MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 27, NO. 1, 2020
journal homepage: www.mais-journal.ru

COMPUTER SYSTEM ORGANIZATION

Estimation of Length of Node-to-Node Paths Distribution in the Global Network

A. I. Kononova[1], A. V. Gorodilov[2]

DOI: 10.18255/1818-1015-2020-1-6-21

[1]National Research University of Electronic Technology, 1 Shokin sq., Moscow, Zelenograd 124498, Russia.
[2]Russkaya Moda (Russian fashion), 10 Rannyaya/Early str, Yaroslavl 150034, Russia.

MSC2020: 68M10
Research article
Full text in Russian

Received January 17, 2020
After revision February 25, 2020
Accepted February 28, 2020

The experiment aimed at finding a distribution of path lengths between nodes in the global network and an estimation of parameters of that distribution is described.In particular, the method of measurement of path length with traceroute utility of the GNU/Linux system and limitations on the selection of nodes imposed by traceroute are described. The measurement results are provided and high values of skewness and kurtosis for all resulting distributions are noted. Simulation model of this experiment was developed to test the experiment validity in the determination of distribution parameters in the global network. This model is also described. It is shown that high values of skewness and kurtosis of the measured distributions are not the result of the measurement technique, therefore the global network could not be described by the Barabási–Albert model. Several most viable hypotheses explaining differences in skewness and kurtosis of experimentally obtained path-length distribution estimations and values derived from the Barabási–Albert model are listed. Results of different hypotheses simulations are provided. It is shown that the most fitting hypothesis is that definitive influence on skewness and kurtosis of path-length distribution estimations is caused by the quasi pre-fractal structure of the global network.

Keywords: global network; routing; node-to-node distance distribution; experiment; Barabási–Albert model

INFORMATION ABOUT THE AUTHORS

Alexandra I. Kononova correspondence author	orcid.org/0000-0002-4178-3828. E-mail: illinc@bk.ru PhD.
Alexey V. Gorodilov	orcid.org/0000-0003-2887-8547. E-mail: kaverina@mail.ru PhD.

For citation: A. I. Kononova and A. V. Gorodilov, “Estimation of Length of Node-to-Node Paths Distribution in the Global Network”, *Modeling and analysis of information systems*, vol. 27, no. 1, pp. 6-21, 2020.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 27, № 1, 2020
сайт журнала: www.mais-journal.ru

COMPUTER SYSTEM ORGANIZATION

К вопросу об оценках распределения длин путей между узлами в глобальной сети

А. И. Кононова[1], А. В. Городилов[2]

DOI: 10.18255/1818-1015-2020-1-6-21

[1]Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «Московский институт электронной техники», пл. Шокина, 1, Москва, Зеленоград, 124498, Россия.
[2]Русская мода, ул. Ранняя, 10, Ярославль, 150034, Россия.

УДК 004.94
Научная статья
Полный текст на русском языке

Получена 17 января 2020 г.
После доработки 25 февраля 2020 г.
Принята к публикации 28 февраля 2020 г.

Описан эксперимент по оцениванию распределения длин путей между узлами в глобальной сети и его характеристик. В частности, показана методика измерения длины пути при помощи утилиты GNU/Linux traceroute и ограничения выбора узлов, налагаемые этим инструментом. Приведены результаты измерений, отмечены высокие значения асимметрии и эксцесса для всех полученных распределений. Описана имитационная модель эксперимента, разработанная для проверки корректности полученных оценок распределения длин путей между узлами в глобальной сети. Приведены результаты моделирования измерений. Показано, что высокие значения асимметрии и эксцесса измеренных распределений не обусловлены только методикой измерения, таким образом, глобальная сеть не описывается моделью Барабаши—Альберт. Перечислены основные гипотезы о причинах отличия асимметрии и эксцесса полученных экспериментально оценок распределения длин путей между узлами в глобальной сети от значений, соответствующих модели Барабаши—Альберт. Описаны результаты моделирования различных гипотез. Показано, что наиболее правдоподобной из них является предположение об определяющем влиянии квазипредфрактальной структуры глобальной сети на асимметрию и эксцесс оценок распределения длин путей между узлами.

Ключевые слова: глобальная сеть; маршрутизация; распределение длин путей; исследование структуры; безмасштабная модель Барабаши—Альберт

ИНФОРМАЦИЯ ОБ АВТОРАХ

Александра Игоревна Кононова
автор для корреспонденции
orcid.org/0000-0002-4178-3828. E-mail: illinc@bk.ru
канд. техн. наук, доцент.

Алексей Владиславович Городилов
orcid.org/0000-0003-2887-8547. E-mail: kaverina@mail.ru
канд. техн. наук, доцент.

Для цитирования: A. I. Kononova and A. V. Gorodilov, “Estimation of Length of Node-to-Node Paths Distribution in the Global Network”, *Modeling and analysis of information systems*, vol. 27, no. 1, pp. 6-21, 2020.

Введение

В настоящее время передача данных через глобальную сеть используется, прямо или косвенно, практически во всех приложениях. В частности, для повышения качества передачи мультимедийных данных в приложении IP-телефонии в 2011 году разрабатывалась методика передачи данных с учётом особенностей сети [1].

Данные в рамках разработанной методики передаются через пиринговую сеть, построенную аналогично файлообменному протоколу BitTorrent, что позволило обеспечить обмен данных между узлами с серыми IP-адресами, прямая передача между которыми в рамках TCP/IP невозможна. Соответственно, каждый узел, участвующий в этой сети, хранит постоянно обновляемый список узлов-ретрансляторов (их адреса и характеристики). При этом количество хранимых узлов не должно быть как слишком малым (это необходимо для успешной передачи данных), так и слишком большим (хранение полной копии структуры сети в памяти каждого узла создаст не только высокую нагрузку на память узла, но и высокий уровень служебного трафика для поддержания актуальности этой копии). Кроме того, сам набор хранимых узлов не должен быть однородным — для повышения скорости информационного обмена бóльшая часть хранимых узлов должна физически находиться в сегментах сети, смежных с текущим, но для поддержания целостности сети каждый узел должен хранить в памяти несколько узлов, удалённых от текущего, причём не смежных между собой.

Для подбора [2] наилучшего размера и структуры списка хранимых узлов в 2011–2012 годах была начата серия экспериментов по оценке распределения длин путей (РДП) сетевого уровня модели TCP/IP глобальной сети при помощи программы traceroute. Изначально оценивалась только средняя длина пути; в дальнейшем был рассмотрен вопрос о прочих характеристиках РДП, а также о корректности полученных оценок.

Понятие длины пути между узлами

Рассмотрим процесс передачи данных по сети (будем рассматривать сетевой уровень четырёхуровневой модели TCP/IP). Пусть необходимо переслать пакет данных от узла α к узлу β. Тогда маршрут передачи данных, скорее всего, будет включать некоторое количество промежуточных узлов, служащих ретрансляторами пакета.

Определение 1. *Пусть маршрут передачи данных от узла α к узлу β, $\alpha \neq \beta$, включает k пересылок узел-узел (hops):*

$$\alpha \longrightarrow \alpha_1 \longrightarrow \ldots \longrightarrow \alpha_{k-1} \longrightarrow \beta \tag{1}$$

тогда будем называть число k длиной пути от узла α к узлу β *и обозначать $\lambda(\alpha, \beta)$:*

$$\lambda(\alpha, \beta) = k. \tag{2}$$

Также положим

$$\lambda(\alpha, \alpha) = 0. \tag{3}$$

При использовании стека протоколов TCP/IP маршрут передачи данных на сетевом уровне выбирается близким к оптимальному; но, поскольку разные промежуточные узлы оптимизируют разные параметры передачи, в общем случае маршрут передачи данных от узла β к узлу α отличается от маршрута передачи от α к β. Соответственно, возможна ситуация $\lambda(\alpha, \beta) \neq \lambda(\beta, \alpha)$. При этом сетевой уровень любой современной сети не содержит петель.

