А. К. Горшенин, С. А. Горбунов, Д. Ю. Волканов, О кластеризации объектов сетевой вычислительной инфраструктуры на основе анализа статистических аномалий в трафике, *Информ. и её примен.*, 2023, том 17, выпуск 3, 76–87

DOI: 10.14357/19922264230311

О КЛАСТЕРИЗАЦИИ ОБЪЕКТОВ СЕТЕВОЙ ВЫЧИСЛИТЕЛЬНОЙ ИНФРАСТРУКТУРЫ НА ОСНОВЕ АНАЛИЗА СТАТИСТИЧЕСКИХ АНОМАЛИЙ В ТРАФИКЕ*

А. К. Горшенин[1], С. А. Горбунов[2], Д. Ю. Волканов[3]

Аннотация: Рассматривается задача выявления статистических аномалий (т. е. существенных превышений от типичных значений полученного и исходящего трафика) нагрузки на узлы сетевой вычислительной инфраструктуры. Рост нагрузки в реальных системах ведет к необходимости регулярного масштабирования вычислительных ресурсов и хранилищ, а также перенаправления потоков данных. Предложена процедура выявления статистических аномалий в сетевом трафике с использованием аппроксимации наблюдений обобщенным гамма-распределением для дальнейшей кластеризации объектов сетевой вычислительной инфраструктуры с целью оценки потребности в ресурсах. Все вычислительные статистические процедуры, описанные в статье, реализованы с использованием языка программирования R и применены к сетевому трафику, полученному в рамках моделирования на специализированном архитектурно-программном стенде. Предложенные подходы могут быть использованы и для более широкого класса телекоммуникационных задач.

Ключевые слова: сетевая инфраструктура; сетевой трафик; обобщенное гамма-распределение; вычислительная статистика; проверка статистических гипотез; выявление аномалий; кластеризация

DOI: 10.14357/19922264230311 **EDN:** XHTMVI

1 Введение

В современной сетевой вычислительной инфраструктуре по мере развития информационных ресурсов растет нагрузка на вычислительные ресурсы инфраструктуры [1, 2]. Этот рост периодически вызывает необходимость масштабирования вычислительных ресурсов и ресурсов хранилищ данных на узлах сетевой вычислительной инфраструктуры или же перенаправления потоков данных. На сегодняшний день эта проблема чаще всего решается в «ручном» режиме на основе опыта сетевых и системных администраторов. Для автоматизации процессов управления масштабированием ресурсов на вычислительных узлах сетевой вычислительной инфраструктуры необходимы методы объективного выявления аномалий в нагрузке на такие узлы и оценки размера аномалий для определения объемов необходимых ресурсов [3–6]. Под аномалиями в данной работе понимаются существенные превышения типичных значений полученного и исходящего трафика. Они могут быть вызваны, например, работой несанкционированного программного обеспечения на узле или нехваткой ресурсов для обработки пользовательского трафика. Один из возможных путей решения этой проблемы заключается в мониторинге и анализе сетевого трафика в сетевой вычислительной инфраструктуре и выявления в нем статистических аномалий, возникающих в том числе в результате влияния набора случайных факторов [7, 8].

Хорошо известна возможность статистического описания процессов в телекоммуникационных сетях с использованием различных семейств гамма-распределений: классических гамма — для распределений времени обслуживания [9]; гамма–гамма — для аппроксимации некоторых характеристик в сетях сотовой связи [10], подводных коммуникационных системах [11], глобальных вычислительных сетях [12]; конечных гамма-смесей — для описания тонкой структуры информационных потоков [13–17]; обобщенных гамма (Generalized Gamma, GG) — для распределений времени пребывания пользователя в ячейке сотовой сети [18].

*Работа выполнена при поддержке Программы развития МГУ, проект № 23-Ш03-03. При обработке и анализе трафика использовалась инфраструктура Центра коллективного пользования «Высокопроизводительные вычисления и большие данные» (ЦКП «Информатика») ФИЦ ИУ РАН (г. Москва).

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук; Московский государственный университет имени М. В. Ломоносова, agorshenin@frccsc.ru

[2]Московский государственный университет имени М. В. Ломоносова; Московский центр фундаментальной и прикладной математики, s.gorbunov.cmc@gmail.com

[3]Московский государственный университет имени М. В. Ломоносова, volkanov@asvk.cs.msu.ru

В работе предложен статистический подход к решению задачи кластеризации объектов сетевой вычислительной инфраструктуры по уровню нагруженности с точки зрения информационного обмена (передаваемого и получаемого трафика) на основе процедуры выявления аномальных наблюдений в сетевом трафике с использованием специального GG-теста. Продемонстрированы примеры применения разработанной методики для данных сетевого трафика, полученных в рамках моделирования на специализированном архитектурно-программном стенде кафедры автоматизации систем вычислительных комплексов факультета вычислительной математики и кибернетики МГУ имени М. В. Ломоносова.

2 Статистическая модель трафика на основе обобщенного гамма-распределения

В качестве анализируемых данных в статье используются наборы, полученные на специализированном стенде, на котором моделируются различные сценарии реальных сетевых взаимодействий. Рассматриваются суточные данные (с пятиминутной агрегацией). В каждый момент времени известно суммарное число отправленных и полученных бит (для удобства дальнейшего анализа данные нормированы значением 2^{20} — около 131 Кбайт). За указанный промежуток времени моделировалось взаимодействие для 1 920 объектов в сети. На рис. 1 приведен пример данных входящего трафика для одного из объектов за все время наблюдений, а на рис. 2 — данные сразу для всех объектов, но в некоторый фиксированный момент времени.

Ранее авторами было установлено высокое статистическое согласие данных мобильного трафика сотового оператора с семейством обобщенных гамма-распределений [19] — это сосредоточенное на положительной полупрямой трехпараметрическое семейство вероятностных распределений, определяемое плотностью вида

$$f(x; r, \gamma, \mu) = \frac{|\gamma|\mu^r}{\Gamma(r)}\, x^{\gamma r - 1} e^{-\mu x^\gamma},$$

где $x > 0$, $r > 0$, $\mu > 0$, $\gamma \in \mathbb{R} \setminus \{0\}$. Его выбор может объясняться и тем фактом, что указанное семейство содержит практически все самые популярные абсолютно непрерывные распределения, сосредоточенные на положительной полупрямой, в том числе распределения с тяжелыми хвостами (некоторые примеры приведены в таблице). Естественным образом возникает идея использования

