, то получим компакт Q_p, элементы которого — целые p-адические числа. Они допускают запись в p-ичной системе счисления, т.е. целое p-адическое число c однозначно записывается в виде бесконечной влево последовательности $\cdots a_n \cdots a_2\ a_1\ a_0, 0 \le c \le p-1$. Эту последовательность рассматривают как сумму сходящегося ряда $a_0 + a_1 p + a_2 p^2 + ...$ $... + a_n p^n + ...$, сходимость которого вытекает из условия $\left\| a_n p^n \right\| \le p^{-n}$. Ультраметрическое расстояние соответствует степени делимости рационального числа на два. Чем лучше число делится на два, тем оно ближе к нулю: например, $8 = 2^3$ ближе к нулю, чем $1/2 = 2^{-1}$; $16 = 2^4$ ближе к нулю, чем 8; 480 ближе к нулю, чем 16; 384 ближе к нулю, чем 480 и т. д.

Целые p-адические числа можно складывать, вычитать и умножать, так как в общем случае они образуют кольцо, однако в отличие от Z в Q_p отсутствует естественный порядок, понятия положительного и отрицательного чисел в Q_p не имеют смысла, т.е. $-1 = \lim (p^n - 1)$ при $n \to \infty$. Важным свойством p-адических чисел является то, что многие из них в отличие от обычных целых чисел обратимы, т.е. если $sgn_p\, a \neq 0$, то a^{-1} — целое положительное число. Кроме того, все рациональные числа со знаменателем, взаимно простым с p, являются целыми p-адическими. Нетрудно видеть, насколько эффективным может быть применение p-адических чисел к анализу нечетких чисел, для которых вопросы обратимости не решены относительно идентификации аномального трафика в условиях неопределенности.

Арифметические действия с p-адическими числами выполняются, как с обычными десятичными дробями, только все операции начинаются с последней цифры. Наиболее полный набор процедур, связанных с p-адическим анализом, содержится в разделе Number пакета математического моделирования Mapple. Важное свойство p-адического анализа состоит в том, что естественным аналогом отрезка [0,1] является множество Q_p.

Определение. Норма называется неархимедовой, если для всех x и y выполнено неравенство $\| x + y \| \leq \max (\| x \|, \| y \|)$.

Расстояние, индуцированное неархимедовой нормой, называется ультраметрикой. Неравенство треугольника для обычной функции расстояния $d(x, z) \leq d(x, y) + d(y, z)$ трансформируется в сильное неравенство треугольника $d(x, z) \leq \max (d(x, y), d(y, z))$. Соответствующие МП называются ультраметрическими пространствами. Следует заметить, что функция $|\,.\,|_p$ может принимать только дискретное множество значений, а именно $\{p^n, n \in \mathbf{Z} \cup \{0\}\}$. Если $a, b \in \mathbf{N}$, то $a \equiv b \,(\mathrm{mod}\, p^n)$ тогда и только тогда, когда $|a - b|_p \leq 1/\, p^n$. Таким образом, a можно представить в виде сходящегося (по p-адической норме) ряда

$$a = \sum_{n=0}^{\infty} a_n p^n,$$

а множествоцелых p-адических чисел — в виде

$$\mathbf{Z}_p = \left\{ \sum_{i=0}^{\infty} a_i p^i \right\}, \; \mathbf{Z}_p = \{a \in \mathbf{Q}_p : |a|_p \leq 1\}.$$

Все p-адические числа имеют иерархическую струкутуру [16, 17] и, вместе с тем, все иерархические структуры описываются p-адическими