Кроме того, со временем глобальная сеть изменяет свою структуру, и как маршрут передачи данных, так и его длина $\lambda(\alpha, \beta)$ могут изменяться.

Определение 2. *Множество вершин (узлов) графа G будем обозначать как V_G, множество рёбер графа G — как E_G.*

Определение 3. *Под* распределением длин путей между узлами (РДП) *графа G будем понимать ряд чисел $\zeta(x)$, для каждой физически возможной длины пути x, то есть для каждого целого $x \geqslant 0$ показывающий, как часто длина x встречается среди всех возможных путей в G:*

$$\zeta(x) = p\big(\lambda(\alpha, \beta) = x \mid \alpha, \beta \in V_G\big). \tag{4}$$

В дальнейшем, чтобы подчеркнуть отличие распределения $\zeta(x)$ от его оценок, будем называть его полным РДП *графа G.*

Непосредственно измерить полное РДП для глобальной сети (то есть измерить и проанализировать длины $\lambda(\alpha, \beta)$ для всех возможных пар узлов α и β) невозможно как из-за её гигантских размеров (по различным оценкам [3] глобальная сеть содержит от 10^8 до 10^{10} узлов), так и из-за отсутствия в стеке TCP/IP средств для измерения расстояния между двумя произвольно взятыми узлами.

1. Измерение длины пути в GNU/Linux

Для исследования маршрутов сетевого уровня в GNU/Linux предназначена утилита traceroute. При запуске в командной строке узла α команды `traceroute` β результатом будет маршрут передачи данных $\alpha \longrightarrow \beta$, причём одна строка вывода traceroute соответствует одному узлу маршрута. Далее путём синтаксического анализа можно определить длину пути $\lambda(\alpha, \beta)$. Для этого использовались возможности оболочки Bash и утилиты GNU/Linux (в частности, wc, grep, sed), передача результатов измерений с различных корневых узлов осуществлялась при помощи svn.

Определение 4. *Будем далее называть* корневым узлом *узел α, на котором выполняется traceroute, и* оконечным узлом — *узел β, аргумент команды traceroute.*

Соответственно, описанную ниже схему оценивания РДП глобальной сети будем называть схемой корневой узел-оконечные узлы (КУОУ).

Таким образом, для измерения длины пути $\lambda(\alpha, \beta)$ необходимо иметь право на запуск программ на узле α (корневом) и знать IP-адрес или имя узла β (оконечного). Также необходима возможность соединения α с β, то есть оконечный узел β должен иметь белый IP-адрес либо находиться в одной подсети с α.

Ограничения на выбор оконечного узла гораздо мягче, чем для корневого. Вследствие этого количество возможных оконечных узлов может быть намного больше, чем корневых. В дальнейшем будем обозначать множество оконечных узлов $Q = \{\beta_1, \beta_2, \dots \beta_n\} \subseteq V_G$.

Соответственно, задавшись парой из корневого узла α и множества оконечных узлов Q и измеряя при помощи traceroute расстояния $\lambda(\alpha, \beta_j)$ от α до каждого из оконечных $\beta_j \in Q$, можно получить некоторую оценку РДП глобальной сети. Эта оценка зависит от выбора корневого узла, от множества оконечных узлов и от времени (от текущей конфигурации глобальной сети).

1.1. Результаты измерений

Для запуска программ было использовано три доступных корневых узла — домашние и рабочие компьютеры участников исследования (обозначаемые далее соответственно как A, B и C).

Также для оценивания РДП по методу КУОУ необходимо множество уникальных IP-адресов оконечных узлов. Так как полученные результаты планировалось использовать для организации передачи данных между компьютерами, аналогичными A, B, C, в качестве оконечных узлов рассматривались пользовательские компьютеры, а для серых IP-адресов — шлюзы их провайдеров. Подобные узлы участвуют в файлообмене по протоколу BitTorrent.

Изначально предполагалось, что характеристики РДП могут зависеть от географического положения оконечных узлов, что отчасти связано с используемым языком общения. Таким образом, было рассмотрено три крупных торрент-трекера, доступных на момент начала исследования и различающихся как по специфике распространяемого контента и языку общения, так и по физическому расположению большинства пользователей: pornolab.net, rutracker.org и thepiratebay.org.

Путём мониторинга IP-адресов их пользователей, принимающих участие в информационном обмене, были получены три множества уникальных IP-адресов размерами соответственно 564, 5222 и 540, обозначаемые далее как P, R и T. Эти множества затем и использовались как множества оконечных узлов.

От корневых узлов A и C было выполнено по четыре разнесённых во времени измерения расстояний до каждого доступного множества оконечных узлов; от узла B, из-за ограниченного времени доступа — только по одному измерению. Таким образом, всего было получено 27 оценок РДП, характеризуемых идентификатором i, составленным из множества оконечных узлов $Q \in \{P, R, T\}$, корневого узла $\alpha \in \{A, B, C\}$ и порядкового номера измерения для пары (α, Q):

$$i \in \{PA_0, PA_1, PA_2, PA_3, PB_0, PC_0, PC_1, PC_2, PC_3, RA_0, \dots RC_3, TA_0, \dots TC_3\} \tag{5}$$

Измерения с одним порядковым номером j из-за протяжённости процесса измерения (более суток), а также по техническим причинам, были выполнены не строго одновременно. Тем не менее, они выполнялись в сопоставимое время (в течение одного месяца). Измерения с порядковыми номерами j и $j + 1$ разделяет не менее восьми месяцев.

Результаты измерений представлены на рис. 1 и 2.

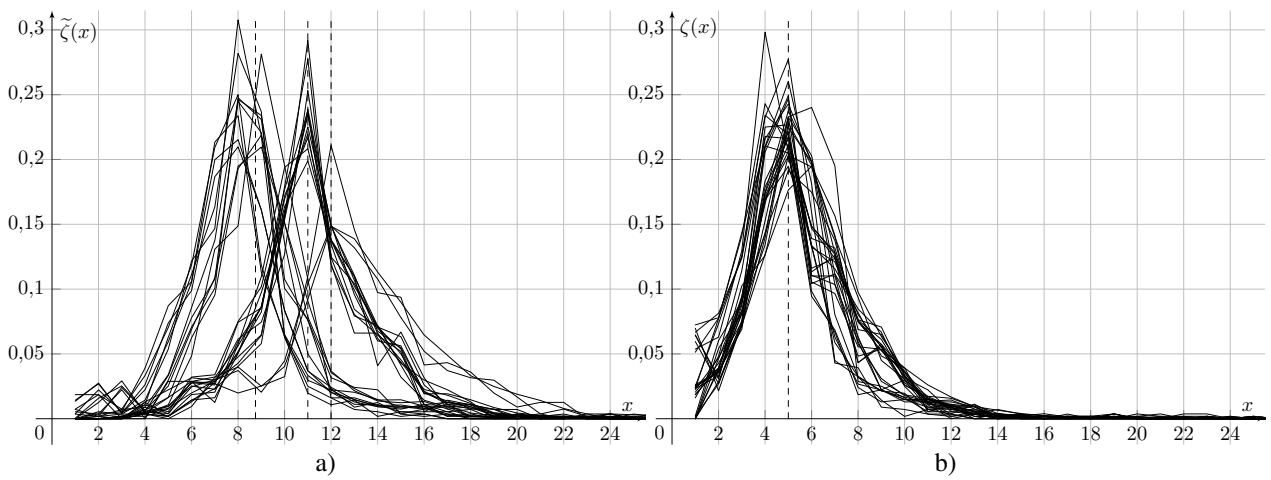

Fig. 1. Experimental evaluation of the NND distribution: a) including subnet of the root node provider, b) excluding subnet of the root node provider