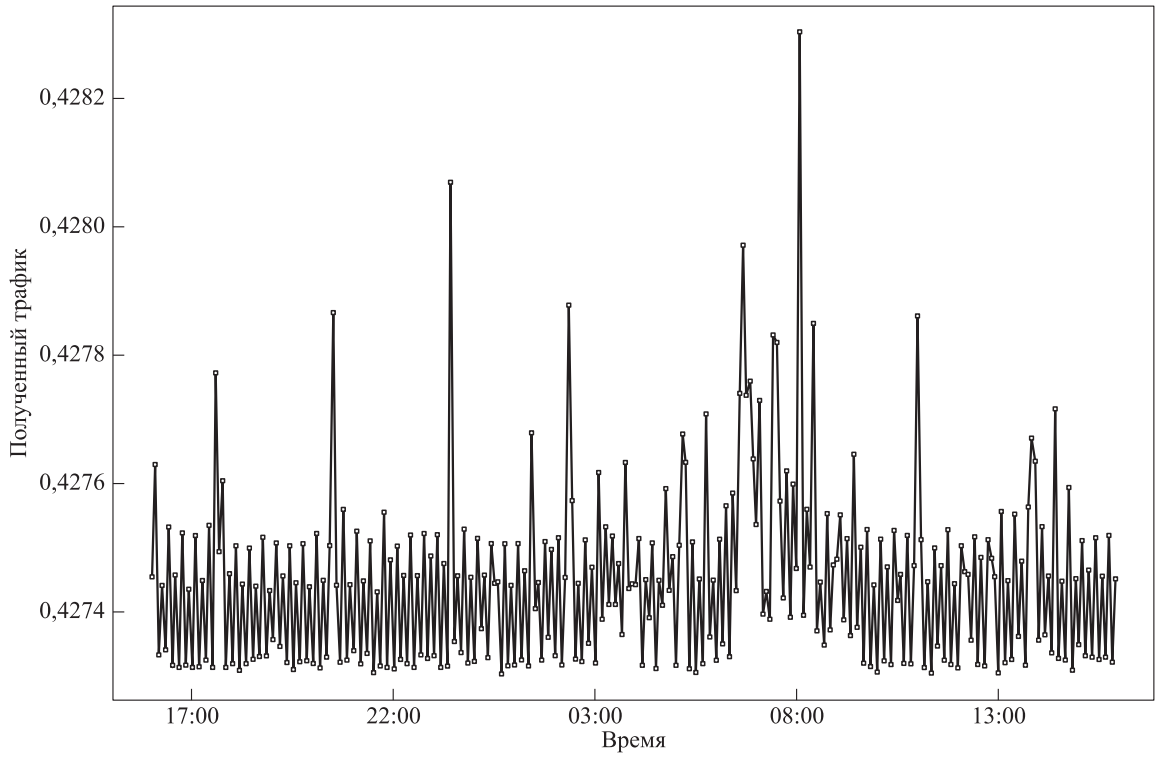

Рис. 1 Временной ряд полученного трафика для одного объекта

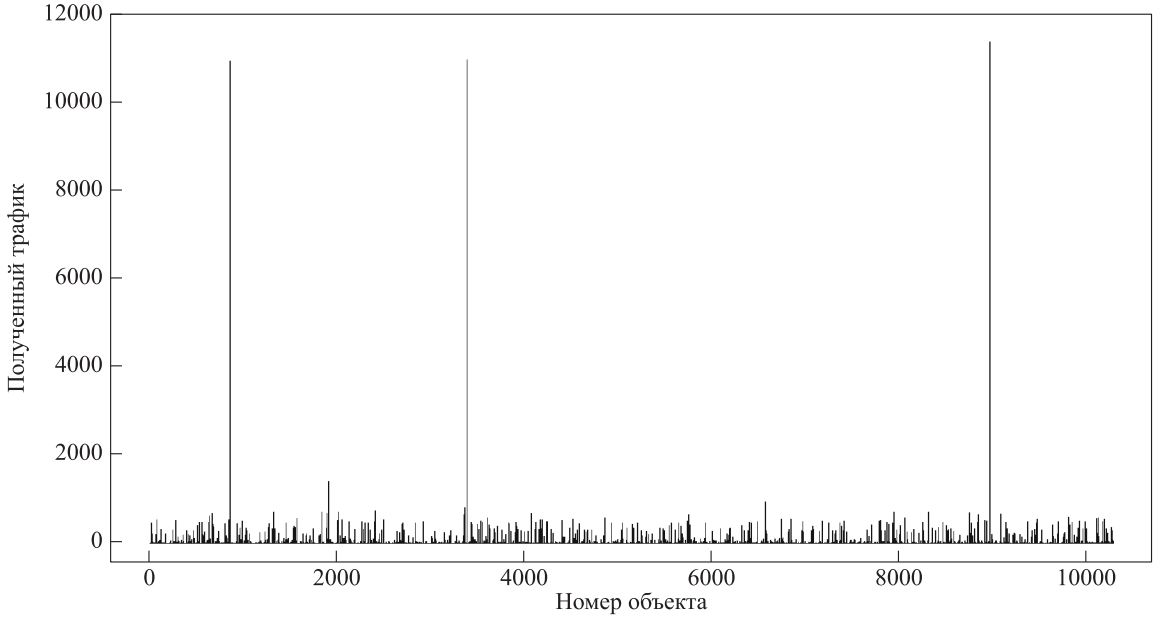

Рис. 2 Выборка полученного трафика за одну отметку времени для всех объектов

подобной статистической модели для иных информационных систем.

Для оценивания параметров GG-распределения можно применять различные методы, начиная от классического метода максимального правдоподобия. Однако в этом случае для анализируемых данных наблюдались существенные ошибки при аппроксимации хвостов распределения. Поэтому в качестве вычислительной процедуры был реализован алгоритм на основе минимизации l^2-нормы между значениями эмпирической и теоретической функций распределения, вычисленными в узлах некоторой сетки $y = (y_1, \ldots, y_m)$. Для оценки параметров по выборке $\mathbb{X} = (X_1, \ldots, X_n)$ решается оптимизационная задача следующего вида:

$$(\hat{r}, \hat{\gamma}, \hat{\mu}) = \underset{r>0,\mu>0,\gamma\in\mathbb{R}\setminus\{0\}}{\arg\min} \sum_{i=1}^{m} \left[\int_0^{y_i} \frac{|\gamma|\mu^r}{\Gamma(r)}\, x^{\gamma r-1} e^{-\mu x^\gamma}\, dx - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i < y_i) \right]^2.$$

Типичный пример функций, полученных в результате применения различных методов, продемонстрирован на рис. 3. Оптимизационный подход позволил не только точнее, но и примерно в 30 раз быстрее оценивать параметры для всех анализируемых данных по сравнению с методом максимального правдоподобия.

Некоторые частные случаи обобщенного гамма-распределения

Семейство	**Значения параметров**
Гамма-распределение	$\gamma = 1$
Обратное гамма-распределение	$\gamma = -1$
Распределение Леви	$\gamma = -1,\ r = 0{,}5$
Показательное распределение	$\gamma = 1,\ r = 1$
Распределение Эрланга	$\gamma = 1,\ r \in \mathbb{N}$
Распределение хи-квадрат	$\gamma = 1,\ \mu = 0{,}5$
Распределение Накагами	$\gamma = 2$
Полунормальное распределение	$\gamma = 2,\ r = 0{,}5$
Распределение Рэлея	$\gamma = 2,\ r = 1$
Хи-распределение	$\gamma = 2,\ \mu = 1/\sqrt{2}$
Распределение Максвелла	$\gamma = 2,\ r = 1{,}5$
Распределение Фреше (распределение экстремальных значений II типа)	$r = 1,\ \gamma < 0$
Распределение Вейбулла–Гнеденко (распределение экстремальных значений III типа)	$r = 1,\ \gamma > 0$

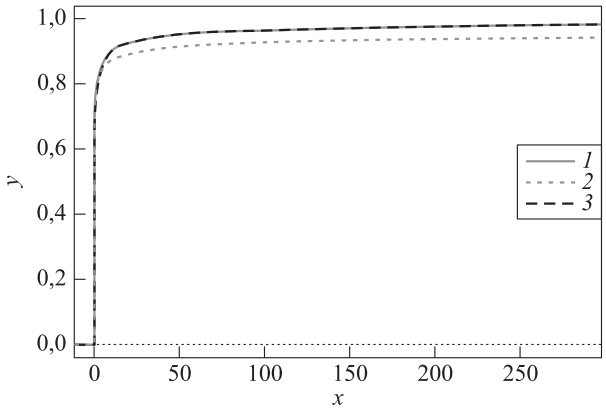

Рис. 3 Сравнение эмпирической функции распределения (*1*) и функций, полученных двумя методами оценки параметров — максимального правдоподобия (*2*) и функциональной оптимизации (*3*)

3 Метод кластеризации объектов сетевой вычислительной инфраструктуры на основе анализа аномалий трафика

В реальных задачах аномальность трафика определяется не только его абсолютным значением, но и соотношением с другими наблюдениями. В данном разделе рассмотрим процедуру, которая позволяет статистически корректно учитывать это.