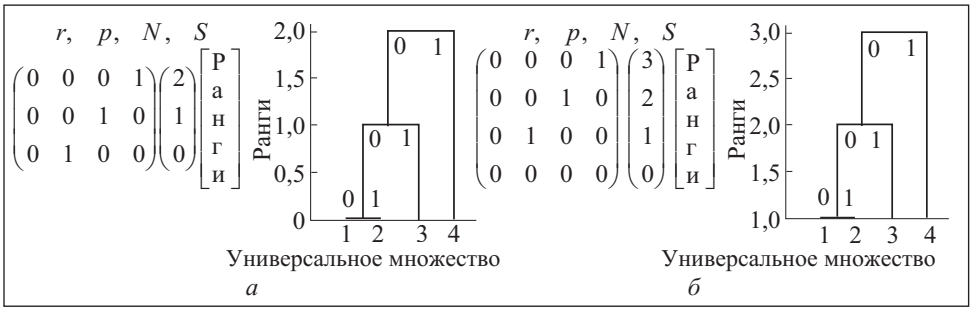


Рис. 5. Матричное и графическое (иерархическое дерево) представление стандартных (четких) чисел $7_{10} = 0 + 2^0 + 2^1 + 2^2$ (*а*) и $14_{10} = 0 + 2^1 + 2^2 + 2^3$ (*б*) в 2-адическом базисе

числами. На рис. 5 представлены числа 7_{10} и 14_{10} в 2-адическом базисе. В работах [16, 17] предложено каждое БДр характеризовать четырьмя параметрами: ранг *r*; *p*-адическое число,

$$p = \sum_{n=0}^{\infty} a_n p^n;$$

фрактальная размерность,

$$N = \lim_{\varepsilon \to 0+} \frac{\log N(\varepsilon)}{\log(1/\varepsilon)},$$

где $N(\varepsilon)$ — число кубов (квадратов), необходимых для покрытия всего множества кубами (квадратами) $\{B_i\}$ с величиной ребра (стороны), не превышающей ε; структуру *S*. Заметим, что фрактальная размерность *N* вычисляется для бинарного изображения матрицы БДр.

Для принятого числа сканирований трафика можно построить зависимости $r(j), p(j), N(s), S(j), j = 1, c$, и таким образом получить закономерности изменения данных параметров трафика для нормального и аномального состояний. Фрактальная размерность изображения БДр, соответствующих числам 7 и 14 в 2-адическом базисе, совпадает с топологической, в чем можно убедиться, вычислив эту величину с помощью программы fractdim ( ), входящей в состав стандартного пакета МатЛаб.

Идентификация аномальных состояний трафика. Для идентификации аномальных состояний трафика КС требуется выполнение необходимого числа сканирований, которое следует минимизировать, так как большое их число приводит к многомерным массивам сверхвысокой размерности, а малое — к искажению реальной структуры трафика (структу-

ра типа одно сканирование на один кластер не адекватна структуре типа 20 сканирований на один кластер и так далее). В работе [18] введено понятие масштабируемого пространства, которое позволяет в известной мере прояснить вопрос выбора рационального числа сканирований в концепции идентификации аномалий на основе структурных свойств компонент трафика или всего трафика в целом.

Теория масштабируемого пространства основана на том, что важные для исследователя свойства сигнала (понимаемого обобщенно) существуют только в определенном диапазоне масштабов. Если этот масштаб неизвестен, применение классических методов анализа с вычислением интегральных характеристик не даст желаемых результатов. Например, сканирование трафика КС, не совпадающее во временном масштабе со сканированием злоумышленника, не позволит идентифицировать его достаточно быстро.

При априорно неизвестном масштабе важных особенностей трафика КС наиболее эффективно мультимасштабное описание трафика: различное число кластеров, разнообразные метрики, альтернативные критерии построения иерархической структуры. Окончательной целью мультимасштабного описания является построение иерархии особенностей поведения трафика на основе анализа их взаимосвязи в различных масштабах, в частности на основе фрактальных характеристик его структур.

Существенное значение имеет визуальное представление структурных характеристик. В соответствии с концепцией масштабируемого пространства требуется выполнение следующих правил:

1. Максимально возможное число сканирований и определение элементарных объектов трафика (компонент, числа пакетов), от которых зависит тип иерархической структуры трафика, доступных для восприятия человеком (должны быть рассчитаны исходя из разрешающей способности соответствующего органа восприятия).

2. При огрублении масштаба восприятия, например при замене n сканирований одним усредненным или случайно взятым из рассматриваемого множества элементарные объекты сливаются, объединяясь в объекты более высокого уровня иерархии.

Второе правило справедливо для любых объектов. Бывают ситуации, когда число объектов при огрублении уровня восприятия монотонно убывает, приводя к появлению новых объектов (в данном случае кластеров), не являющихся объектами предыдущего уровня иерархии. Такие ситуации недопустимы и должны быть исключены, так как при этом нарушается аксиома причинности.