Рис. 1. Экспериментально полученные оценки РДП: a) включая подсеть провайдера корневого узла, b) исключая подсеть провайдера корневого узла

На рис. 1 показано два варианта оценок РДП. Они были получены из следующих соображений. Все корневые узлы $\alpha \in \{A, B, C\}$ имели серые IP-адреса, а большинство оконечных $\beta \in Q$ — белые. Соответственно, маршрут передачи данных делился на две части: маршрут от α до шлюза его провайдера α_w (при этом все узлы $\alpha, \alpha_1, \dots \alpha_{w-1}$, находящиеся за шлюзом, имеют серые IP-адреса) и маршрут от шлюза провайдера α_w до β:

$$\alpha \longrightarrow \alpha_1 \longrightarrow \dots \alpha_{w-1} \longrightarrow \alpha_w \dots \longrightarrow \alpha_{k-1} \longrightarrow \beta \tag{6}$$

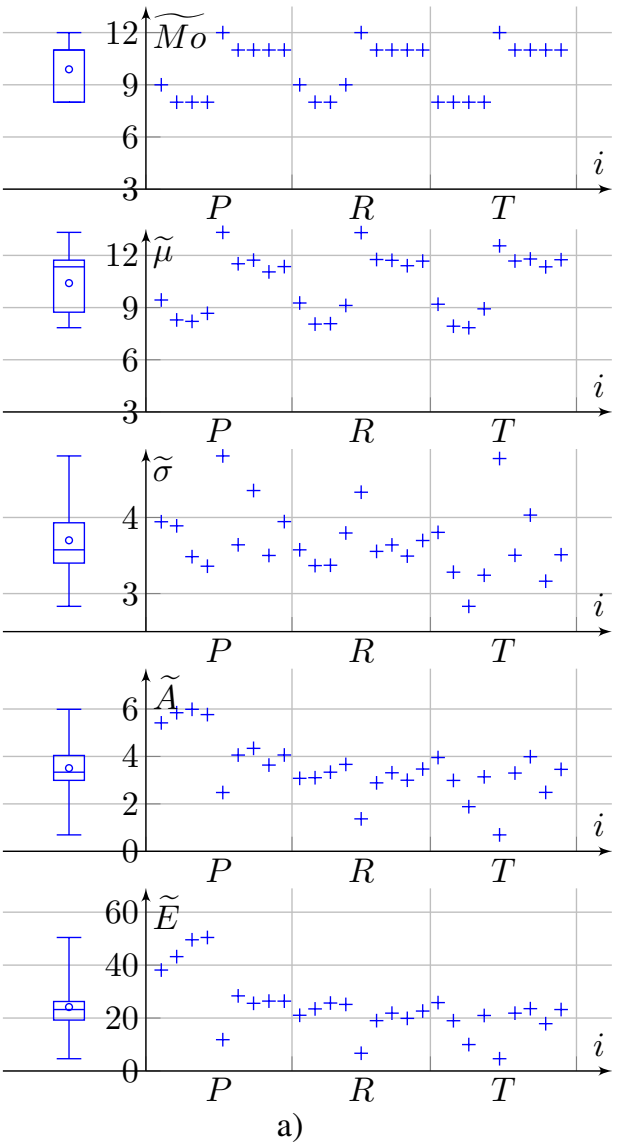

Fig. 2. Characteristics of experimentally obtained estimates of NND distribution: a) including a subnet of the root node provider, b) excluding a subnet of the root node provider

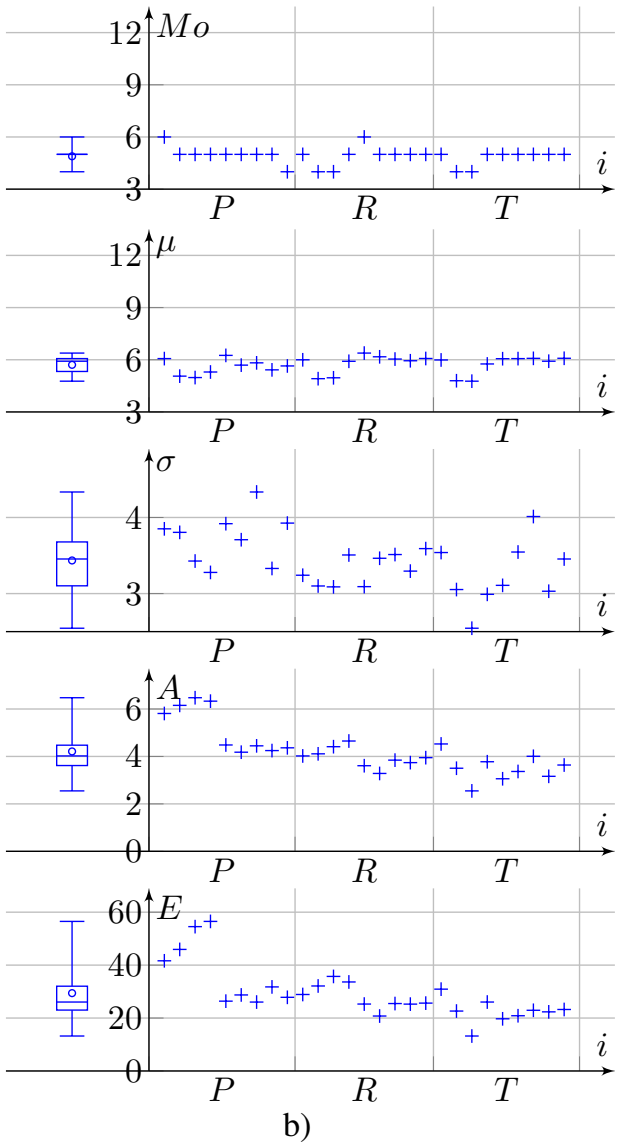

Рис. 2. Характеристики экспериментально полученных оценок РДП: a) включая подсеть провайдера корневого узла, b) исключая подсеть провайдера корневого узла

Для каждой пары α, β с помощью регулярных выражений определялось положение шлюза провайдера, при этом получались два значения длины пути: полная длина, включающая подсеть провайдера корневого узла

$$\tilde{\lambda}(\alpha, \beta) = k, \tag{7}$$

и скорректированная длина, исключающая подсеть провайдера корневого узла (то есть включающая только пересылки между узлами с белыми IP-адресами):

$$\lambda(\alpha, \beta) = k - w. \tag{8}$$

Для оконечных узлов β из подсети провайдера α принималось $\lambda(\alpha, \beta) = 1$.

После окончания измерений оба множества полученных значений длин путей $\tilde{\lambda}$ и λ обрабатывались GNU Octave. Соответственно, для каждого измерения i получались две оценки РДП:

– $\tilde{\zeta}_i(x)$, включающая подсеть провайдера корневого узла (рис. 1, a) и 2, a);
– $\zeta_i(x)$, исключающая подсеть провайдера корневого узла (рис. 1, b) и 2, b).

На рис. 2, a) и b) показаны соответственно характеристики оценок РДП $\tilde{\zeta}_i(x)$ и $\zeta_i(x)$ (мода Mo, среднее μ, среднеквадратическое отклонение σ, асимметрия A и эксцесс E). Значение эксцесса E вычисляется так, чтобы для нормального распределения он был бы равен нулю. Оси абсцисс соответствует идентификатор измерения i согласно (5). Таким образом, каждая точка справа от оси ординат соответствуют значению характеристики одной оценки РДП. Разброс этих значений иллюстрируют ящики с усами слева от оси ординат:

– окружность показывает усреднённое по всем измеренным оценкам РДП (среднее арифметическое) значение рассматриваемой характеристики;
– нижняя граница уса показывает минимальное значение характеристики;
– нижняя граница ящика показывает первую (нижнюю) квартиль;
– горизонтальная линия внутри ящика показывает медиану (вторую квартиль);
– верхняя граница ящика показывает третью (верхнюю) квартиль;
– верхняя граница уса показывает максимальное значение характеристики.