Сначала рассмотрим статистическую процедуру выявления аномальных наблюдений для данных, описываемых обобщенным гамма-распределением. Впервые такого рода статистический тест был предложен в статье [20] для метеорологических рядов (осадков). Учитывая особенности анализируемых данных (параметр γ для трафика может иметь как положительные, так и отрицательные значения), обобщим описанную в упомянутой статье процедуру.

Пусть $V_1, \ldots, V_m$ — независимая выборка из обобщенного гамма-распределения с некоторыми параметрами $r > 0$, $\gamma \neq 0$, $\mu > 0$, $V_1 \geqslant V_j$, $\forall j \geqslant 2$. Рассмотрим статистику вида

$$\hat{\mathcal{R}} = \left(\frac{(m-1)V_1^{\gamma}}{V_2^{\gamma} + \cdots + V_m^{\gamma}} \right)^{\operatorname{sgn}(\gamma)}.$$

Тогда при условии, что верна гипотеза H_0: «значение V_1 не является аномально большим»,

$$\hat{\mathcal{R}} \sim \begin{cases} F(r,\ (m-1)r) & \text{для } \gamma > 0; \\ F((m-1)r,\ r) & \text{для } \gamma < 0, \end{cases}$$

где через F обозначено распределение Снедекора–Фишера (F-распределение) с соответствующими параметрами.

Для того чтобы выявить аномальные наблюдения в выборке из обобщенного гамма-распределения $\mathbb{V} = \{V_1, \ldots, V_m\}$ с параметрами (r, γ, μ), необходимо зафиксировать уровень значимости α, для каждого V_i, $i \in \{1, \ldots, m\}$, рассчитать значение статистики $\hat{\mathcal{R}}$ и сравнить это значение с α-квантилями F-распределения с соответствующими параметрами. Наблюдение признается аномальным тогда и только тогда, когда наблюдаемое значение статистики больше квантиля F-распределения.

На практике для выборок большого объема подсчет знаменателя статистики $\hat{\mathcal{R}}$ для каждого наблюдения может заметно увеличить время работы алгоритма, поэтому рациональнее будет один раз посчитать $S = V_1^{\gamma} + \cdots + V_m^{\gamma}$, а затем для i-го наблюдения вычислять значение статистики следующим образом:

$$\hat{\mathcal{R}} = \left(\frac{(m-1)V_i^{\gamma}}{S - V_i^{\gamma}} \right)^{\operatorname{sgn}(\gamma)}.$$

На рис. 4 приведены примеры выявления аномальных наблюдений при уровне значимости $\alpha = 0{,}05$ для фиксированного момента времени (выделены маркерами) с помощью описанного метода.

Для кластеризации можно использовать такую разметку наблюдений для выборок за каждый момент времени. Причем формировать выборки можно как на непересекающихся последовательных интервалах времени, так и воспользовавшись методом скользящего окна.

Сначала разметим аномальные наблюдения описанным выше GG-тестом, формируя выборки непересекающимися часовыми окнами, и нанесем получившуюся разметку на временны́е ряды по конкретным объектам. Если рассматривать разметку GG-теста на выборках, по которым он обучался (рис. 5), то можно заметить, что существует некоторое пороговое значение, разделяющее аномальные и обычные значения.

Однако при рассмотрении той же разметки для конкретного объекта (рис. 6) такого порога может и не оказаться. Так происходит из-за того, что выборки, на которых размечаются аномалии, формируются окнами по времени, поэтому разметка содержит в себе информацию о среднем тренде по всем объектам (рис. 7). На примере объекта с рис. 6 можно выделить две зоны с большим числом аномалий: в районе 22:00, где трафик рос быстрее среднего по всем объектам, и в районе 14:00, где трафик конкретного объекта достиг максимума, а перцентили имели нисходящий тренд. Имея разметку по каждому объекту, можно провести бинарную кластеризацию, задавая некоторый порог доли данных, которые были размечены аномальными.

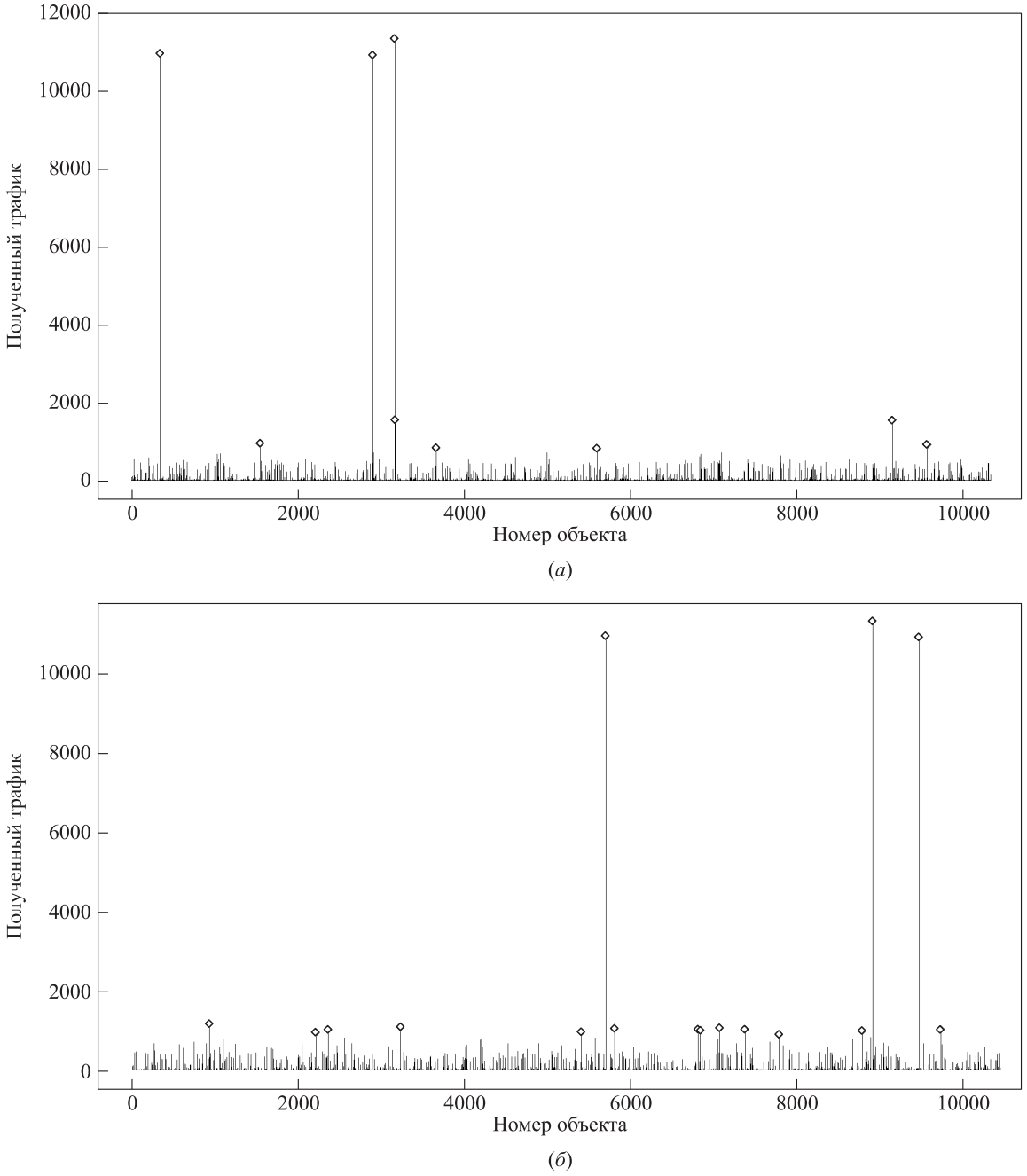

Рис. 4 Аномалии в полученном (*а*) и отправленном (*б*) трафике

Для визуализации полученных результатов составим признаковое описание объектов: каждый объект опишем медианой, квартилями и интерквартильным размахом по полученному и отправленному трафику по всем имеющимся наблюдениям.

