ВЕСТНИК НАЦИОНАЛЬНОГО ИССЛЕДОВАТЕЛЬСКОГО ЯДЕРНОГО УНИВЕРСИТЕТА “МИФИ”, 2019, том 8, № 4, с. 380–394

ПРИКЛАДНАЯ МАТЕМАТИКА И ИНФОРМАТИКА

АППРОКСИМАЦИЯ ИЗМЕРЕНИЙ ИНТЕРНЕТ-ТРАФИКА В МАГИСТРАЛЬНОМ КАНАЛЕ СУММОЙ ЛОГНОРМАЛЬНЫХ РАСПРЕДЕЛЕНИЙ

© 2019 г. Валерий В. Иванов[1,2], Виктор В. Иванов[2,3], А. В. Крянев[3,2], И. И. Татаринов[4]

[1] *ЗАО “МПОТК “ТЕХНОКОМПЛЕКТ”, Дубна, 141981, Россия*

[2] *Объединенный институт ядерных исследований, Дубна, 141980, Россия*

[3] *Национальный исследовательский ядерный университет “МИФИ”, Москва, 115409, Россия*

[4] *АО “Лаборатория Касперского”, Москва, 123060, Россия*

Поступила в редакцию 20.06.2019 г.
После доработки 20.06.2019 г.
Принята к публикации 25.06.2019 г.

В настоящей работе представлены два численных подхода для аппроксимации измерений сетевого трафика, регистрируемого в магистральном канале, на основе традиционного метода наименьших квадратов (МНК) и коэффициента детерминации R^2. Для дополнительной оценки точности аппроксимации анализируемых данных логнормальным распределением в настоящей статье анализировалась динамика зависимости максимума интенсивности сетевого трафика от размера окна агрегации. В первом и втором подходах был достигнут высокий уровень соответствия данных наблюдения логнормальному закону. При этом, после включения в аппроксимирующую функцию дополнительных слагаемых, точность аппроксимации заметно возрастает. Показано, что зависимость коэффициента детерминации от размера окна агрегации для анализируемых сетевых пакетов позволяет контролировать точность аппроксимации данных наблюдения логнормальным законом.

Ключевые слова: сетевой трафик, интенсивность, аппроксимация, логнормальное распределение, окно агрегации, коэффициент детерминации

DOI: 10.1134/S2304487X19040047

1. ВВЕДЕНИЕ

В работе [1] нами было показано, что также, как для локальной сети среднего размера, при агрегировании измерений интернет-трафика в магистральном канале формируются статистические распределения, которые с высокой точностью аппроксимируются логнормальными распределениями [2, 3].

Анализ плотностей вероятности распределения (ПВР) скорости передачи сетевых пакетов, которые в отличие от ПВР скорости передачи данных (ПВР интенсивности сетевого трафика) не зависят от размера сетевых пакетов, показал, что рассматриваемые распределения также с высокой точностью согласуются с логнормальным законом. Важно отметить, что указанные распределения могут использоваться для оценки стабильности работы сетевого оборудования.

Кроме того, нами наблюдалось более сложное поведение сетевого трафика для сетевых пакетов, полученных в дуплексном режиме, и было показано, что рассматриваемое распределение с приемлемой точностью аппроксимируется суммой двух логнормальных распределений [1].

В работе рассмотрены два численных подхода для аппроксимации измерений сетевого трафика, регистрируемого в магистральном канале, на основе: 1) традиционного метода наименьших квадратов (МНК); 2) не столь известного широкой аудитории – коэффициента детерминации R^2 [4, 5], который в общем случае можно рассматривать как универсальную меру зависимости одной случайной величины от множества других.

Для дополнительной оценки точности аппроксимации анализируемых данных логнормальным распределением нами также анализировалась динамика зависимости максимума интенсивности сетевого трафика от размера окна агрегации.

2. АППРОКСИМАЦИЯ ИЗМЕРЕНИЙ СЕТЕВОГО ТРАФИКА МЕТОДОМ МНК

Экспериментальные данные ПВР интенсивности сетевого трафика, представленные набором данных $[x_i, y_i]$ (где x_i – скорость передачи данных, y_i – количество данных, переданных с заданной скоростью), аппроксимировались логнормальным распределением по формуле:

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

где x – переменная, σ и μ – параметры распределения, a – нормировочный множитель.

Для определения коэффициентов a, σ, μ распределения (1) использовался метод наименьших квадратов, т.е. подбирались такие коэффициенты a, σ, μ, чтобы выполнялось условие

$$S_{ols} = \sum_{k=1}^{N}[y_k - f(x_k)]^2 = \min. \quad (2)$$

Вычисление коэффициентов a, σ, μ выполнялось итерационно с использованием метода градиентного спуска [6] и его модификаций: на начальном этапе использовался метод наискорейшего спуска [7], на заключительном – метод оврагов [8].

Непосредственная реализация процедуры определения коэффициентов a, σ, μ состояла в следующем. Начальные коэффициенты a_0, σ_0, μ_0 выбирались произвольно, но таким образом, чтобы выполнялось равенство

$$\int_0^\infty f(x)dx = 1,$$

например,

$$a_0 = 1.0, \quad \sigma_0 = 1.0, \quad \mu_0 = \ln(x_{last}),$$

где x_{last} – координата x самой правой экспериментальной точки.