Анализ полученных результатов показал, что исключение из рассмотрения подсети провайдера корневого узла приводит к уменьшению разброса моды Mo и среднего μ и мало влияет на среднеквадратическое отклонение σ, асимметрию A и эксцесс E.

Кроме того, если изначально ожидалось, что определяющее влияние на РДП оказывает выбор оконечных узлов (соответственно, измерения на рис. 2 сгруппированы согласно (5) по множествам оконечных узлов), то рис. 1 и 2 показывают, что на форму и характеристики измеренного РДП влияет в основном корневой узел (даже после исключения подсети его провайдера).

Значения асимметрии и эксцесса как $\tilde{\zeta}_i(x)$, так и $\zeta_i(x)$ составляют (здесь и далее указано среднее значение без учёта выбросов, а в скобках — пятидесятипроцентный интервал):

$$\begin{cases} A_i \approx 3{,}5 \quad (3 \ldots 4{,}5) \\ E_i \approx 24 \quad (20 \ldots 30) \end{cases} \tag{9}$$

то есть обе эти характеристики существенно больше нуля. Ни для какого измерения i ни $\tilde{\zeta}_i(x)$, ни $\zeta_i(x)$ не имеют отрицательной асимметрии либо эксцесса.

Тем не менее, полученные по методу КУОУ оценки РДП могут отличаться от полного РДП глобальной сети. Так как точно измерить полное РДП глобальной сети не представляется возможным, для оценки корректности КУОУ необходимо рассмотреть модель сети, полное РДП которой доступно, и имитировать проведённый эксперимент для неё.

2. Имитационное моделирование погрешности эксперимента

Определение 5. *Будем называть количество вершин (узлов)* $|V_G|$ *графа* G размером графа.

Для проверки корректности полученных оценок РДП и его характеристик было произведено имитационное моделирование погрешности эксперимента по схеме КУОУ. Моделью глобальной сети был выбран граф G, вершины которого соответствуют узлам сети, а рёбра — возможности прямой пересылки данных на сетевом уровне. С учётом отсутствия петель на сетевом уровне G должен быть связным деревом.

Целью моделирования является сопоставление полного РДП $\zeta(x)$ и его оценок $\zeta_i(x)$, а также их характеристик, в частности, асимметрии и эксцесса.

Таким образом, размер G должен позволять:

– найти расстояния $\lambda(\alpha, \beta)$ для всех возможных пар вершин $\alpha, \beta \in V_G$ и рассчитать полное РДП $\zeta(x)$;
– выполнить имитацию схемы КУОУ, получить несколько оценок РДП $\zeta_i(x)$ и сравнить их с полным РДП.

Размеры множеств оконечных узлов, использованных в экспериментальном исследовании, составляют от 500 до 5000 узлов. Соответственно, минимальным размером, позволяющим выполнить оценивание РДП по схеме КУОУ, было принято $|V_G| = 10^3$ вершин [4]. Максимальным размером, при котором можно выполнить расчёт за приемлемое время, после распараллеливания оказался $|V_G| = 10^6$ вершин.

Граф G представлялся в памяти компьютера в виде сжатой матрицы смежности, то есть вектором ρ_G из $|V_G|$ списков, каждый из которых соответствует вершине графа и содержит номера смежных с ней вершин. Для расчёта длин путей использован поиск в ширину, позволяющий для заданной вершины $\alpha \in V_G$ рассчитать расстояния до всех прочих вершин графа G; его сложность для представления графа в виде ρ_G равна $O(|V_G| + |E_G|)$ [5]. При расчёте полного РДП поиск в ширину необходимо выполнить для всех вершин, таким образом, сложность возрастает в $|V_G|$ раз. Для дерева сложность поиска в ширину составляет $O(|V_G|)$, так что сложность расчёта полного РДП составляет $O(|V_G|^2)$.

Для каждого графа рассчитывалось несколько оценок РДП $\zeta_i(x)$ по схеме КУОУ. Для каждой оценки i корневая вершина $\alpha \in V_G$, соответствующая корневому узлу в эксперименте и оконечные вершины $Q \subset V_G$, соответствующие оконечным узлам, выбирались заново случайным образом. Количество оконечных вершин $|Q|$ во время предварительных вычислений варьировалось в диапазоне 500 … 5000. Так как не было выявлено существенного различия результатов, для итогового моделирования было принято $|Q| = 500$.

Рост графа и расчёт распределения длин путей реализованы на C++ (стандарт C++11). Для ускорения расчёт распараллелен с помощью OpenMP, что позволило вырастить дерево из $|V_G| = 10^6$ вершин, обработать его и сохранить результаты в текстовом файле менее чем за двое суток. Дальнейшие расчёты проводились в GNU Octave.

Одному прогону r модели роста γ для $|V_G| = 10^v$ соответствует одно дерево $G_r^{\gamma,v}$ и, соответственно, одно полное РДП $\zeta^{G_r^{\gamma,v}}(x)$, а также несколько его оценок $\zeta_i^{G_r^{\gamma,v}}(x)$ (здесь i — порядковый номер оценки). Таким образом, дереву $G_r^{\gamma,v}$ соответствует один набор характеристик полного РДП $\left(Mo(G_r^{\gamma,v}), \mu(G_r^{\gamma,v}), \sigma(G_r^{\gamma,v}), A(G_r^{\gamma,v}), E(G_r^{\gamma,v})\right)$ и множество наборов характеристик его оценок $\left(Mo_i(G_r^{\gamma,v}), \mu_i(G_r^{\gamma,v}), \sigma_i(G_r^{\gamma,v}), A_i(G_r^{\gamma,v}), E_i(G_r^{\gamma,v})\right)$. Для краткости будем писать $\zeta(x)$ и $\zeta_i(x)$, а также Mo, μ, σ, A, E и $Mo_i, \mu_i, \sigma_i, A_i, E_i$, если это не приведёт к неоднозначности.

2.1. Модель Барабаши—Альберт

Чаще всего для моделирования растущих сетей [3, 6] применяется безмасштабная модель Барабаши—Альберт [7] (далее будем обозначать эту модель γ = BA), полагающая, что узлы присоединяются к сети постепенно, следуя *принципу предпочтительного присоединения,* то есть вероятность p_j присоединения новой вершины v_{n+1} к каждой из существующих вершин $v_j, j \in \{1, \dots n\}$ пропорциональна количеству $|v_j|$ уже имеющихся связей:

$$p_j = \frac{|v_j|}{\sum_{k=1}^{n} |v_k|}. \tag{10}$$

Соответственно, первыми РДП и их оценки по схеме КУОУ были рассчитаны именно для безмасштабных деревьев $G_r^{\mathrm{BA},v}$ (рис. 3 и 4).

На рис. 3, а) показаны полные РДП для различных деревьев Барабаши—Альберт из 10^5 вершин ($G_r^{\mathrm{BA},5}$). Видно, что они мало отличаются друг от друга, всегда унимодальны и имеют схожую с нормальной вершину и тонкие хвосты.