Применим метод главных компонент, чтобы перевести признаковое описание на двумерную плоскость. Установлено, что первые две главные компоненты суммарно описывают 95,8% вариации, причем одна приписывает больший вес перцентилям (см. ось абсцисс на рис. 8), а другая – интерквартильному размаху (см. ось ординат на рис. 8). На рис. 8 отмечены объекты, признанные аномальными по полученному (*1*), отправленному (*2*) и обоим видам трафика (*3*).

По оси абсцисс на рис. 8 отложены значения первой главной компоненты, которая имеет выраженную положительную корреляцию с перцентилями, по оси ординат — значения второй главной компоненты с выраженной отрицательной корреляцией с размахами. При фиксированном уров-

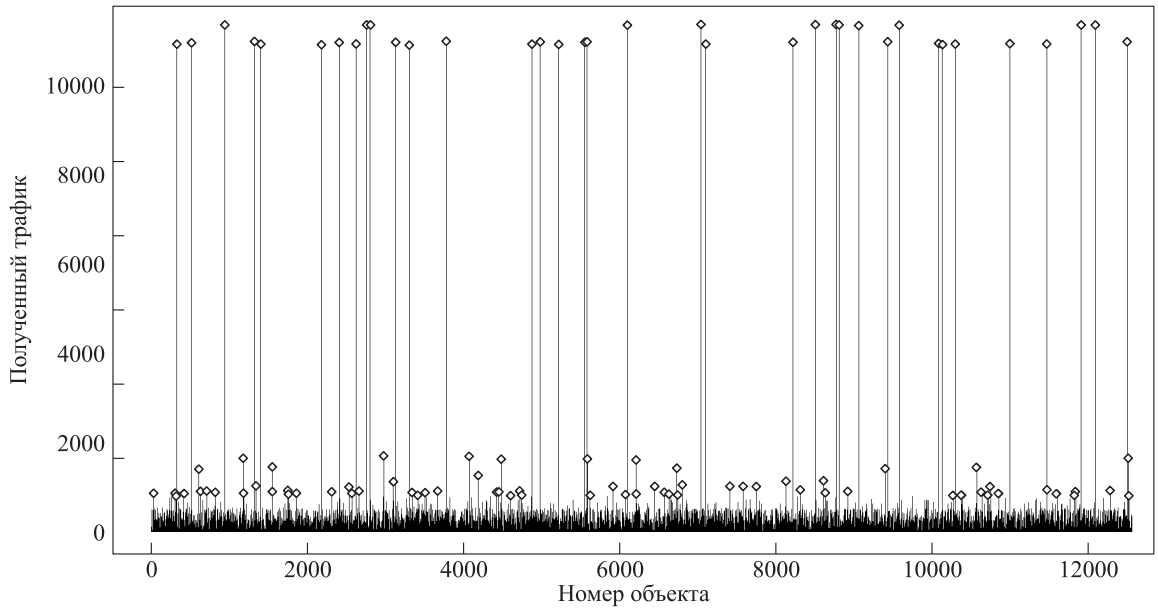

Рис. 5 Аномалии на часовом горизонте наблюдения

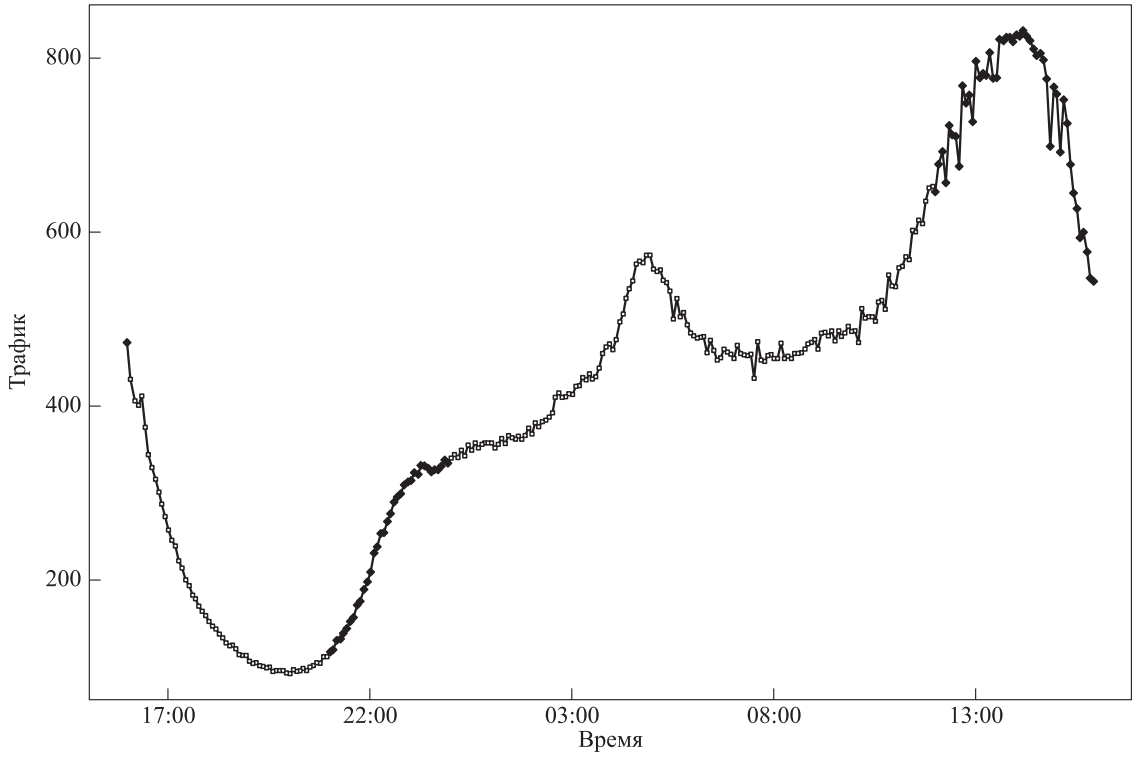

Рис. 6 Аномалии на временном ряде конкретного объекта

не перцентилей (первой компоненты) аномальны объекты, у которых меньше размах, и, наоборот, при фиксированном уровне второй компоненты — у которых больше значения перцентилей.

Недостаток метода непересекающихся окон заключается в том, что каждое наблюдение попадает только в одно окно, поэтому наблюдение для текущего окна может быть признано аномальным, хотя при увеличении интервала времени оно таковым уже не является (в частности, см. обсуждение подхода в статье [20]) (см. рис. 5).

Если же использовать скользящее окно некоторой ширины w, то каждое наблюдение (за исключением расположенных на краях) попадет в w окон.