В дальнейшем, коэффициенты σ, μ вычислялись итерационно методом градиентного спуска как

$$\sigma_n = w_\sigma(\sigma_{n-1}),$$
$$\mu_n = w_\mu(\mu_{n-1}),$$

где w_σ, w_μ – функции градиентного спуска по коэффициентам σ, μ, соответственно, а коэффициент a определялся как

$$a_n = \frac{\sum_{k=1}^{N} y_i g(x_k, \sigma_n, \mu_n)}{\sum_{k=1}^{N}(g_k)^2},$$

где

$$g(x_k, \sigma_n, \mu_n) = \frac{1}{\sqrt{2\pi}\sigma_n x_k}\exp\left[-\frac{1}{2\sigma_n^2}(\ln x_k - \mu_n)^2\right]$$

и для заданных коэффициентов σ_n, μ_n величина $S_{ols}(\sigma_n, \mu_n, a_n)$ (2) минимальна, т.е., для того, чтобы

$$S_{ols}(a) = \sum_{k=1}^{N}[y_k - ag(x_k)]^2 = \min,$$

требуется выполнение равенства

$$S'_{ols}(a) = 0.$$

Откуда

$$a = \frac{\sum_{k=1}^{N} y_k g(x_k)}{\sum_{k=1}^{N}(g_k)^2}.$$

Каждая итерация вычисления коэффициентов a_n, σ_n, μ_n включала следующие шаги:

1. Для определения коэффициента σ_n:

a. на основании вычисленного на предыдущей итерации коэффициента σ_{n-1} выбирались коэффициенты:

$$\sigma_{n_1} = \sigma_{n-1}(1 - \Delta_\sigma),$$
$$\sigma_{n_2} = \sigma_{n-1},$$
$$\sigma_{n_3} = \sigma_{n-1}(1 + \Delta_\sigma),$$

b. где Δ_σ – шаг спуска (максимальный шаг приближения σ_n к искомому значению σ_{min}) и вычислялись коэффициенты МНК:

$$S_{n_1} = S_{ols}(\sigma_{n_1}, \mu),$$
$$S_{n_2} = S_{ols}(\sigma_{n_2}, \mu),$$
$$S_{n_3} = S_{ols}(\sigma_{n_3}, \mu),$$

c. предполагая, что точки $\{\sigma_{n_1}, S_{n_1}\}$, $\{\sigma_{n_2}, S_{n_2}\}$, $\{\sigma_{n_3}, S_{n_3}\}$ лежат на некоторой кривой $S_{ols}(\sigma) = p_1\sigma^2 + p_2\sigma + p_3$, определялся коэффициент $\sigma_{n_{min}}$, при котором указанная кривая достигала бы минимума как

$$\sigma_{n_{min}} = \frac{1}{2}\times$$
$$\times\frac{\sigma_{n_1}^2(S_{n_2} - S_{n_3}) + \sigma_{n_2}^2(S_{n_3} - S_{n_1}) + \sigma_{n_3}^2(S_{n_1} - S_{n_2})}{\sigma_{n_1}(S_{n_2} - S_{n_3}) + \sigma_{n_2}(S_{n_3} - S_{n_1}) + \sigma_{n_3}(S_{n_1} - S_{n_2})},$$

d. представляющее собой решение уравнения:

$$S'_{ols}(\sigma) = 0,$$

$$\sigma_{n_{\min}} = -\frac{p_2}{2p_1},$$

e. с коэффициентами, получаемыми как решение системы уравнений:

$$S_{n_i} = p_1\sigma_{n_i}^2 + p_2\sigma_{n_i} + p_3, \quad i = 1, 2, 3,$$

f. после чего вычислялось соответствующее знач

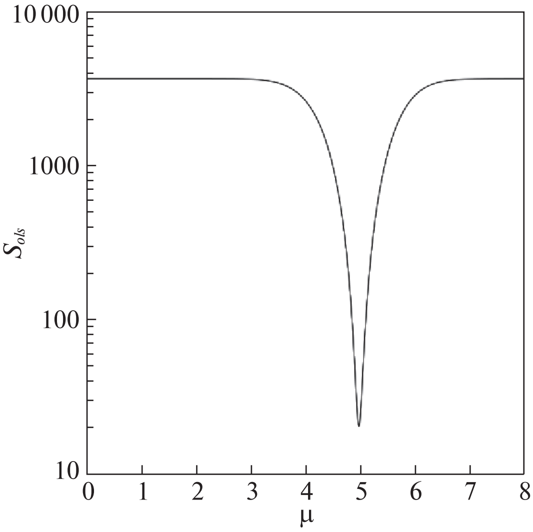

Рис. 1. Зависимость коэффициента МНК S_{ols} от коэффициента μ (один глобальный минимум при большом шаге Δμ).

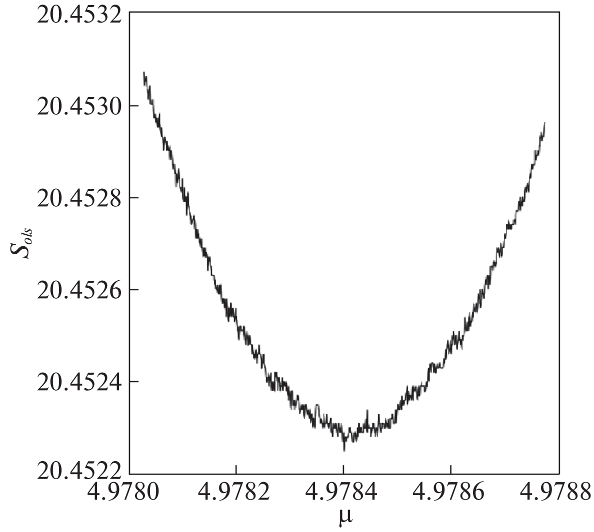

Рис. 2. Зависимость коэффициента МНК S_{ols} от коэффициента μ (множество локальных минимумов при малом шаге Δμ).