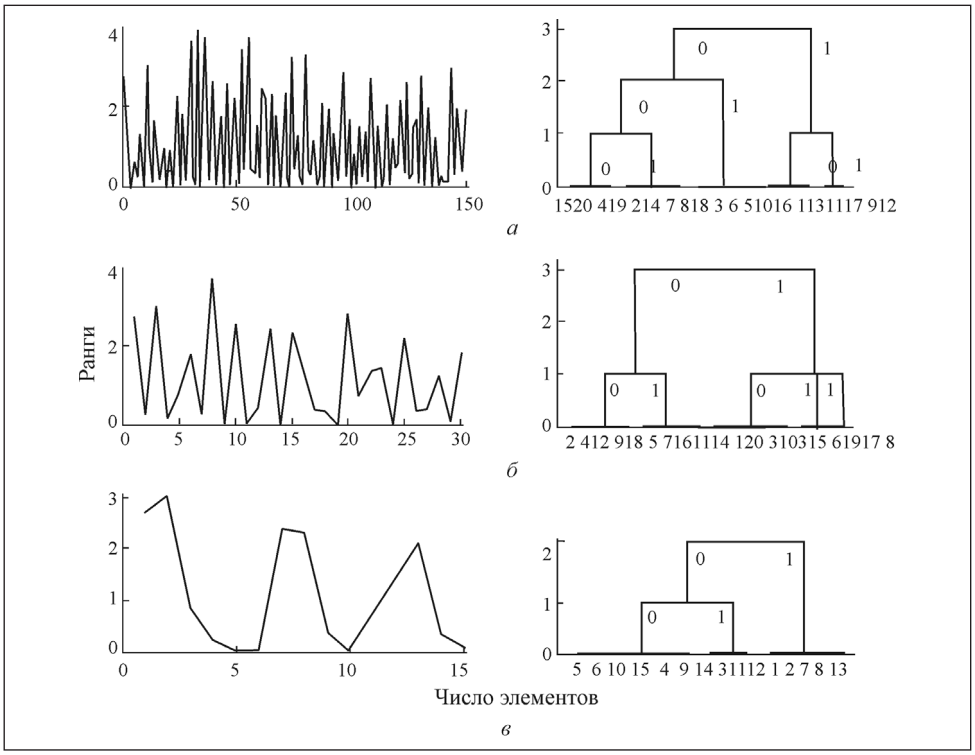


Рис. 6. Примеры построения упрощаемых версий структурного описания трафика, сканированного в интервале от 1 до 150 усл. ед. времени, последовательно представленного рядом из 150 (*а*), 30 (*б*) и 15 (*в*) значений: фрактальные числа и фрактальные размерности соответственно: *а* — 91 и 1,1123; *б* — 101 и 1,0745; *в* — 32 и 0,9745

Следует заметить, что теория масштабируемого пространства требует выполнения аксиомы изотропности, в соответствии с которой закон объединения объектов в иерархическую структуру не должен зависеть ни от пространственного направления, ни от масштаба уровня иерархии. При исследовании трафика смысл аксиомы изотропности несколько изменяется и свидетельствует о невозможности исследования различных частей трафика с помощью различных метрик.

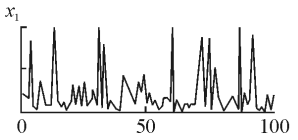
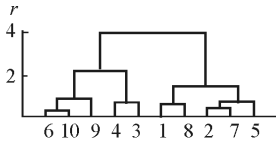
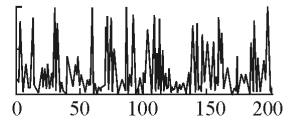
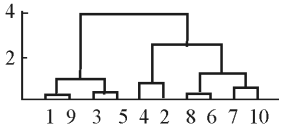
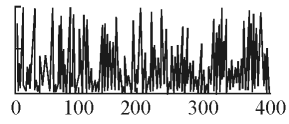
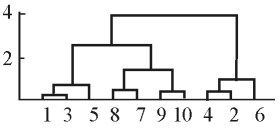
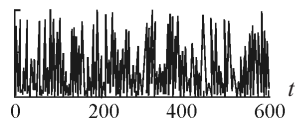
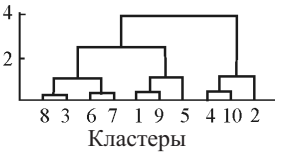
В общем случае операция разложения многомерного трафика в семейство упрощающихся версий структурного описания возможна до тех пор, пока не начнется существенное уменьшение (> 20 %) размерности *p*-адической матрицы.