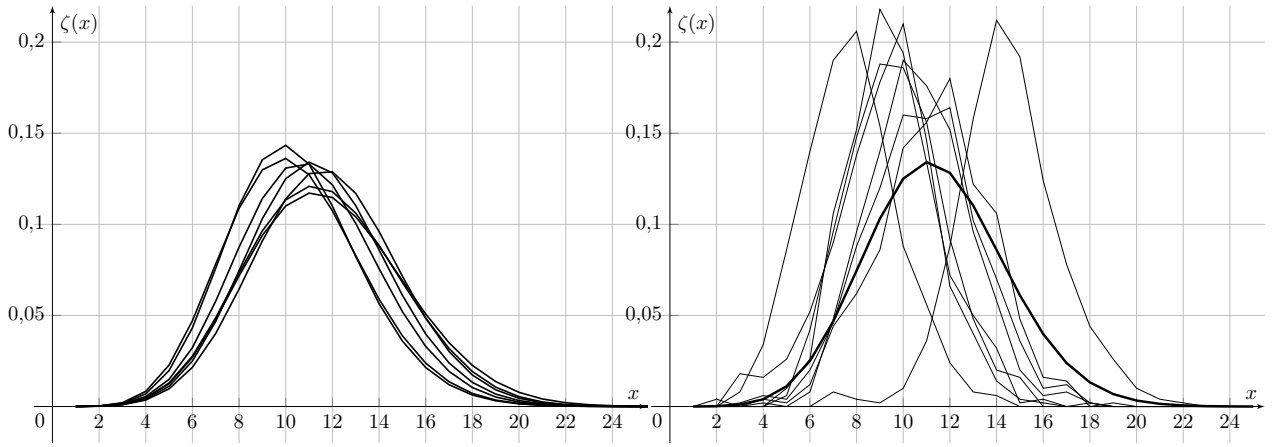

Fig. 3. Full NND distributions dispersion for different scale-free trees $G_r^{\mathrm{BA},5}$ (a); full NND distribution $\zeta(x)$ (thick line) and its estimates $\zeta_i(x)$ (thin lines) for the $G_4^{\mathrm{BA},5}$ (b)

Рис. 3. Разброс полных РДП для различных деревьев Барабаши—Альберт $G_r^{\mathrm{BA},5}$ (a); полное РДП $\zeta(x)$ (жирная линия) и его оценки $\zeta_i(x)$ (тонкие линии) для $G_4^{\mathrm{BA},5}$ (b)

На рис. 3, b) для одного из этих деревьев (полученного на четвёртом прогоне, то есть $G_4^{\mathrm{BA},5}$) показано полное РДП и его оценки. Большинство оценок $\zeta_i(x)$ рис. 3, b) выглядит более островершинными, чем полное $\zeta(x)$, но менее островершинными, чем экспериментально полученные оценки РДП глобальной сети (рис. 1, a, b).

Рассмотрим подробнее характеристики РДП. Так как общее количество полученных в результате моделирования оценок РДП для различных деревьев измеряется сотнями, характеристики каждой отдельной оценки РДП $\zeta_i(x)$ не могут быть приведены в рамках статьи. Соответственно, далее будем рассматривать различные деревья $G_r^{\mathrm{BA},v}$ и разброс характеристик оценок, полученных для каждого из них (для экспериментально полученных оценок РДП, полученных для глобальной сети, аналогичный разброс показывают ящики с усами в левой части рис. 2 a) и b).

На рис. 4 показаны характеристики полных РДП и их оценок для множества деревьев разного размера. Одна вертикаль соответствует одному прогону (одному дереву $G_r^{\mathrm{BA},v}$). Прогоны сгруппированы по размеру дерева $|V_G| = 10^v$ (показан вверху рисунка).

Внутри группы ось абсцисс соответствует номеру прогона модели r (показан вверху рисунка под $|V_G|$). Ниже показаны характеристики РДП дерева $G_r^{\mathrm{BA},v}$ и его оценок. Для каждой характеристики χ (где χ может быть модой Mo, средним μ, среднеквадратическим отклонением σ, асимметрией A и эксцессом E) и каждого дерева $G_r^{\mathrm{BA},v}$:

- чёрной точкой показано значение характеристики полного РДП $\chi(G_r^{\mathrm{BA},v})$;
- ящиком с усами на той же вертикали в тех же осях показан разброс характеристик оценок РДП $\chi_i(G_r^{\mathrm{BA},v})$ по i: среднее арифметическое по всем оценкам i (окружность), минимальное значение (нижняя граница уса), первая квартиль (нижняя граница ящика), медиана (линия внутри ящика), третья квартиль (верхняя граница ящика) и максимальное значение (верхняя граница уса).

Прежде всего из рис. 4 видно, что измерение РДП по схеме КУОУ в общем случае не позволяет получить полное РДП графа. Характеристики получаемых оценок значительно отличаются от характеристик полных РДП соответствующих деревьев.

Мода Mo_i и средняя длина пути μ_i для оценок РДП для деревьев небольших размеров (для $|V_G|$, сравнимых с объёмом выборки оконечных узлов) могут рассматриваться как оценки Mo и μ

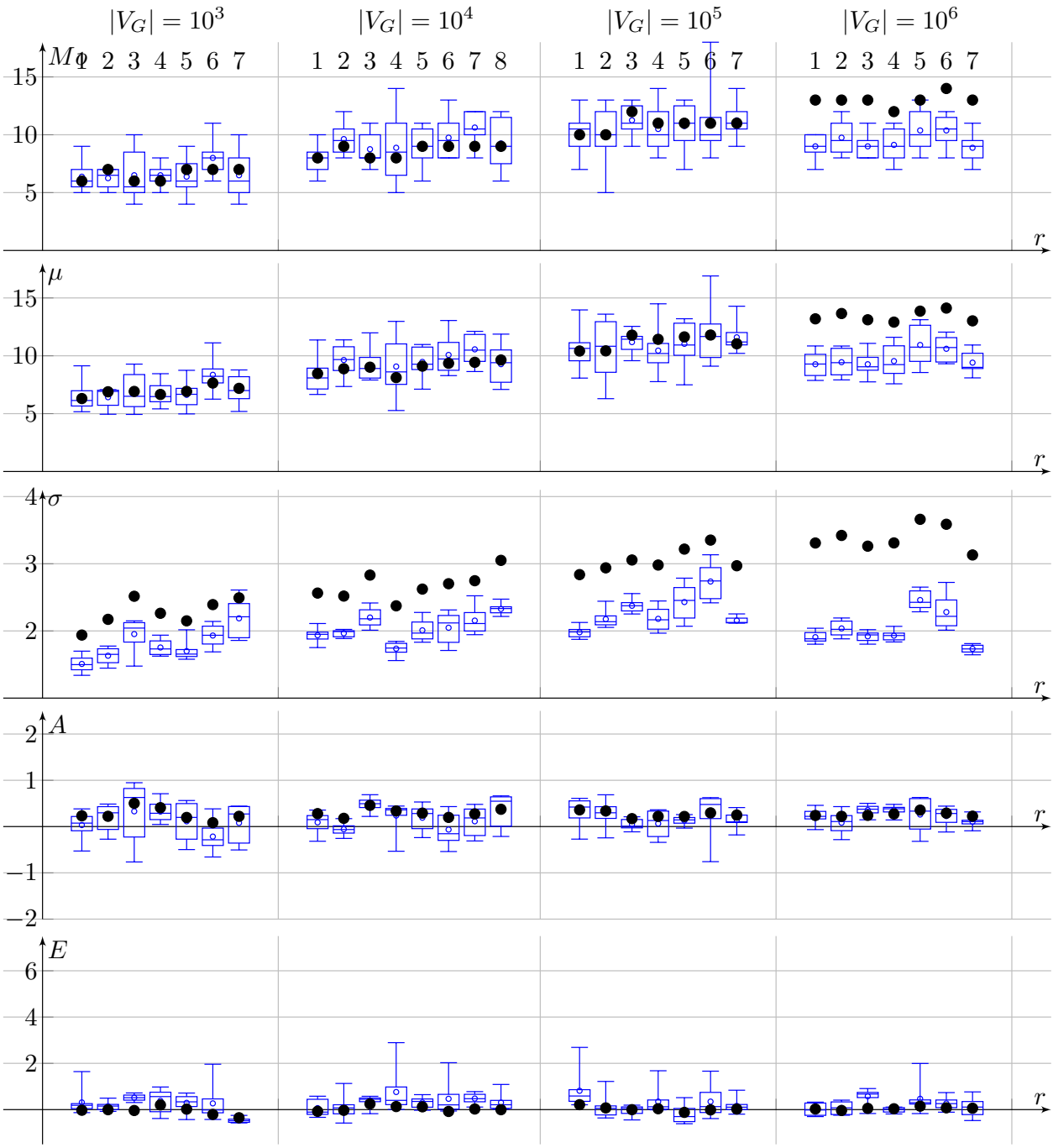

Fig. 4. Characteristics of full NND distribution in the scale-free trees $G_r^{\mathrm{BA},v}$ (black dots) and dispersion of the characteristics of the estimates (boxplots)

Рис. 4. Характеристики полного РДП в деревьях Барабаши—Альберт $G_r^{\mathrm{BA},v}$ (чёрные точки) и разброс характеристик его оценок (ящики с усами)

полного РДП, но уже для $|V_G| = 10^6$ значения Mo_i и μ_i существенно ниже Mo и μ соответственно. Таким образом, по полученным в результате эксперимента для глобальной сети по схеме КУОУ значениям Mo_i и μ_i также нельзя судить даже о порядке величины Mo и μ полного РДП глобальной сети.