ение коэффициента МНК $S_{n_{\min}} = S_{ols}(\sigma_{n_{\min}}, \mu)$,

g. из набора коэффициентов σ_{n_1}, σ_{n_2}, σ_{n_3}, $\sigma_{n_{min}}$ выбирался такой σ_n, для которого значение коэффициента МНК $S_{ols}(\sigma_n)$ было минимальным;

2. аналогичная процедура проводилась при определении коэффициента μ_n;

3. проверялось условие прекращения поиска коэффициентов a, σ, μ:

a. если вычисленный коэффициент МНК $S_n < S_{n-1} - \varepsilon$, где ε – некоторая заданная ошибка для определения коэффициента МНК $S_{\min}$, то выполнялась $(n+1)$-я итерация с теми же значениями шагов спуска Δ_σ, Δ_μ, что и для n-ой итерации;

b. если вычисленный коэффициент МНК удовлетворял следующему условию $S_n \geq S_{n-1} - \varepsilon$, значения шагов спуска корректировались согласно

$$\Delta_\sigma = \Delta_\sigma r_\sigma,$$

$$\Delta_\mu = \Delta_\mu r_\mu;$$

c. и выполнялась $(n+1)$-я итерация с новыми значениями шагов спуска Δ_σ, Δ_μ;

d. если последние L итераций шаг спуска уменьшался, то поиск коэффициентов a, σ, μ прекращался, а коэффициенты a_n, σ_n, μ_n, полученные на последней итерации, считались искомыми.

Величины шагов спуска Δ_σ, Δ_μ и количество итераций L подбирались индивидуально. Для рассмотренной нами задачи был использован следующий набор параметров:

$$\Delta_\sigma = \Delta_\mu = 0.1,$$

$$r_\sigma = r_\mu = 1.01,$$

$$L = 128.$$

Преимуществом метода градиентного спуска при поиске минимума функции $f(x)$ является быстрое схождение x_n к $x_{\min} \pm \varepsilon$ (см. рис. 1). Однако при этом с повышением точности метода (т.е. при уменьшении величины ε в результате уменьшения шага спуска Δx) x_n будет стремиться к одному из локальных минимумов, не совпадающему с искомым глобальным минимумом (см. рис. 2) (так называемая “проблема оврагов”).

Для ускорения процесса нахождения искомых коэффициентов σ_n и μ_n, соответственно, к $\sigma_{\min}$ и $\mu_{\min}$ применялось постепенное уменьшение шагов спуска Δσ и Δμ.

Для преодоления локальных минимумов применялось прогнозирование поведения кривой (например, в виде параболы), на которой лежат значения соответствующих коэффициентов, и выбор такого коэффициента, при котором указанная кривая обеспечивает минимум.

Дополнительно, для поиска наиболее близких значений коэффициентов σ_n к $\sigma_{\min}$ и μ_n к $\mu_{\min}$, анализировалось поведение коэффициента МНК S_{ols} при разных начальных значениях коэффициентов a_0, σ_0, μ_0. А именно, вычислялось несколько коэффициентов МНК S_{ols_i} для разных значений μ_{0_i}

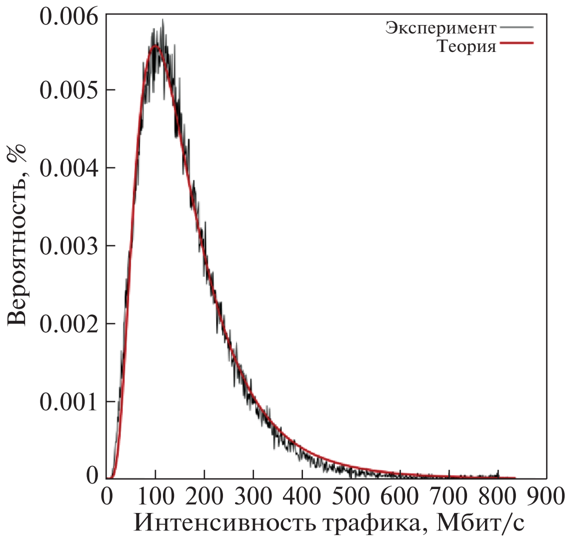

Рис. 3. ПВР интенсивности сетевого трафика, аппроксимированного логнормальным распределением (1).

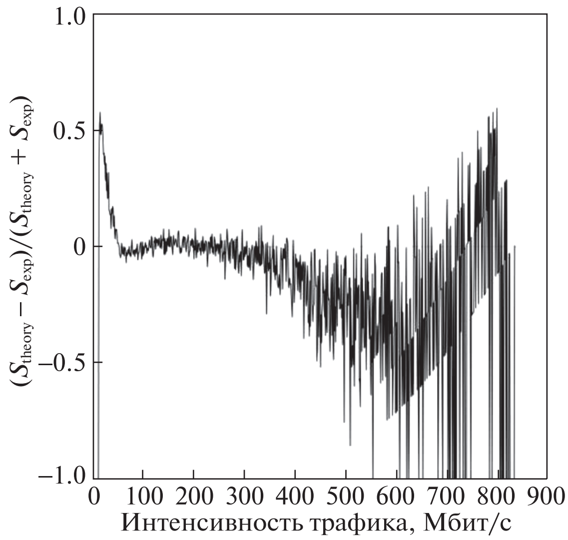

Рис. 4. Поведение случайной величины R_s (3), описывающей степень соответствия ПВР интенсивности сетевого трафика логнормальному закону (1).

и выбирались такие a, σ, μ, для которых коэффициент МНК S_{ols} был минимальным.

В результате применения описанного алгоритма достигалась высокая точность аппроксимации ПВР интенсивности сетевого трафика логнормальным распределением (1).

На рис. 3 представлен результат аппроксимации ПВР интенсивности сетевого трафика, полученного в симплексном режиме, логнормальным распределением (1). На рис. 4 показано поведение случайной величины

$$R_s = \frac{S_{theory} - S_{\exp}}{S_{theory} + S_{\exp}}, \quad (3)$$

равной отношению разности между значением ПВР интенсивности сетевого трафика $S_{\exp}$ и аппроксимирующим его логнормальным распределением S_{theory} на их сумму $(S_{theory} + S_{\exp})$. Переменная R_s представляет собой относительную ошибку, характеризующую точность соответстия между измерениями сетевого трафика и аппроксимирующей их функцией.