На рис. 6 приведены временные ряды (ВР) и их БДр при иерархической кластеризации по методу НУС. Как видно из рис. 6, представление ВР из 150 элементов 30-ю элементами с точки зрения структурной бли-

зости вполне допустимо, так как фрактальные числа и фрактальные размерности отличаются не более чем на 10 %, что для приближенных расчетов вполне допустимо. В то же время, представление ВР из 150 элементов 15-ю элементами с точки зрения структурной близости недопустимо, так как фрактальные числа и фрактальные размерности отличаются более чем на 10 %, что недопустимо даже для приближенных расчетов.

Адекватная формализация динамических кластеров рассматривается с помощью построения усложняющихся версий структурного описания посредством наращивания числа элементов в ВР (табл. 1). Анализируя данные, представленные в табл. 1, видим, что гипотеза эквивалентности структур — $S^{(1)} \approx S^{(2)} \approx S^{(3)} \approx S^{(4)}$ — при заданном числе кластеров (в данном случае их 10) выполняется, что свидетельствует о корректности моделирования.

Таблица 1. **Версии структурного описания**

Нормальный трафик	Иерархическая структура	Характеристика трафика
		Базовый трафик $x_1(t) = x(t)_{t=0}^{t_1}$ и соответствующая структура $S^{(1)}(k)$ (число кластеров задано)
		Динамически измененный трафик $x_2(t) = x_1(t) \cup x(t)_{t=t_1}^{t_2}$ и соответствующая структура с заданным ранее числом кластеров $S^{(2)}(k)$ (гипотеза $S^{(2)} = S^{(1)} \cup \delta_{S^{(1)}}$)
		Динамически измененный трафик $x_3(t) = x_2(t) \cup x(t)_{t=t_2}^{t_3}$ и соответствующая структура с заданным ранее числом кластеров $S^{(3)}(k)$ (гипотеза $S^{(3)} = S^{(2)} \cup \delta_{S^{(2)}}$)
		Динамически измененный трафик $x_4(t) = x_3(t) \cup x(t)_{t=t_3}^{t_4}$ и соответствующая структура с заданным ранее числом кластеров $S^{(4)}(k)$ (гипотеза $S^{(4)} = S^{(3)} \cup \delta_{S^{(3)}}$)

Рассмотрим пример идентификации аномального состояния трафика на основе структурних характеристик. Известны потоки данных, относимые экспертами к классу нормальных и аномальных. Матрицы данных имеют размерность 500 × 9, где 500 — число пакетов, 9 — число компонент трафика. Фрагмент потока данных, состоящий из пяти пакетов, приведен в табл. 2. Компоненты трафика: $X = \{X_i\}_{i=1}^{9}$, X_1 — Protocol ID — протокол, связанный с событием (TCP = 0, UDP = 1, ICMP = 2, unknown = 3); X_2 — номер порта источника; X_3 — номер порта хоста назначения; X_4 — IP-адрес источника; X_5 — IP-адрес приемника; X_6 — ICMP Type — тип ICMP-пакета (Echo Request or Null); X_7 — ICMP-Code — кодовое поле из ICMP-пакета (None or Null); X_8 — Raw Data Length — длина данных в