Среднеквадратическое отклонение σ_i всех оценок РДП существенно ниже среднеквадратического отклонения σ полного РДП для всех деревьев, причём при увеличении количества узлов $|V_G| = 10^v$ растёт и систематическая погрешность оценивания.

Тем не менее рис. 4 показывает и то, что для каждого $|V_G|$ значения Mo_i, μ_i и σ_i оценок РДП имеют не вполне произвольные значения, а коррелируют со значениями Mo, μ и σ полного РДП.

Рассмотрим асимметрию и эксцесс полученных РДП. Асимметрия A полных РДП деревьев Барабаши—Альберт $\zeta(x)$ имеет небольшое положительное значение, а эксцесс E близок к нулю:

$$\begin{cases} A \approx 0{,}3 & (0{,}2 \dots 0{,}4) \\ E \approx 0 & (-0{,}1 \dots +0{,}2) \end{cases}, \tag{11}$$

С ростом количества $|V_G|$ вершин в дереве эти величины стабилизируются.

Отличие асимметрии A_i и эксцесса E_i оценок РДП от значений A и E полных РДП составляет сотни процентов, что связано прежде всего с близостью A и E к нулю. При этом асимметрия при оценивании по схеме КУОУ немного занижается, а эксцесс — немного завышается:

$$\begin{cases} A_i \approx 0{,}2 & (0{,}0 \dots 0{,}3) \\ E_i \approx 0{,}3 & (-0{,}2 \dots +0{,}8) \end{cases}. \tag{12}$$

Таким образом:

- с ростом количества $|V_G|$ вершин в дереве не растут ни асимметрия A и эксцесс E полных РДП деревьев Барабаши—Альберт, ни их оценки по КУОУ A_i и E_i;
- хотя бы одна оценка E_i для большинства прогонов отрицательна;
- ни одна оценка A_i и E_i ни для какого прогона не достигает полученных в результате эксперимента значений (9) и даже не приближается к ним.

Таким образом, различие между асимметрией и эксцессом экспериментально полученных для сетевого уровня глобальной сети данных и распределениями длин путей в деревьях Барабаши—Альберт обусловлено не случайными отклонениями, не методикой измерения и не отличием в размерах.

Соответственно, различие обусловлено выбранной моделью роста дерева, то есть модель Барабаши—Альберт не отражает свойств глобальной сети [8].

2.2. Модифицированные модели

Отличие экспериментально полученных оценок A_i и E_i от предсказываемых безмасштабной моделью Барабаши—Альберт может быть обусловлено различными аспектами роста реальной глобальной сети. Тем не менее, сама концепция постепенного роста сети и предпочтительного присоединения логична, так что была предпринята попытка внести в модель Барабаши—Альберт небольшие изменения и исследовать их влияние на асимметрию и эксцесс РДП выращиваемых графов.

Были выдвинуты три гипотезы о том, какой именно фактор оказывает определяющее влияние на значения асимметрии и эксцесса оценок РДП:

1. Ограничение степени узла.

   В оригинальной модели Барабаши—Альберт предполагается, что количество связей $|\nu_j|$ вершины может быть сколь угодно большим. Это верно для связей, не создающих нагрузку на узлы (таких, как web-ссылки), но неверно для сетевого уровня. Предположим, что количество связей вершины ограничено сверху некоторым числом M:

   $$|\nu_j| \leqslant M \tag{13}$$

   Таким образом, новая вершина не может соединяться ребром с вершиной, уже имеющей M связей; среди остальных вершин выбор происходит по принципу предпочтительного присоединения.

2. Удаление узлов.
   Узлы не только добавляются к глобальной сети, но и отключаются от неё. Рассмотрим модель, в которой на каждом шаге либо (с некоторой вероятностью p) удаляется случайно выбранная вершина, либо добавляется новая (с вероятностью $1 - p$). При удалении вершины её связи перераспределяются между оставшимися, чтобы сохранить связность.
3. Квазипредфрактальная структура сети.
   Глобальная сеть не гомогенна. Она состоит из множества подсетей различных провайдеров, причём каждый провайдер, предоставляющий связь конечным пользователям, сам пользуется услугами более крупного (магистрального) провайдера.
   Опишем рост графа G как совокупности нескольких подграфов G_s:

$$G = \bigcup_{s=1}^{m} G_s \tag{14}$$

На каждом шаге либо к одному из G_s присоединяется новая вершина, либо создаётся новый подграф G_{m+1}. Вероятности этих событий пропорциональны размерам G_s и G_{m+1}, то есть также соответствуют принципу предпочтительного присоединения. Новая вершина к подграфу G_s присоединяется аналогично модели Барабаши—Альберт. При создании нового подграфа G_{m+1} для него также по принципу предпочтительного присоединения выбирается магистральный провайдер — один из существующих подграфов G_s, далее G_{m+1} соединяется с G_s.
Таким образом, соединение подграфов G_s в графе G аналогично соединению вершин в каждом из подграфов. При точном совпадении структуры граф G можно было бы назвать предфрактальным [9].

Определение 6. *Будем называть* квазипредфрактальным *граф G, состоящий из нескольких подграфов G_s, соединённых рёбрами, причём модель роста всех G_s одинакова и аналогична модели роста самого G.*

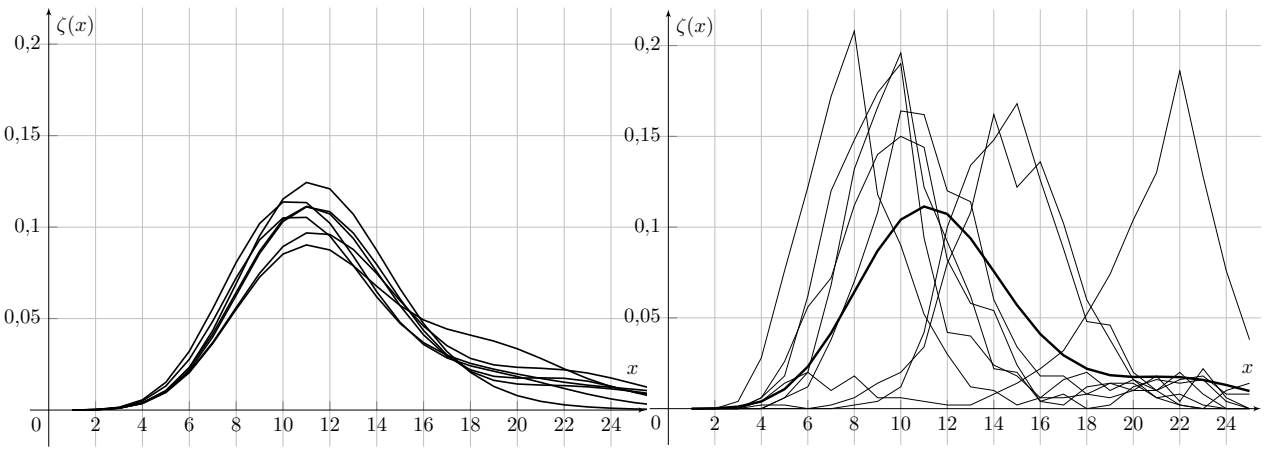

Fig. 5. Full NND distributions scattering in the quasi-pre-fractal trees $G_r^{\mathrm{PF},5}$ (a); full NND distribution $\zeta(x)$ (thick line) and its estimates $\zeta_i(x)$ (thin lines) for the $G_1^{\mathrm{PF},5}$ (b)