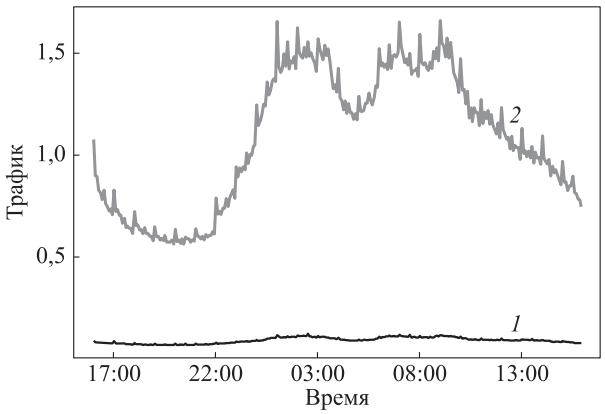

Рис. 7 Медиана (*1*) и верхний квартиль полученного трафика для всех объектов: *2* — 75% перцентиль

Таким образом, наблюдение может быть признано аномальным от 0 до w раз и можно говорить о некотором сопоставлении уровней аномальности конкретного наблюдения относительно других. Будем считать аномальное наблюдение:

- абсолютно аномальным (признано аномальным ровно w раз);
- относительно аномальным (признано аномальным от $\lceil w/2 \rceil$ до $w-1$ раз);
- промежуточно аномальным (признано аномальным от 1 до $\lceil w/2 \rceil - 1$ раз).

На рис. 9 приведены примеры разметки некоторых объектов в режиме скользящего окна шириной 1 ч.

В этом случае наибольшие по абсолютной величине значения не обязательно признаются аномальными, так как они могут соответствовать периодам общей высокой нагрузки на сеть. Для каждого наблюдения V_j по каждому объекту имеется некоторый уровень аномальности $z_j \in \{0, \ldots, w\}$, и в качестве правила кластеризации можно использовать некоторую функцию от $\{z_1, \ldots, z_n\}$. Например, на рис. 10 представлена кластеризация на основе решающего правила $\mathbb{I}\left(\{(1/n)\sum_i z_i > 0{,}5\}\right)$.

На рис. 8 не были отображены 3 объекта с наибольшими значениями главных компонент, располагающиеся в правом верхнем углу на рис. 10, так как подход на основе непересекающихся окон отнес к аномальным по общему трафику 37 объектов, большая часть из которых продемонстрирована. Метод скользящего окна в качестве аномальных сразу по обоим типам трафика отметил только эти 3 наблюдения, что действительно позволяет говорить об их существенных отличиях от других объектов.

4 Заключение

В работе рассмотрен статистический подход к выявлению аномальных нагрузок на узлах сетевой

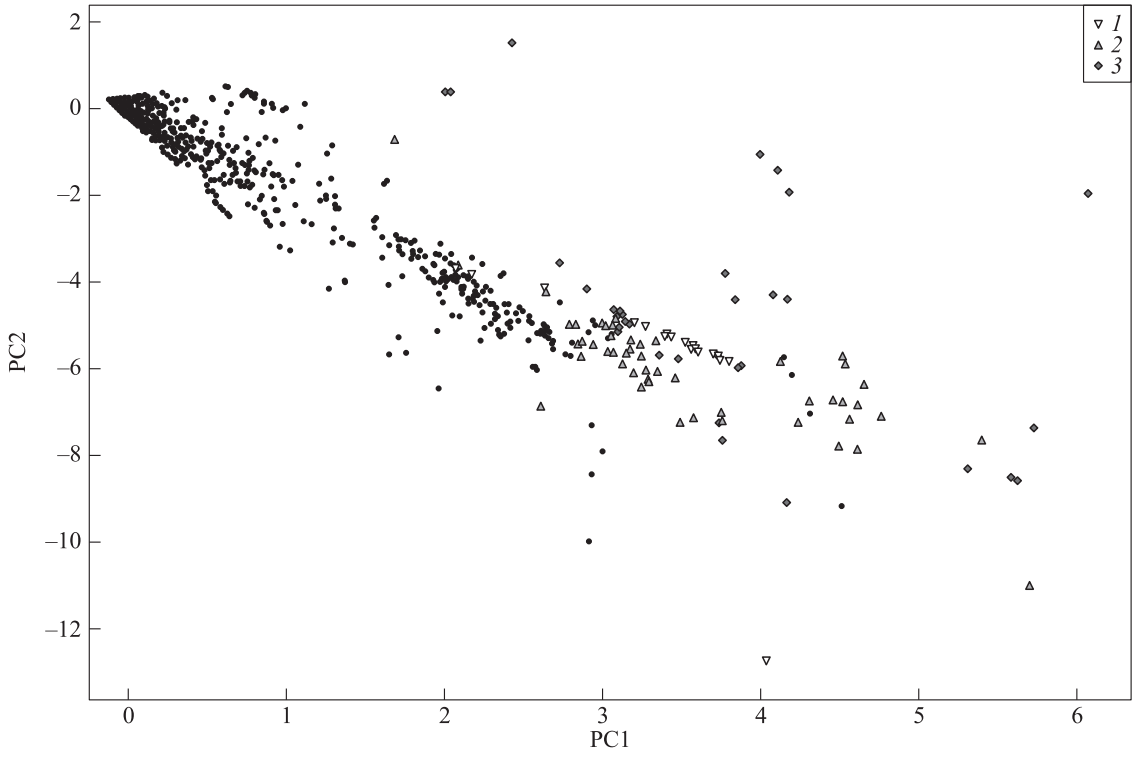

Рис. 8 Кластеризация объектов на группы с высоким входящим, исходящим и общим трафиком: *1* — аномален по полученному трафику; *2* — аномален по отправленному трафику; *3* — аномален по обоим типам трафика