Таблица 2. **Фрагмент потока данных**

$X = \{X_i\}_{i=1}^{9}$	Трафик, состоящий из пакетов				
	1	2	3	4	5
			Аномальный		
X_1	0	0	0	0	0
X_2	1788.00	1731.00	1668.00	1739.00	1791.00
X_3	6419.00	6478.00	6884.00	6219.00	6619.00
X_4	853890539.00	880594725.00	858785769.00	921460420.00	909228438.00
X_5	–930935380.00	–1039198847.00	–971792093.00	–1020585109.00	–1015310165.00
X_6	1.00	1.00	1.00	1.00	1.00
X_7	1.00	1.00	1.00	1.00	1.00
X_8	457.01	116.26	344.08	201.68	480.41
X_9	872.00	2912.00	2659.00	3012.00	2960.00
			Нормальный		
X_1	0	0	0	0	0
X_2	2396.00	3010.00	3104.00	2359.00	3255.00
X_3	102.00	88.00	82.00	112.00	108.00
X_4	2338390598.00	1595431598.00	1760174623.00	2139448656.00	2275207188.00
X_5	–1686689687.00	–1882232220.00	–1639528692.00	–1933497671.00	–1599454967.00
X_6	1.00	1.00	1.00	1.00	1.00
X_7	1.00	1.00	1.00	1.00	1.00
X_8	481.00	503.00	567.00	415.00	539.00
X_9	4399.00	3887.00	4295.00	5144.00	5349.00

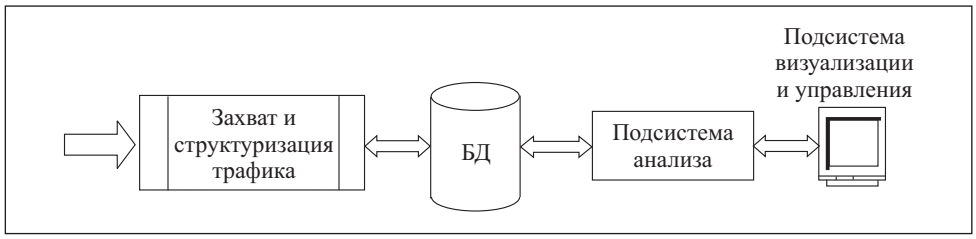


Рис. 7. Схема расчетной модели имитации и анализа трафика КС

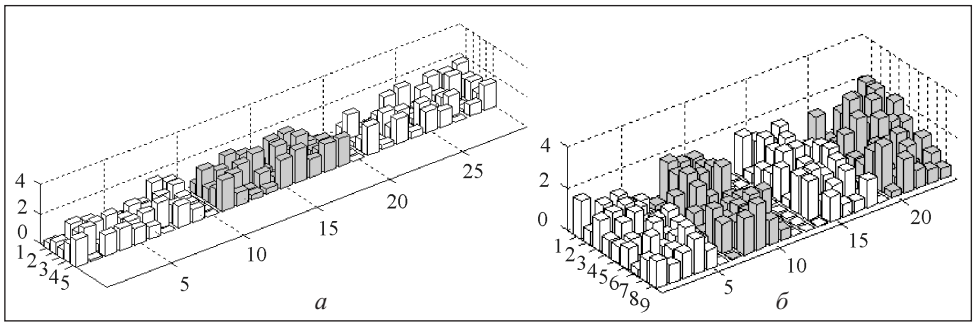


Рис. 8. Расчетная модель многомерного трафика КС: *а* — поток данных структурирован попакетно; *б* — поток данных структурирован покомпонентно (иерархические структуры построены для строк и столбцов матрицы данных)

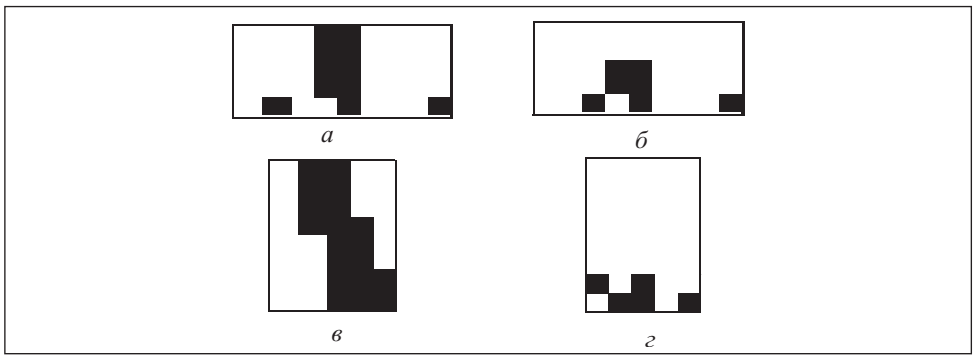


Рис. 9. Визуализация аномального (*а*, *в*) и нормального (*б*, *г*) трафиков при различных способах структуризации матрицы данных: *а*, *б* — покомпонентная иерархическая кластеризация; *в*, *г* — попакетная иерархическая кластеризация; фрактальные числа и фрактальные размерности соответственно: *а* — 8230 и 0,0011; *б* —15,00 и 0,8856; *в* —15,00 и 0,8856; *г* —7,0 и 0,7740

пакете; X_9 — Raw Data — порция данных в пакете. Расчетная модель имитации и анализа трафика КС приведена на рис. 7, а покомпонетная и попакетная структуризация потока данных — в табл. 2.