Рис. 5. Разброс полных РДП для квазипредфрактальных деревьев $G_r^{\mathrm{PF},5}$ (a); полное РДП $\zeta(x)$ (жирная линия) и его оценки $\zeta_i(x)$ (тонкие линии) для $G_1^{\mathrm{PF},5}$ (b)

Соответственно, в исследовании были рассмотрены три модели роста, каждая из которых реализует только одну из описанных гипотез. Модели, реализующие ограничение степени узла и удаление узлов из сети, приводят к результатам, качественно не отличающимся от модели Барабаши—Альберт, и, соответственно, не рассматриваются в данной статье.

Квазипредфрактальная модель, описывающая рост сети как рост множества связанных друг с другом подсетей (будем обозначать её γ = PF), приводит к отличным от Барабаши—Альберт результатам (рис. 5 и 6).

Полные РДП $\zeta(x)$ на рис. 5 отличаются от рис. 3 наличием более или менее толстого правого хвоста. Его толщина обусловлена соотношением размеров подграфов G_s и порядком их соединения.

Полное РДП квазипредфрактального дерева может иметь:

- более или менее толстый правый хвост, аналогичный рис. 5 (с крупным подграфом соседствует несколько меньших — наиболее распространённый среди приведённых результатов вариант);
- один или несколько ярко выраженных побочных максимумов (два или три конкурирующие подграфа сопоставимых размеров — $G_2^{\mathrm{PF},4}$, рис. 7, a);
- тонкий хвост, как для РДП деревьев Барабаши—Альберт (сформировался только один крупный подграф; прочие, если и существуют, ничтожно малы — $G_1^{\mathrm{PF},6}$, рис. 7, b).

В последнем случае, когда граф $G_r^{\mathrm{PF},v}$ фактически совпадает со своим крупнейшим подграфом, его можно считать выращенным по модели Барабаши—Альберт. Соответственно, и асимметрия и эксцесс полного РДП, и их оценки будут близки к нулю и иметь малый разброс. Действительно,

$$\begin{cases} A_i(G_1^{\mathrm{PF},6}) \approx A(G_1^{\mathrm{PF},6}) \approx 0{,}3 \\ E_i(G_1^{\mathrm{PF},6}) \approx E(G_1^{\mathrm{PF},6}) \approx 0 \end{cases}. \tag{15}$$

В случае нескольких максимумов и, соответственно, нескольких крупных подграфов, асимметрия и эксцесс полного РДП могут быть как положительными, так и отрицательными, в зависимости от взаимного расположения максимумов. Так, $E(G_2^{\mathrm{PF},4}) < 0$.

Так как подграфы $G_s \subset G_r^{\mathrm{PF},v}$ и связи между ними организуются случайно в процессе роста графа $G_r^{\mathrm{PF},v}$, форма полных РДП для разных прогонов r различается сильнее, чем для модели Барабаши—Альберт. Соответственно, все характеристики на рис. 6 (организованном аналогично рис. 4) имеют большой разброс, не уменьшающийся с увеличением размеров деревьев $|V_G|$.

Максимально возможные значения асимметрии и эксцесса полных РДП квазипредфрактальных деревьев растут с увеличением $|V_G|$. Так, для квазипредфрактальных деревьев из $10^5 \ldots 10^6$ вершин усреднённые асимметрия и эксцесс без учёта выбросов:

$$\begin{cases} A \approx 1 \quad (0{,}5 \ldots 1{,}2) \\ E \approx 1{,}2 \quad (0{,}5 \ldots 1{,}8) \end{cases}, \tag{16}$$

и растут с ростом $|V_G|$ (при этом увеличивается и разброс распределений длин путей).

Из рис. 6 видно, что характеристики оценок РДП по схеме КУОУ для квазипредфрактальных деревьев отличаются от характеристик полных РДП ещё больше, чем для деревьев Барабаши—Альберт, что подтверждает сделанный ранее вывод о том, что результаты, полученные по КУОУ экспериментально, не могут рассматриваться как характеристики РДП глобальной сети.

При оценивании по схеме КУОУ асимметрия и эксцесс унимодальных полных РДП с толстым хвостом чаще завышаются, чем занижаются, причём большие значения A и E завышаются сильнее. Так, для $|V_G| \in [10^5, 10^6]$ вершин:

$$\begin{cases} A_i \approx 1{,}2 \quad (1{,}0 \ldots 1{,}8) \\ E_i \approx 2{,}7 \quad (1{,}7 \ldots 4{,}3) \end{cases}, \tag{17}$$