Из рис. 3 и рис. 4 видно, что, несмотря на достаточно хорошее приближение экспериментальных данных аппроксимирующей кривой (1), между ними наблюдается заметное расхождение в области значений интенсивности сетевого трафика 300–600 Мбит/с. Причины указанного расхождения и то, как их можно уменьшить будут рассмотрены ниже.

3. АППРОКСИМАЦИЯ РАСПРЕДЕЛЕНИЙ С НЕСКОЛЬКИМИ ПИКАМИ

Как отмечалось выше, в работе нами наблюдалось более сложное поведение ПВР интенсивности сетевого трафика, которое, в случае присутствия нескольких четко выраженных пиков, можно аппроксимировать суммой нескольких логнормальных распределений.

На рис. 5 приведен такой пример, который был получен, в частности, при работе сетевого оборудования в дуплексном режиме (анализировались как получаемые, так и передаваемые сетевые пакеты).

Экспериментальные данные мультипикового ПВР интенсивности сетевого трафика аппроксимировались суммой нескольких логнормальных распределений по формуле:

$$f(x) = \sum_{i=1}^{M} \left\{ \frac{a_i}{\sqrt{2\pi}\sigma_i x} \exp\left[-\frac{1}{2\sigma_i^2} (\ln x - \mu_i)^2 \right] \right\}, \quad (4)$$

где M – количество логнормальных распределений, каждое из которых описывается своими коэффициентами $\{a_i, \sigma_i, \mu_i\}$, и суммой которых аппроксимируется наблюдаемое распределение экспериментальных данных.

Для определение коэффициентов $\{a_i, \sigma_i, \mu_i\}$ логнормального распределения (4) здесь также использовался метод наименьших квадратов, т.е. подбирались такие коэффициенты $\{a_i, \sigma_i, \mu_i\}$, чтобы выполнялось условие (2).

Вычисление коэффициентов $\{a_i, \sigma_i, \mu_i\}$ выполнялось итерационно с использованием метода градиентного спуска способом, аналогичным

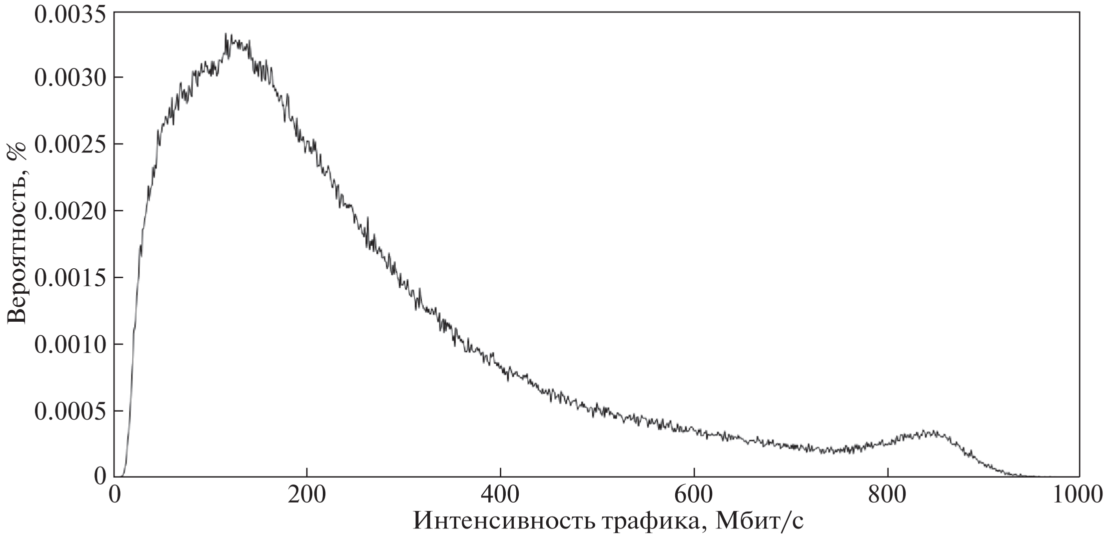

Рис. 5. Двухпиковое ПВР интенсивности сетевого трафика.

описанному выше. Непосредственно алгоритм определения коэффициентов $\{a_i, \sigma_i, \mu_i\}$ состоял в следующем.

Начальные значения коэффициентов a_{i0}, σ_{i0}, μ_{i0} выбирались произвольно с условием, чтобы выполнялось равенство

$$\int_0^\infty \sum_{i=1}^{M} \frac{a_{i0}}{\sqrt{2\pi}\sigma_{i0} x} e^{-\frac{(\ln x - \mu_{i0})^2}{2\sigma_{i0}^2}} dx = 1,$$

откуда

$$\sum_{i=1}^{M} a_{0i} = 1,$$

например,

$$a_0 = \frac{1}{M}, \quad \sigma_0 = 1.0, \quad \mu_0 = \ln(x_{last}),$$

где x_{last} – это координата самой правой экспериментальной точки.

В дальнейшем коэффициенты $\{\sigma_i, \mu_i\}$ вычислялись итерационно методом градиентного спуска как

$$\sigma_{in} = w_\sigma(\sigma_{in-1}),$$
$$\mu_{in} = w_\mu(\mu_{in-1}),$$

а коэффициент a_i определялся как

$$a_{in} = \frac{\sum_{k=1}^{N} g(x_k, \sigma_n, \mu_n)(y_i - h(x_k))}{\sum_{k=1}^{N} (g_k)^2},$$

где

$$g(x_k, \sigma_n, \mu_n) = \frac{1}{\sqrt{2\pi}\sigma_n x_k} \exp\left[-\frac{1}{2\sigma_n^2}(\ln x_k - \mu_n)^2\right],$$

$$h(x_k) = \sum_{j \neq i}^{M} \frac{a_j}{\sqrt{2\pi}\sigma_j x_k} \exp\left[-\frac{1}{2\sigma_j^2}(\ln x_k - \mu_j)^2\right].$$

Таким образом, чтобы для заданных коэффициентов σ_{in}, μ_{in} величина коэффициента МНК $S_{ols}(\sigma_{in}, \mu_{in}, a_{in})$ была минимальной.