Рассмотрим свойства структурированного трафика с целью определить возможность идентификации аномального состояния на основании структурных характеристик. При этом необходимо выполнить построение БДр для двух случаев структуризации потока данных. Затем следует вычислить фрактальные числа и фрактальные размерности полученных иерархических структур для сравнения и идентификации аномальных состояний и показать возможность визуальной идентификации.

Визуальное изображение структурной модели трафика, полученное на основании кодирования дендрогрограмм, фрактальные числа и фрактальные размерности для рассмотренных случаев приведены на рис. 8, где темным цветом выделены пакеты потока данных, в которых компоненты трафика находятся вне интервала нормальных значений: $(\forall x_j)\, x_j \notin [x_j^{\min} x_j^{\max}]$, $j = 1,5$. После кодирования БДр и построения бинарной матрицы выполнено затемнение клеток матрицы, в которых находится единица. Таким образом, получено визуальное изображение структурной модели трафика, а также фрактальные числа и фрактальные размерности для рассмотренных случав структуризации (рис. 9). Из рис. 9 видно, насколько визуально различны изображения, идентифицируемые как аномальное и нормальное состояния. Такой поход позволяет достаточно просто идентифицировать состояние КС, а полученные числовые оценки усиливают эффект визуализации.

Выводы

1. Попытки распознавания современных схем организации атак и других противозаконных действий с помощью известных методов приводят к тому, что исходная задача становится практически неразрешимой. Современное состояние исследований в этом направлении можно охарактеризовать как особое внимание к многомерному трафику КС и его структуре, так как только учет всех компонент и их взаимосвязей позволяет идентифицировать аномалию.

2. Предложенный новый метод идентификации аномальных состояний трафика основан на анализе его структурных свойств — различие составляет приблизительно два порядка.

Проведенные практические эксперименты с трафиком реальной сети и моделями реального трафика свидетельствуют о высокой эффективности предложенной методики.

The paper deals with the problem of identification of anomalous conditions of computer system traffic based on its presentations in the manner of multivariate time series, which componentwise and packagewise structuring enables to get hierarchical structured features of dataflow, used herein after for identifications. A possibility of application of *p*-adical models for the analysis of traffic, when using the preceding dataflow as dynamic pattern conditions, is shown.The examples which evidence for efficiency of the offered methodology are presented.