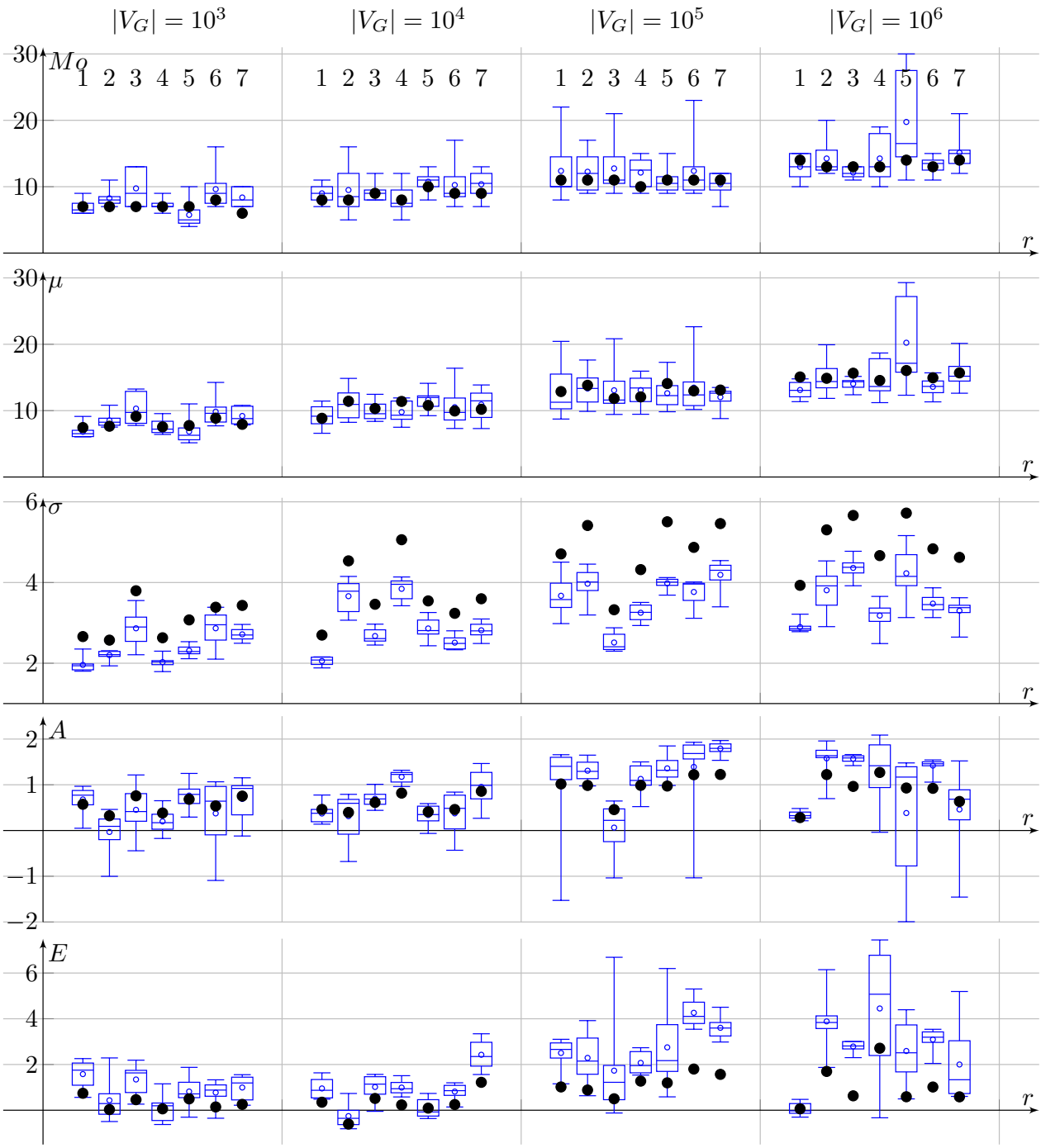

Fig. 6. Characteristics of full NND distribution in the quasi-pre-fractal trees $G_r^{\mathrm{PF},v}$ (black dots) and scattering of characteristics of estimates (boxplots)

Рис. 6. Характеристики полного РДП в квазипредфрактальных деревьях $G_r^{\mathrm{PF},v}$ (чёрные точки) и разброс характеристик его оценок (ящики с усами)

и эти величины растут с ростом $|V_G|$. Также необходимо отметить, что для большинства прогонов $E_i > A_i$ (для асимметрии A и эксцесса E полных РДП это выполняется реже).

Таким образом, для $|V_G| \in [10^8, 10^{10}]$ вершин оценки асимметрии и эксцесса для некоторых квазипредфрактальных деревьев будут приближаться к экспериментально полученным значениям (9).

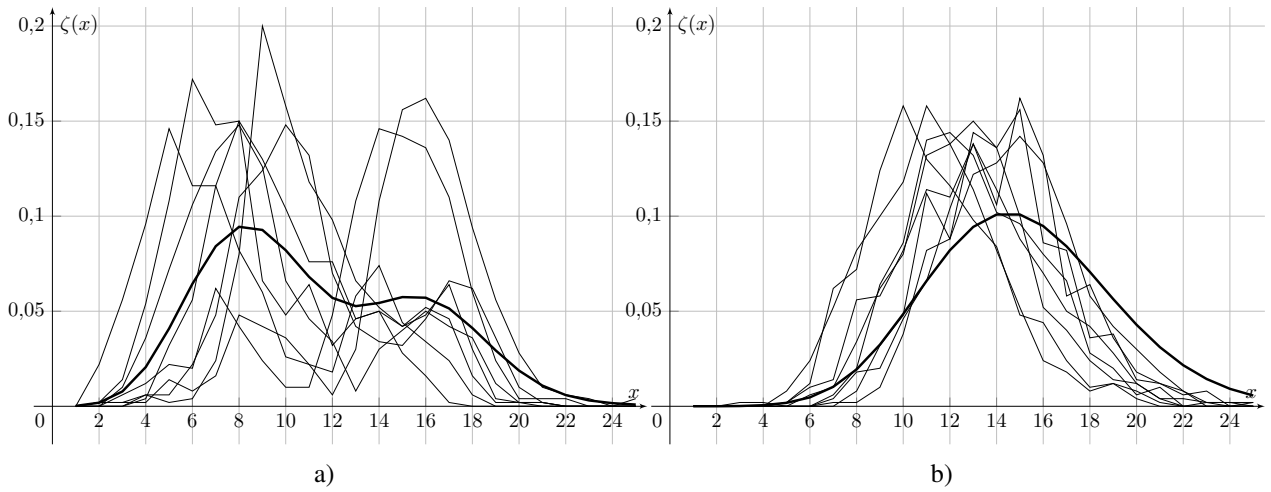

Fig. 7. Full NND distribution $\zeta(x)$ (thick line) and its estimates $\zeta_i(x)$ (thin lines) for: a) $G_2^{\mathrm{PF},4}$, b) $G_1^{\mathrm{PF},6}$

Рис. 7. Полное РДП $\zeta(x)$ (жирная линия) и его оценки $\zeta_i(x)$ (тонкие линии): a) для $G_2^{\mathrm{PF},4}$, b) для $G_1^{\mathrm{PF},6}$

Заключение

Для оценивания РДП сетевого уровня глобальной сети была разработана схема КУОУ, использующая инструментальные средства, стандартные для большинства дистрибутивов GNU/Linux, и проведена серия измерений, начатых в 2011 году; результаты измерений были обработаны с помощью GNU Octave. Полученные оценки РДП имеют высокие положительные значения асимметрии и эксцесса.

Чтобы установить возможную ошибку, вносимую таким оцениванием, было проведено имитационное моделирование измерений. Его результаты показали, что погрешность оценок характеристик РДП по КУОУ может составлять 100% и выше, так что полученные в результате измерений данные не могут рассматриваться как характеристики РДП сетевого уровня глобальной сети или их адекватные оценки.

Тем не менее, также имитационное моделирование показало, если рост сети описывается безмасштабной моделью Барабаши—Альберт, то при оценивании РДП по схеме КУОУ не могут быть получены настолько высокие значения асимметрии и эксцесса, как у полученных экспериментально оценок РДП. Таким образом, проведённые измерения в сочетании с имитационным моделированием позволяют уверенно утверждать, что сетевой уровень глобальной сети не описывается безмасштабной моделью Барабаши—Альберт.

Разработанная квазипредфрактальная модель роста в некотором приближении может быть использована для описания сетевого уровня глобальной сети. Имитационное моделирование показывает, что в квазипредфрактальном дереве, состоящем из одной крупнейшей подсети и множества более мелких, размер которого сопоставим с глобальной сетью, при оценивании РДП по КУОУ будут получены высокие значения асимметрии и эксцесса, сопоставимые с полученными экспериментально.

References

[1] A. Gorodilov, A. Kononova, and V. Shangin, "Osobennosti peredachi dannyh v decentralizovannyh piringovyh setyah", *Proceedings of Universities. Electronics*, vol. 98, no. 6, pp. 95–97, 2012.

[2] A. P. Shiryaev, A. V. Dorofeev, A. R. Fedorov, L. G. Gagarina, and V. V. Zaycev, "LDA models for finding trends in technical knowledge domain", in *2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, IEEE, 2017, pp. 551–554.

[3] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.

[4] A. Fronczak, P. Fronczak, and J. A. Hołyst, "Average path length in random networks", *Physical Review E*, vol. 70, no. 5, p. 056 110, 2004.

[5] R. Sedgewick, *Algorithms in C, Part 5: Graph Algorithms*, 2002.

[6] H. Jeong, B. Tombor, Z. N. Albert R.and Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks", *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.

[7] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks", *Reviews of modern physics*, vol. 74, no. 1, pp. 47–97, 2002.

[8] A. Gorodilov A.V.and Kononova, "Simulation as tool of error assessment of experimental measurement of path-lengths distribution in global network", *Sistemy komp'yuternoj matematiki i ih prilozheniya: materialy XX Mezhdunarodnoj nauchnoj konferencii*, vol. 1, pp. 34–41, 2019.

[9] R. A. Kochkarov, D. A. Pavlov, and D. A.-Z. Hubieva, "Fractal and preefactal graphs, basic definitions and symbols", *Scientific journal of KubSAU (Polythematic online scientific journal of Kuban State Agrarian University)*, no. 134, pp. 174–188, 2017.