Имеем

$$S_{ols}(a) = \sum_{k=1}^{N} (y_k - f(x_k))^2 = \min,$$

где

$$f(x_k) = ag(x_k) + h(x_k).$$

Т.е.

$$S'_{ols}(a) = 0,$$

откуда имеем

$$a_i = \frac{\sum_{k=1}^{N} g(x_k)(y_k - h_i(x_k)))}{\sum_{k=1}^{N} (g_k)^2}.$$

Каждая итерация вычисления коэффициентов a_{in}, σ_{in}, μ_{in} включала следующие шаги:

1. Все коэффициенты $\sigma_i, i \in [1, M]$ определялись согласно процедуре, отвечающей шагу 1 алгоритма, описанного в разделе 2;

2. Все коэффициенты $\mu_i, i \in [1, M]$ определялись согласно процедуре, отвечающей шагу 2 алгоритма, описанного в разделе 2;

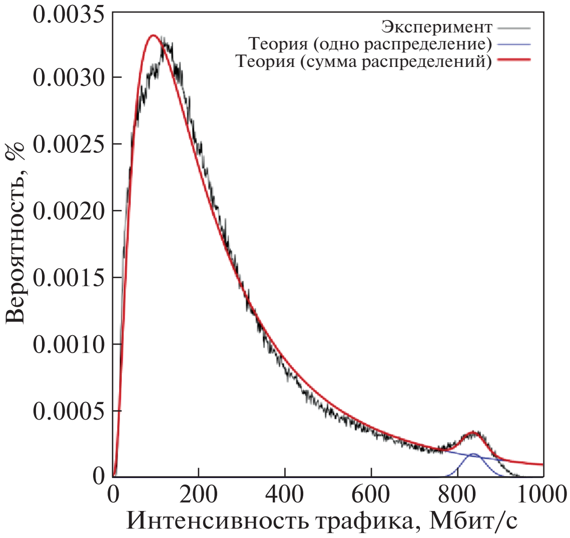

Рис. 6. ПВР интенсивности сетевого трафика, аппроксимированное сумой логнормальных распределений (4).

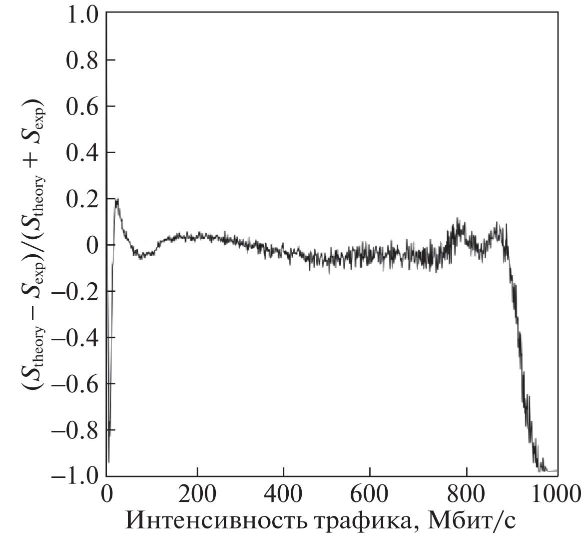

Рис. 7. Поведение случайной величины R_s (3), описывающей степень соответствия ПВР интенсивности сетевого трафика сумме логнормальных распределений (4).

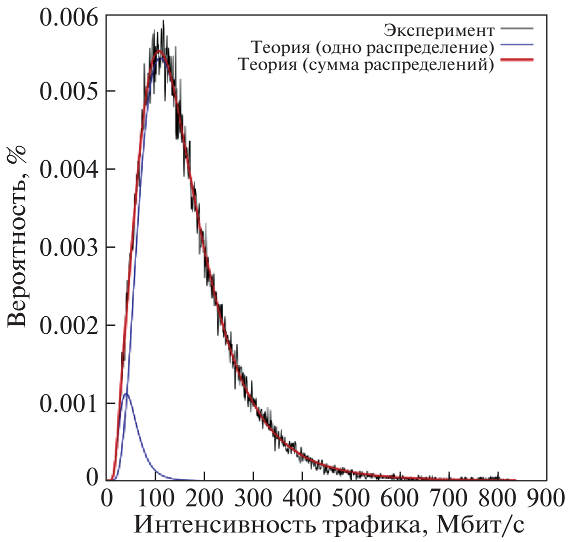

Рис. 8. ПВР интенсивности сетевого трафика, аппроксимируемое суммой двух логнормальных распределений (4).

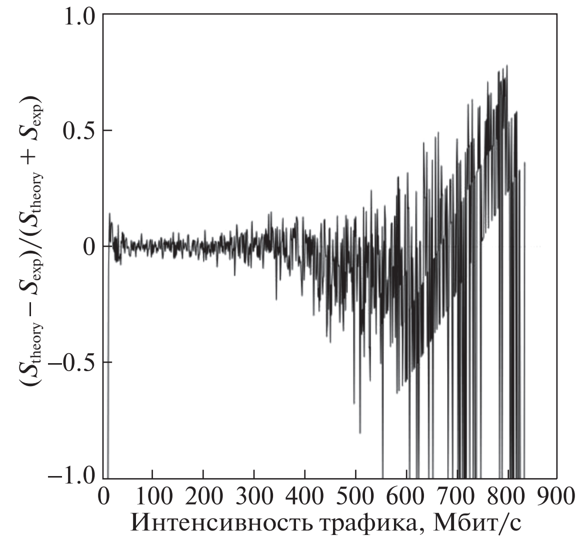

Рис. 9. Поведение случайной величины R_s (3), описывающей степень соответствия ПВР интенсивности сетевого трафика сумме двух логнормальных распределений (4).