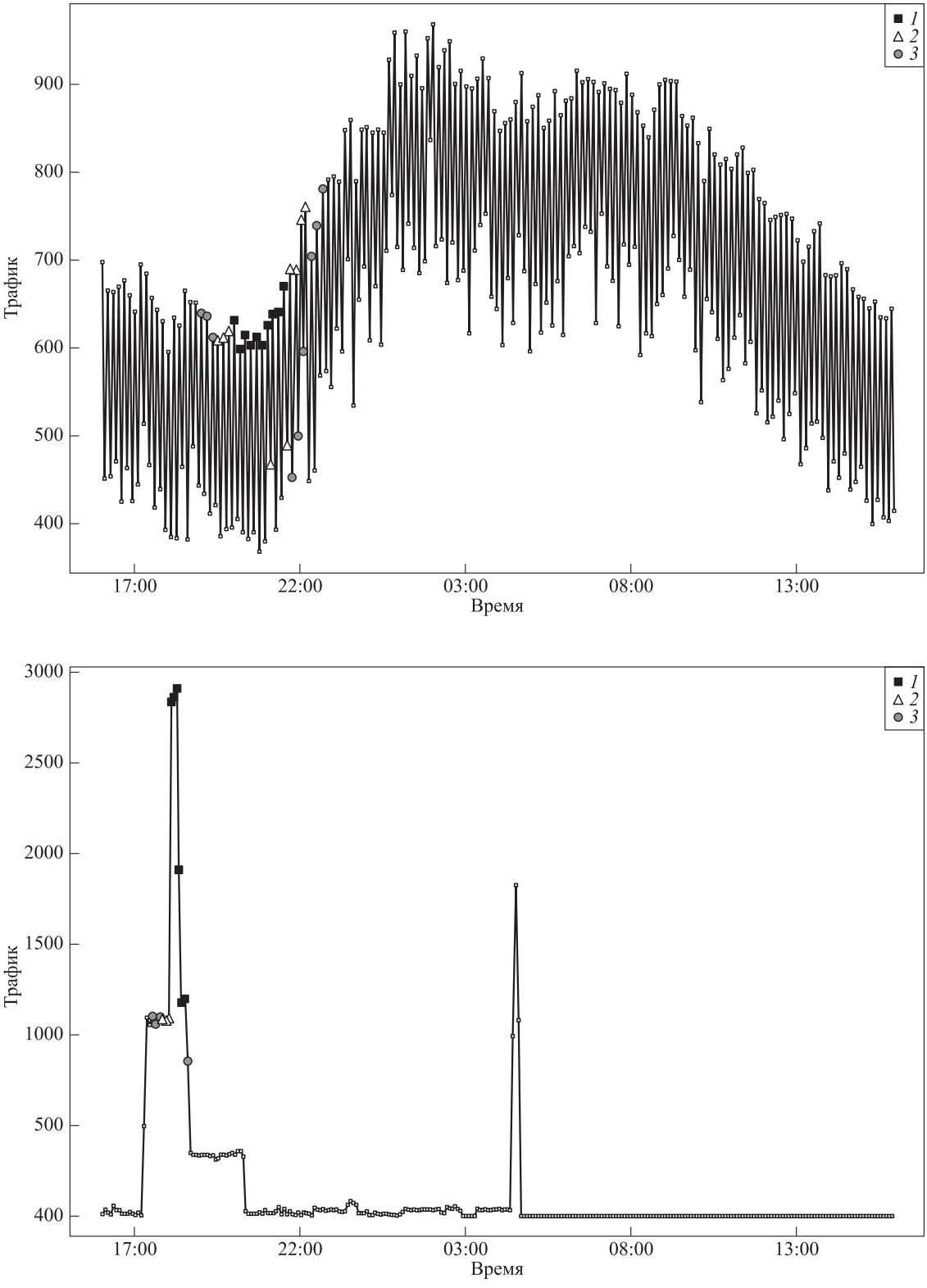

Рис. 9 Примеры разметки временных рядов трафика, скользящее окно: *1* — абсолютно аномальные; *2* — относительно аномальные; *3* — промежуточно аномальные

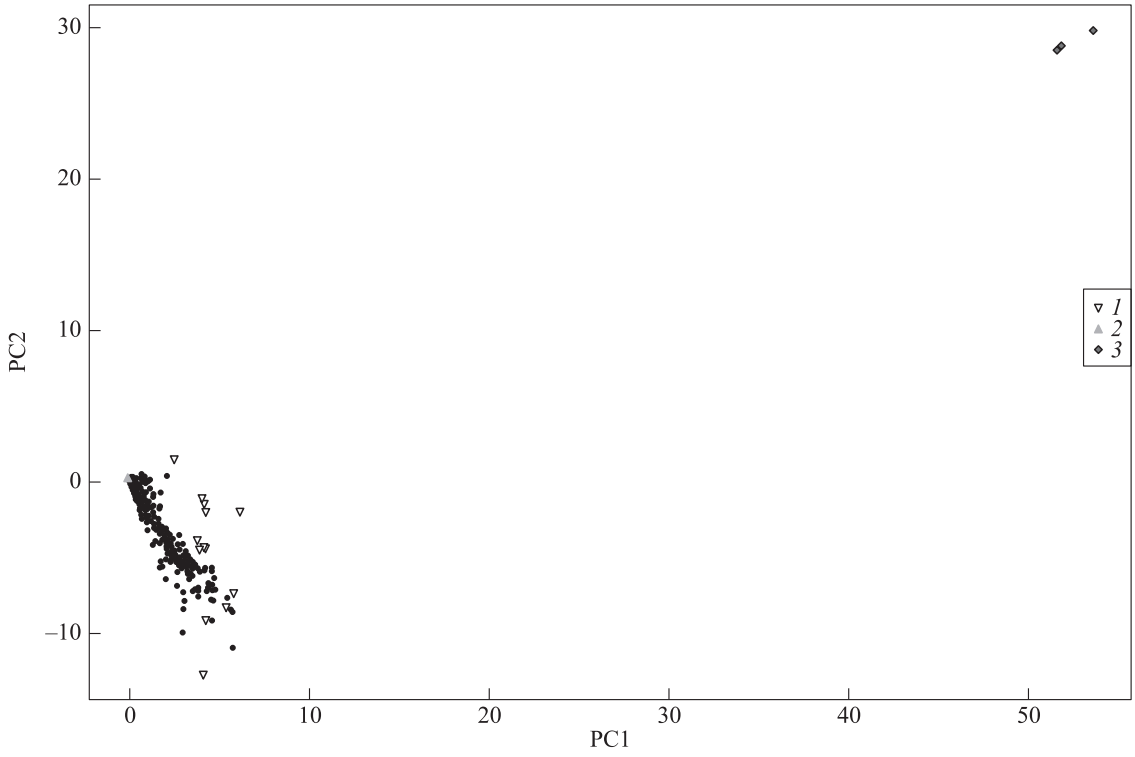

Рис. 10 Кластеризация объектов на группы с высоким входящим, исходящим и общим трафиком с использованием скользящего окна: *1* — аномален по полученному трафику; *2* — аномален по отправленному трафику; *3* — аномален по обоим типам трафика

вычислительной архитектуры. С использованием языка программирования R реализована процедура определения аномальных наблюдений в трафике в рамках хорошо соответствующего реальным данным предположения о возможности описания наблюдений обобщенным гамма-распределением. На основе анализа появления таких наблюдений в узлах может быть реализована процедура кластеризации объектов сети. Дальнейшие направления исследований в данной области могут быть связаны с построением статистических моделей для иных характеристик вычислительных узлов (например, загрузки процессора или используемой памяти) и решение описанной задачи кластеризации в расширенном признаковом пространстве.

Предложенная методика имеет определенный потенциал и для других задач в области телекоммуникаций, например связанных с распределением ресурсов для виртуальных машин [21, 22]. Алгоритмы такого рода могут быть реализованы в виде сервисов отдельных аналитических систем [23, 24] или в рамках цифровых платформ [25].

Авторы выражают признательность члену-корреспонденту РАН Р. Л. Смелянскому за ценные советы, касающиеся телекоммуникационной составляющей статьи, а также профессору В. Ю. Королеву за плодотворные обсуждения вопросов моделирования реальных данных с использованием различных семейств вероятностных распределений.

Литература

1. *Смелянский Р. Л.* Иерархические периферийные вычисления // Моделирование и анализ информационных систем, 2019. Т. 26. Вып. 1. С. 146–169. doi: 10.18255/1818-1015-2019-1-146-169.
2. *Smeliansky R.* Network powered by computing // Edge computing — technology, management and integration. — IntechOpen, 2023. 21 p. doi: 10.5772/intechopen.110178.
3. *Rossem S.V., Tavernier W., Colle D., Pickavet M., Demeester P.* Automated monitoring and detection of resource-limited NFV-based services // Conference on Network Softwarization: Softwarization Sustaining a Hyper-Connected World: en Route to 5G, NetSoft 2017. — Piscataway, NJ, USA: IEEE, 2017. Art. 8004220. doi: 10.1109/NETSOFT.2017.8004220.
4. *Malak D., Medard M., Andrews J. G.* Spatial concentration of caching in wireless heterogeneous networks // IEEE T. Wirel. Commun., 2021. Vol. 20. Iss. 6. P. 3397–3414. doi: 10.1109/TWC.2021.3049812.
5. *Zhang Z., Lu J., Chen H.* Controller robust placement with dynamic traffic in software-defined networking // Comput. Commun. 2022. Vol. 194. P. 458–467. doi: 10.1016/j.comcom.2022.07.018.