СПИСОК ЛИТЕРАТУРЫ

1. *Jiang D., Qin W., Nie L. et al.* Time-frequency detection algorithm of network traffic anomalies. Intern. Conf. on innovation and information management (ICIIM IPCSIT). — 2012. — Vol. 36. — IACSIT Press, Singapore. — P. 110—116.
2. *Витяев Е.Е. , Ковалерчук Б.Я., Федотов А.М. и др.* Обнаружение закономерностей и распознавание аномальных ситуаций в потоке данных сетевого трафика//Вестник НГУ. Серия: Информационные технологии. — 2008. — **6**, вып. 2. — С. 57—70.
3. *Предотвращение* атак с распределенным отказом в обслуживании (DDoS). Технический отчет: Угрозы DDoS — риски, устранение и лучшие практические приемы. — Интернетресурс: http://www.cisco.com/web/U/products/ps5887/products_white_paper0900aecd8011e927_ .html
4. *Возможности* классификации и идентификации трафика,заложенные в программное обеспечение Cisco IOS. — Интернет-ресурс: http://www.cisco.com/web/about/ac123/%20ac114/ac173/Q3-04/dept_ttips_threat.html
5. *Kim M.-S. , Kang H.-J., Hong S.-Ch. et al.* A flow-based method for abnormal network traffic detection. — Интернет-ресурс: attack-analysis-v5-revision.pdf
6. *De Almeida A.L.F., Favier G., Mota J.C.M.* Tensor-decompositions and applications to wireless communication systems//Telecommunications: advances and trend in transmission, networking and applications. Edited by Charles Casimiro Cavalcante, Ricardo FialhoColares e Paulo Cesar Barbosa. — Fortalesa: Universidade de Fortalesa. - UNI-FOR, 2006. — 187 p.
7. *Sidiropoulos N.D., Kyrillidis A.* Multi-way compressed sensing for sparse low-rank tensors// IEEE signal processing letters.— 2012. — Vol. 19, № 11. — P. 757—760.
8. *Sun J., Tao D., Faloutsos Ch.* Beyond Streams and Graphs: Dynamic Tensor Analysis. — Интернет-ресурс: http://pdf.aminer.org/000/473/322/beyond_ streams_and_ graphs_dynamic_ tensor_analysis.pdf.
9. *Gudkov V., Johnson E.* Multidimensional Network Monitoring for Intrusion Detection.— Интернет-ресурс: http://www.necsi.edu/events/iccs/2002/NAp03_gudkov_iccsFixed02.20pdf.
10. *Northcutt S., Novak J., McLachlan D.* Network intrusion detection. An analyst's handbook. Indiapolis: New Riders Publishing, IN. — 2001. — P.
11. *Минаев Ю.Н, Толстикова Е.В., Филимонова О.Ю., Минаева Ю.И.* Интеллектуальные методы идентификации аномального трафика на основе *p*-адических моделей// Тези доп. V Міжнародної наук.-техн. конф. «Комп′ютерні системи та мережні технології» (CSNT-2012), Київ, 13—15 червня 2012 р. — Киев: изд. НАУ, 2012. — С. 18—20.
12. *Миркин Б.Г.* Методы кластер-анализа для поддержки принятия решений: обзор. Препринт WP7/ 2011/03. — М. : Изд. дом Нац. исслед. ун-та «Высшая школа экономики», 2011. — 88 с.
13. *Документация* МатЛаб: matlab: indexhelper('C:/MatLab7/Toolbox/Stats', '' , 'statistics', 'Cluster Analysis','html/clusterdemo.html)
14. *Kolá̌cek J., Zelinka J.* Kernel Smoothing Toolbox for MATLAB. — Интернет-ресурс: http://www.math.muni.cz/english/science-andresearch/developed-software/232-matlabtoolbox.html.

15. *Каток С.Б. P*-адический анализ в сравнении с вещественным./Пер.с англ. П.А. Колгушкина. — М. : МЦНМО, 2004. —112 с.
16. *Владимиров В.С., Волович И.В., Зеленов Е.И. P*-адический анализ и математическая физика. — М. : Физматлит, 1994. — 352 с.
17. *Хренников А.Ю.* Неархимедов анализ и его приложения. — М. : Физматлит, 2003. — 216 с.
18. *Скурихин А.В.* Применение методов масштабируемого пространства в обработке сигналов. —Интернет-ресурс: http://www.spiiras.nw.ru/rus/conferences/ict/Skurihin110604.ppt.
19. *MySQL*:6.3.3.2. Математические функции. — Интернет-ресурс: http://phpclub.ru/mysql/doc/mathematical-functions.html.

Поступила 20.03.13;
после доработки 21.05.13

МИНАЕВ Юрий Николаевич, д-р техн. наук, профессор кафедры компьютерных систем и сетей Национального авиационного университета. В 1959 г. окончил Харьковский политехнический ин-т. Область научных исследований — интеллектуальный анализ данных, применение интеллектуальных технологий в системах принятия решений.

ФИЛИМОНОВА Оксана Юрьевна, канд. техн. наук, доцент Киевского национального университета строительства и архитектуры. В 1989 г. окончила Киевский инженерно-строительный ин-т. Область научных исследований — интеллектуальный анализ данных.

МИНАЕВА Юлия Ивановна, канд. техн. наук, и. о. доцента кафедры основ информатики Киевского национального университета строительства и архитектуры, который окончила в 2008 г. Область научных исследований — интеллектуальный анализ данных.