3. Проверялось условие прекращения поиска коэффициентов $\{a_i, \sigma_i, \mu_i\}$ способом, описанным в шаге 3 алгоритма, приведенного в разделе 2.

Применение рассмотренного алгоритма к измерениям сетевого трафика, приведенным на рис. 5 позволяет добиться высокой точности аппроксимации ПВР интенсивности сетевого трафика суммой логнормальных распределений (4) (см. рис. 6 и 7).

Следует отметить, что применив к однопиковому распределению, приведенному на рис. 3, аппроксимацию суммой двух логнормальных распределений, используя формулу (4), можно заметно улучшить соответствие реальных измерений теоретической кривой (см. рис. 8 и 9).

В примере с двухпиковым ПВР интенсивности сетевого трафика (см. рис. 6) наблюдаются отклонения аппроксимирующей кривой (4) от измерений сетевого трафика в области главного пика

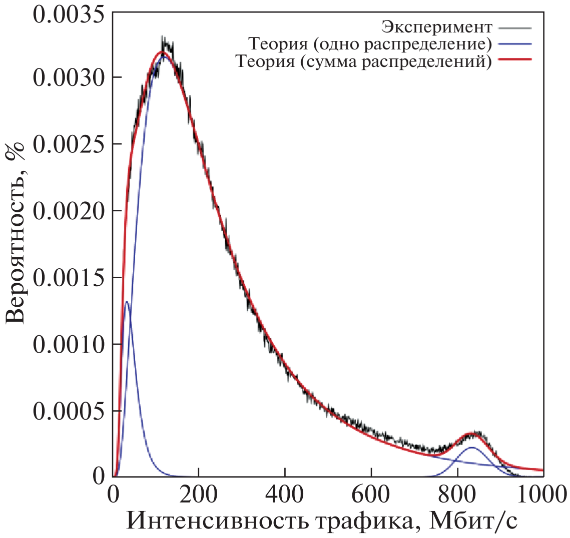

Рис. 10. ПВР интенсивности сетевого трафика, аппроксимированное суммой 3 логнормальных распределений.

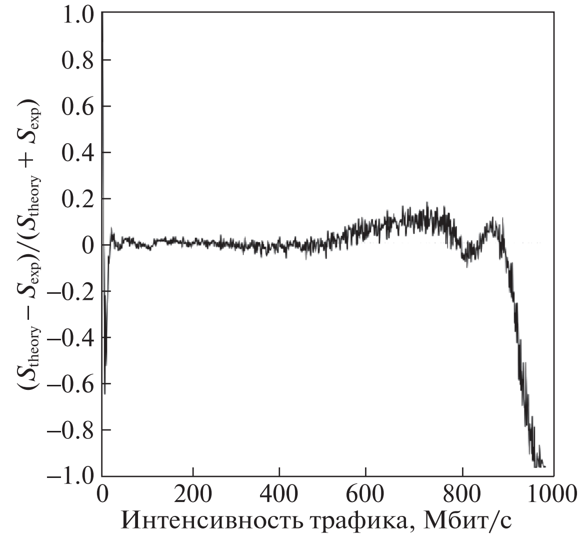

Рис. 11. Поведение случайной величины R_s (3), описывающей степень соответствия ПВР интенсивности сетевого трафика сумме 3 логнормальных распределений (4).

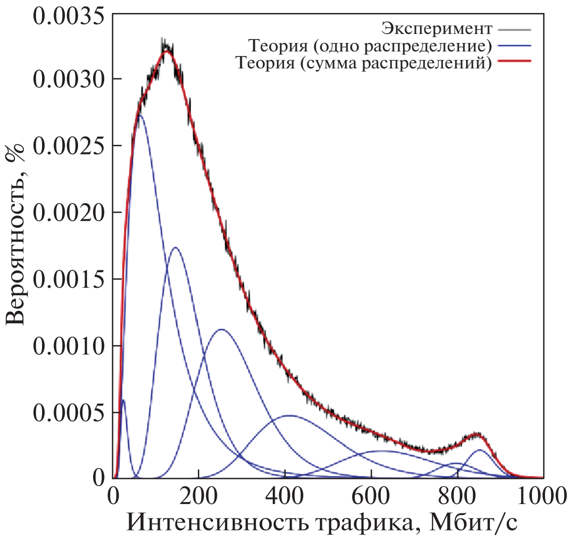

Рис. 12. ПВР интенсивности сетевого трафика, аппроксимированное суммой 8 логнормальных распределений.

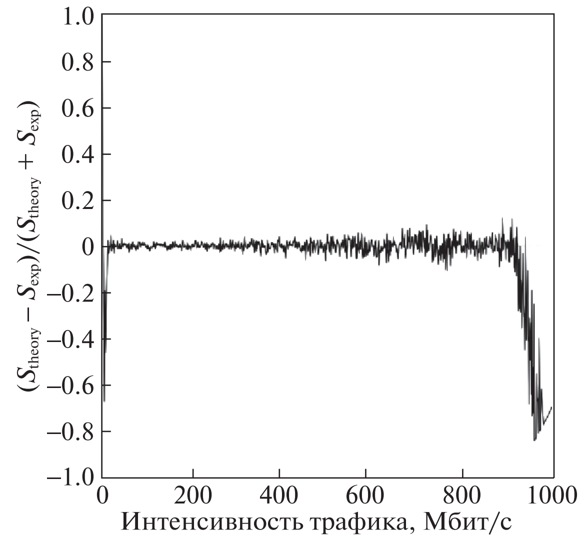

Рис. 13. Поведение случайной величины R_s (3), описывающей степень соответствия ПВР интенсивности сетевого трафика сумме 8 логнормальных распределений (4).

(~100 Мбит/с), при значениях интенсивности сетевого трафика в интервале 300–600 Мбит/с, а также в области правого хвоста распределения ПНР, где экспериментальные значения на “хвосте” спадают очень резко, а теоретическая кривая ведет себя более плавно.