6. *Mesodiakaki A., Zola E., Kassler A.* Robust and energy-efficient user association and traffic routing in B5G HetNets // Comput. Netw., 2022. Vol. 217. Art. 109305. doi: 10.1016/j.comnet.2022.109305.
7. *Hosseinzadeh S., Amirmazlaghani M., Shajari M.* An aggregated statistical approach for network flood detection using gamma-normal mixture modeling // Comput. Commun., 2020. Vol. 152. P. 137–148. doi: 10.1016/j.comcom.2020.01.028.
8. *Abood M. S., Mustafa A. S., Mahdi H. F., Mohammed A.-F. A., Hamdi M. M., Hussein N. A.* The analysis of teletraffic and handover performance in cellular system // 3rd Congress (International) on Human–Computer Interaction, Optimization and Robotic Applications Proceedings. — Piscataway, NJ, USA: IEEE, 2021. P. 1–5. doi: 10.1109/HORA52670.2021.9461300.
9. *Parulekar M., Makowski A. M.* $M|G|\infty$ input processes: A versatile class of models for network traffic // IEEE INFOCOM Proceedings. — Piscataway, NJ, USA: IEEE, 1997. Vol. 2. P. 419–426. doi: 10.1109/INFCOM.1997.644490.
10. *Tabassum H., Dawy Z., Hossain E., Alouini M.-S.* Interference statistics and capacity analysis for uplink transmission in two-tier small cell networks: A geometric probability approach // IEEE T. Wirel. Commun., 2014. Vol. 13. Iss. 7. P. 3837–3852. doi: 10.1109/TWC.2014.2314101.
11. *Noor K., Shahid H., Obaid H. M., Rauf A., Yousaf A., Shahid A.* Hybrid underwater intelligent communication system // Wireless Pers. Commun., 2022. Vol. 125. Iss. 3. P. 2219–2238. doi: 10.1007/s11277-022-09653-7.
12. *Padhan A. K., Kumar S. H., Sahu P. R., Samantaray S. R.* Performance analysis of smart grid wide area network with RIS assisted three hop system // IEEE Transactions Signal Information Processing Networks, 2023. Vol. 9. P. 48–59. doi: 10.1109/TSIPN.2023.3239652.
13. *Gorshenin A., Korolev V., Kuzmin V., Zeifman A.* Coordinate-wise versions of the grid method for the analysis of intensities of non-stationary information flows by moving separation of mixtures of gamma-distribution // 27th European Conference on Modelling and Simulation Proceedings / Eds. W. Rekdalsbakken, R. T. Bye, and H. Zhang. — Dudweiler, Germany: Digitaldruck Pirrot GmbHP, 2013. P. 565–568. doi: 10.7148/2013-0565.
14. *Gorshenin A., Korolev V.* Modelling of statistical fluctuations of information flows by mixtures of gamma distributions // 27th European Conference on Modelling and Simulation Proceedings / Eds. W. Rekdalsbakken, R. T. Bye, and H. Zhang. — Dudweiler, Germany: Digitaldruck Pirrot GmbHP, 2013. P. 569–572. doi: 10.7148/2013-0569.
15. *Gorshenin A., Kuzmin V.* Online system for the construction of structural models of information flows // 7th Congress (International) on Ultra Modern Telecommunications and Control Systems and Workshops Proceedings. — Piscataway, NJ, USA: IEEE, 2015. P. 216–219. doi: 10.1109/ICUMT.2015.7382430.
16. *Gorshenin A., Kuzmin V.* On an interface of the online system for a stochastic analysis of the varied information flows // AIP Conf. Proc., 2016. Vol. 1738. Art. 220009. 4 p. doi: 10.1063/1.4952008.
17. *Горшенин А. К.* О некоторых математических и программных методах построения структурных моделей информационных потоков // Информатика и её применения, 2017. Т. 11. Вып. 1. С. 58–68. doi: 10.14357/19922264170105.
18. *Zonoozi M. M., Dassanayake P., Faulkner M.* Teletraffic modelling of cellular mobile networks // IEEE VTC P., 1996. Vol. 2. P. 1274–1277. doi: 10.1109/VETEC.1996.501517.
19. *Stacy E. W.* A generalization of the gamma distribution // Ann. Math. Stat., 1962. Vol. 33. Iss. 3. P. 1187–1192. doi: 10.1214/aoms/1177704481.
20. *Korolev V. Yu., Gorshenin A. K.* Probability models and statistical tests for extreme precipitation based on generalized negative binomial distributions // Mathematics, 2020. Vol. 8. Iss. 4. Art. 604. doi: 10.3390/math8040604.
21. *Huang B., Chen J., He Q., Wang B., Liu Z., Cheng Y.* HASO: A hot-page aware scheduling optimization method in virtualized NUMA systems // 7th Conference (International) on Information and Communication Systems Proceedings. — Piscataway, NJ, USA: IEEE, 2016. P. 68–73. doi: 10.1109/IACS.2016.7476088.
22. *Tian H., Li S., Wang A., Wang W., Wu T., Yang H.* Owl: Performance-aware scheduling for resource-efficient function-as-a-service cloud // 13th Symposium on Cloud Computing Proceedings. — New York, NY, USA: Association for Computing Machinery, 2022. P. 78–93. doi: 10.1145/3542929.3563470.
23. *Горшенин А. К.* Концепция онлайн-комплекса для стохастического моделирования реальных процессов // Информатика и её применения, 2016. Т. 10. Вып. 1. С. 72–81. doi: 10.14357/19922264160107.
24. *Gorshenin A. K., Kuzmin V. Yu.* Research support system for stochastic data processing // Pattern Recognition Image Analysis, 2017. Vol. 27. No. 3. P. 518–524. doi: 10.1134/S1054661817030117.
25. *Gorshenin A.* Toward modern educational IT-ecosystems: From learning management systems to digital platforms // 10th Congress (International) on Ultra Modern Telecommunications and Control Systems and Workshops Proceedings. — Piscataway, NJ, USA: IEEE, 2018. P. 329–333. doi: 10.1109/ICUMT.2018.8631229.

Поступила в редакцию 15.07.23

TOWARD CLUSTERING OF NETWORK COMPUTING INFRASTRUCTURE OBJECTS BASED ON ANALYSIS OF STATISTICAL ANOMALIES IN NETWORK TRAFFIC

A. K. Gorshenin[1,2], S. A. Gorbunov[2,3], and D. Yu. Volkanov[2]

[1]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation
[2]M. V. Lomonosov Moscow State University, 1 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation
[3]Moscow Center for Fundamental and Applied Mathematics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation

Abstract: The problem of detecting statistical anomalies (that is, outliers in relation to the typical values of upload and download traffic) of the load on the nodes of the network computing infrastructure is considered. The regular scaling in computing resources and storage as well as redirection of data flows is needed due to the increase of load in real systems. The procedure for detecting statistical anomalies in network traffic is proposed using the approximation of observations by the generalized gamma distribution for further clustering of network computing infrastructure objects in order to evaluate resource need. All computational statistical procedures described in the paper are implemented using the R programming language and they are applied for network traffic, simulated using a specialized architectural and software stand. The proposed approaches can also be used for a wider class of telecommunication problems.

Keywords: network infrastructure; network traffic; generalized gamma distribution; computational statistics; statistical hypothesis testing; anomaly detection; clustering

DOI: 10.14357/19922264230311 **EDN:** XHTMVI

Acknowledgments

This work was done with the support of MSU Program of Development, Project No. 23-SCH03-03. The research was carried out using the infrastructure of the Shared Research Facilities "High Performance Computing and Big Data" (CKP "Informatics") of FRC CSC RAS (Moscow).