Аппроксимация двухпикового распределения тремя логнормальными распределениями (рис. 10) позволяет заметно улучшить указанную ситуацию (см. рис. 6).

Аппроксимация рассмотренного нами двухпикового распределения суммой восьми логнормальных распределений (рис. 12) позволяет избавиться от всех указанных выше проблем (рис. 6), в том числе, и от медленного спада “хвоста” ПВР интенсивности сетевого трафика.

Интересно отметить, что описанный нами метод градиентного спуска для вычисления параметров логнормального распределения (или суммы логнормальных распределений) несмотря на

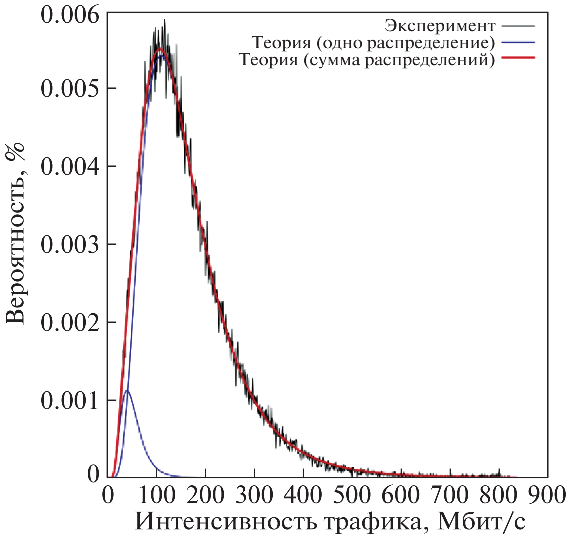

Рис. 14. ПВР интенсивности сетевого трафика, аппроксимированное суммой 2 логнормальных распределений (4).

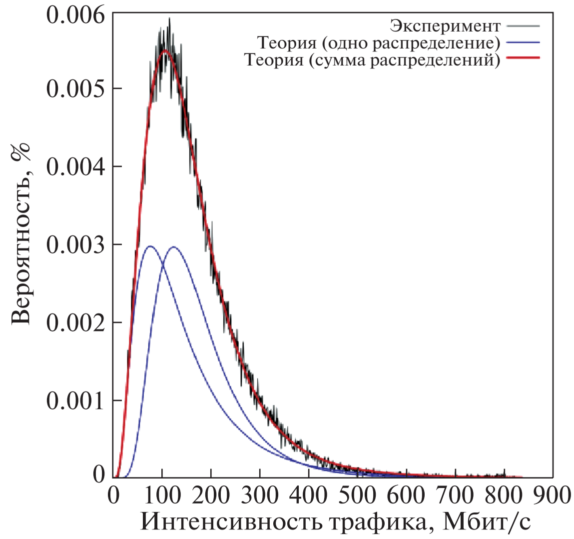

Рис. 15. ПВР интенсивности сетевого трафика, аппроксимированное суммой 2 логнормальных распределений (4).

проблемы, обусловленные наличием локальных минимумов (проблема оврагов), позволяет добиться высокой точности аппроксимации анализируемых распределений. К сожалению, получаемый при этом результат иногда может отвечать разным логнормальным распределениям. Подобный пример представлен нами на рис. 14 и 15.

В этом примере коэффициенты МНК (S_1 и S_2) в обоих случаях отличаются незначительно:

$$S_1 = 0.11209,$$

$$S_2 = 0.11226,$$

$$\frac{S_1 - S_2}{S_1 + S_2} \sim 0.0007.$$

Однако при этом сами пики отличаются уже более существенно (как по интенсивности, так и по положению):

$$x_{11} = 40.67002, \quad a_{11} = 0.05563,$$

$$x_{12} = 110.54602, \quad a_{12} = 0.94571$$

$$x_{11} = 78.02385, \quad a_{11} = 0.50198,$$

$$x_{12} = 125.39898, \quad a_{12} = 0.50198.$$

Отмеченное обстоятельство может представлять определенную сложность при интерпретации анализируемых данных, поскольку каждое логнормальное распределение отвечает определенному режиму передачи данных.

4. АППРОКСИМАЦИЯ ИЗМЕРЕНИЙ СЕТЕВОГО ТРАФИКА НА ОСНОВЕ КОЭФФИЦИЕНТА ДЕТЕРМИНАЦИИ R^2

Определенным ограничением коэффициента МНК S_{ols} (2), используемого для оценки точности аппроксимации, является то, что его величина зависит от разных факторов, например, количества аппроксимируемых точек $[x_i, y_i]$, значения величин $[y_i]$ и т.д. Без учета этих факторов знание величины коэффициента МНК не дает никакой информации о точности аппроксимации анализируемых данных.

Для оценки точности желательно использовать критерий, численное значение которого ограничено в некотором диапазоне, где одна из границ соответствовала бы грубой аппроксимации, а другая – наиболее точной, при которой все экспериментальные точки ложились бы на некоторую теоретическую кривую.

С этой целью мы предлагаем использовать для оценки точности аппроксимации данных наблюдения в качестве критерия коэффициент детерминации R^2 (англ. coefficient of determination R^2).