References

1. Smelyansky, R. L. 2019. Ierarkhicheskie periferiynye vychisleniya [Hierarchical edge computing]. *Modelirovanie i analiz informatsionnykh sistem* [Modeling and Analysis of Information Systems] 26(1):146–169. doi: 10.18255/1818-1015-2019-1-146-169.
2. Smeliansky, R. 2023. Network powered by computing. *Edge computing — technology, management and integration*. IntechOpen. 21 p. doi: 10.5772/intechopen.110178.
3. Rossem, S. V., W. Tavernier, D. Colle, M. Pickavet, and P. Demeester. 2017. Automated monitoring and detection of resource-limited NFV-based services. *Conference on Network Softwarization: Softwarization Sustaining a Hyper-Connected World: en Route to 5G*. Piscataway, NJ: IEEE. 8004220. doi: 10.1109/NETSOFT.2017.8004220.
4. Malak, D., M. Medard, and J. G. Andrews. 2021. Spatial concentration of caching in wireless heterogeneous networks. *IEEE T. Wirel. Commun.* 20(6):3397–3414. doi: 10.1109/TWC.2021.3049812.
5. Zhang, Z., J. Lu, and H. Chen. 2022. Controller robust placement with dynamic traffic in software-defined networking. *Comput. Commun.* 194:458–467. doi: 10.1016/j.comcom.2022.07.018.
6. Mesodiakaki, A., E. Zola, and A. Kassler. 2022. Robust and energy-efficient user association and traffic routing in B5G HetNets. *Comput. Netw.* 217:109305. doi: 10.1016/j.comnet.2022.109305.
7. Hosseinzadeh, S., M. Amirmazlaghani, and M. Shajari. 2020. An aggregated statistical approach for network flood detection using gamma-normal mixture modeling. *Comput. Commun.* 152:137–148. doi: 10.1016/j.comcom.2020.01.028.
8. Abood, M. S., A. S. Mustafa, H. F. Mahdi, A.-F. A. Mohammed, M. M. Hamdi, and N. A. Hussein. 2021. The analysis of teletraffic and handover performance in cellular system. *3rd Congress (International) on Human–Computer Interaction, Optimization and Robotic Applications Proceedings*. Piscataway, NJ: IEEE. 1–5. doi: 10.1109/HORA52670.2021.9461300.
9. Parulekar, M., and A. M. Makowski. 1997. $M|G|\infty$ input processes: A versatile class of models for network traffic. *IEEE INFOCOM Proceedings*. Piscataway, NJ: IEEE. 2:419–426. doi: 10.1109/INFCOM.1997.644490.
10. Tabassum, H., Z. Dawy, E. Hossain, and M.-S. Alouini. 2014. Interference statistics and capacity analysis for

uplink transmission in two-tier small cell networks: A geometric probability approach. *IEEE T. Wirel. Commun.* 13(7):3837–3852. doi: 10.1109/TWC.2014.2314101.

11. Noor, K., H. Shahid, H. M. Obaid, A. Rauf, A. Yousaf, and A. Shahid. 2022. Hybrid underwater intelligent communication system. *Wireless Pers. Commun.* 125(3):2219–2238. doi: 10.1007/s11277-022-09653-7.
12. Padhan, A. K., S. H. Kumar, P. R. Sahu and S. R. Samantaray. 2023. Performance analysis of smart grid wide area network with RIS assisted three hop system. *IEEE Transactions Signal Information Processing Networks* 9:48–59. doi: 10.1109/TSIPN.2023.3239652.
13. Gorshenin, A., V. Korolev, V. Kuzmin, and A. Zeifman. 2013. Coordinate-wise versions of the grid method for the analysis of intensities of non-stationary information flows by moving separation of mixtures of gamma-distribution. *27th European Conference on Modelling and Simulation Proceedings*. Eds. W. Rekdalsbakken, R. T. Bye, and H. Zhang. Dudweiler, Germany: Digitaldruck Pirrot GmbHP. 565–568. doi: 10.7148/2013-0565.
14. Gorshenin, A., and V. Korolev. 2013. Modelling of statistical fluctuations of information flows by mixtures of gamma distributions. *27th European Conference on Modelling and Simulation Proceedings*. Eds. W. Rekdalsbakken, R. T. Bye, and H. Zhang. Dudweiler, Germany: Digitaldruck Pirrot GmbHP. 569–572. doi: 10.7148/2013-0569.
15. Gorshenin, A. K., and V. Kuzmin. 2015. Online system for the construction of structural models of information flows. *7th Congress (International) on Ultra Modern Telecommunications and Control Systems and Workshops Proceedings*. Piscataway, NJ: IEEE. 216–219. doi: 10.1109/ICUMT.2015.7382430.
16. Gorshenin, A., and V. Kuzmin. 2016. On an interface of the online system for a stochastic analysis of the varied information flows. *AIP Conference Proceedings* 1738(1):220009. 4 p. doi: 10.1063/1.4952008.
17. Gorshenin, A. K. 2017. O nekotorykh matematicheskikh i programmnykh metodakh postroeniya strukturnykh modeley informatsionnykh potokov [On some mathematical and programming methods for construction of structural models of information flows]. *Informatika i ee Primeneniya — Inform. Appl.* 11(1):58–68. doi: 10.14357/19922264170105.
18. Zonoozi, M. M., P. Dassanayake, and M. Faulkner. 1996. Teletraffic modelling of cellular mobile networks. *IEEE VTC P.* 2:1274–1277. doi: 10.1109/VETEC.1996.501517.
19. Stacy, E. W. 1962. A generalization of the gamma distribution. *Ann. Math. Stat.* 33(3):1187–1192. doi: 10.1214/aoms/1177704481.
20. Korolev, V. Yu., and A. K. Gorshenin. 2020. Probability models and statistical tests for extreme precipitation based on generalized negative binomial distributions. *Mathematics* 8(4):604. doi: 10.3390/math8040604.
21. Huang, B., J. Chen, Q. He, B. Wang, Z. Liu, and Y. Cheng. 2016. HASO: A hot-page aware scheduling optimization method in virtualized NUMA systems. *7th Conference (International) on Information and Communication Systems Proceedings*. Piscataway, NJ: IEEE. 68–73. doi: 10.1109/IACS.2016.7476088.
22. Tian, H., S. Li, A. Wang, W. Wang, T. Wu, and H. Yang. 2022. Owl: Performance-aware scheduling for resource-efficient function-as-a-service cloud. *13th Symposium on Cloud Computing Proceedings*. New York, NY: Association for Computing Machinery. 78–93. doi: 10.1145/3542929.3563470.
23. Gorshenin, A. K. 2016. Kontseptsiya onlayn-kompleksa dlya stokhasticheskogo modelirovaniya real'nykh protsessov [Concept of online service for stochastic modeling of real processes]. *Informatika i ee Primeneniya — Inform. Appl.* 10(1):72–81. doi: 10.14357/19922264160107.
24. Gorshenin, A. K., and V. Yu. Kuzmin. 2017. Research support system for stochastic data processing. *Pattern Recognition Image Analysis* 27(3):518–524. doi: 10.1134/S1054661817030117.
25. Gorshenin, A. 2018. Toward modern educational IT-ecosystems: From learning management systems to digital platforms. *10th Congress (International) on Ultra Modern Telecommunications and Control Systems and Workshops Proceedings*. Piscataway, NJ: IEEE. 329–333. doi: 10.1109/ICUMT.2018.8631229.

Received July 15, 2023

Contributors

Gorshenin Andrey K. (b. 1986) — Doctor of Science in physics and mathematics, associate professor, principal scientist, head of department, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; associate professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; agorshenin@frccsc.ru

Gorbunov Sergei A. (b. 2000) — master student, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; mathematician, Moscow Center for Fundamental and Applied Mathematics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; shadesilent@yandex.ru

Volkanov Dmitrii Yu. (b. 1979) — Candidate of Science (PhD) in physics and mathematics, associate professor, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; volkanov@asvk.cs.msu.ru