Коэффициент детерминации R^2 – это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными. Коэффициент R^2 рассматривают как универсальную меру зависимости одной случайной величины от множества других:

$$R^2 = 1 - \frac{V(y|x)}{V(y)} = 1 - \frac{\sigma^2}{\sigma_y^2}, \tag{5}$$

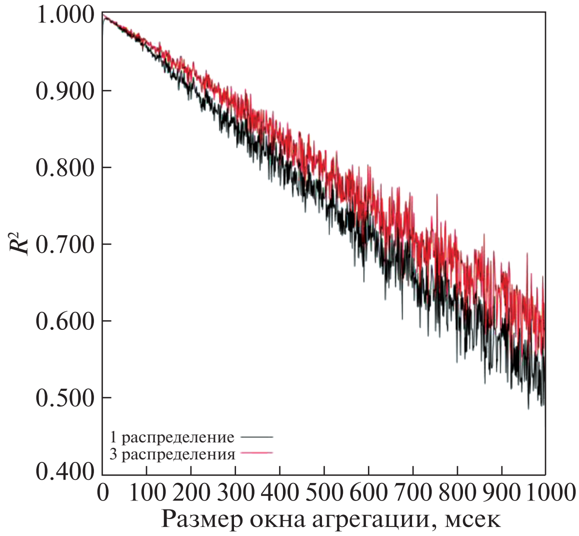

Рис. 16. Зависимость коэффициента R^2 от размера окна агрегации для ПВР интенсивности сетевого трафика, в котором учитывались только TCP пакеты, полученные в дуплексном режиме.

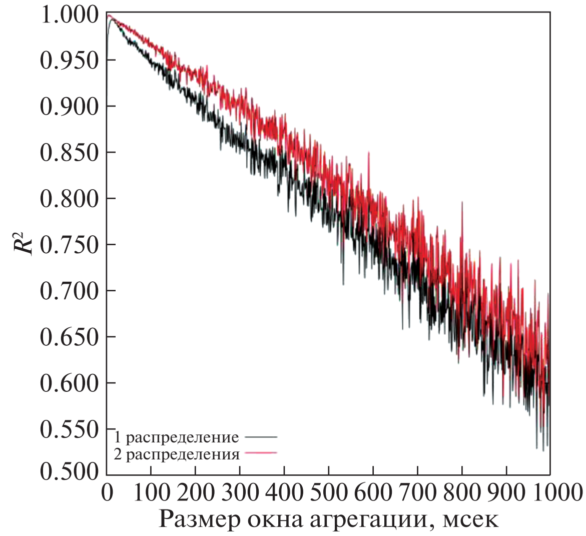

Рис. 17. Зависимость коэффициента R^2 от размера окна агрегации для ПВР интенсивности сетевого трафика, в котором учитывались только TCP пакеты, полученные в симплексном режиме.

где $V(y|x) = \sigma^2$ – условная (по факторам x) дисперсия зависимой переменной (дисперсия случайной ошибки модели).

Если используется выборочная оценка значений соответствующих дисперсий, то из (5) следует формула для выборочного коэффициента детерминации (он обычно и подразумевается под коэффициентом детерминации):

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}, \tag{6}$$

где SS_{res} – сумма квадратов остатков регрессии:

$$SS_{res} = \sum_{k=1}^{N}(y_i - \hat{y}_k)^2 = \sum_{k=1}^{N}[y_k - f(x_k)]^2 = S_{ols},$$

$$SS_{tot} = \sum_{k=1}^{N}(y_i - \bar{y}_k)^2.$$

Откуда получаем выражение коэффициента детерминации R^2 для оценки точности аппроксимации экспериментальных данных определенной функциональной зависимостью:

$$R^2 = 1 - \frac{\sum_{k=1}^{N}[y_k - f(x_k)]^2}{\sum_{k=1}^{N}(y_k - \overline{y_k})^2} = 1 - \frac{S_{ols}}{\sum_{k=1}^{N}(y_k - \overline{y_k})^2}.$$

Коэффициент детерминации R^2 принимает значения от 0 до 1, чем ближе его значение к 1, тем точнее аппроксимация экспериментальных данных теоретической кривой.

Коэффициент R^2 применялся нами для оценки точности аппроксимации ПВР сетевого трафика и ПВР скорости передачи сетевых пакетов, построенных по экспериментальным данным при разных размерах окна агрегации, логнормальным законом: одним, или суммой нескольких логнормальных распределений.

На рис. 16 и рис. 17 приведены графики зависимости коэффициента детерминации R^2 от размера окна агрегации для ПВР интенсивности сетевого трафика, в котором учитывались только TCP пакеты, полученные, соответственно, в дуплексном и симплексном режимах. Для количественной и качественной оценки R^2, как инструмента точности аппроксимации, сравнивались значения R^2, рассчитанные при аппроксимации экспериментальных данных одним и суммой нескольких логнормальных распределений.

• На рис. 16 показаны графики при аппроксимации ПВР интенсивности сетевого трафика одним логнормальным распределением и суммой трех логнормальных распределений. Поскольку ПВР интенсивности сетевого трафика, полученного в дуплексном режиме, содержит два ярко выраженных пика интенсивности (см. рис. 5), то, в отличие от графика, приведенного на рис. 17, для аппроксимации использовалась сумма трех, а не двух логнормальных распределений. При этом, если сумма двух логнормальных распределений обеспечивает высокий уровень аппроксимации двухпиковой ПВР интенсивности сетевого

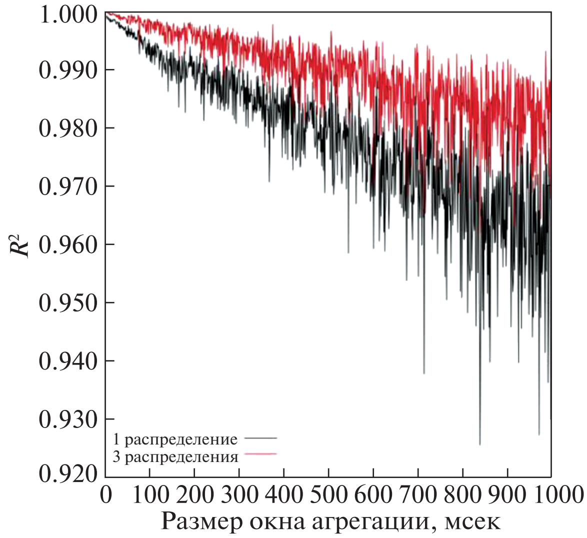

Рис. 18. Зависимость коэффициента R^2 от размера окна агрегации для ПВР скорости передачи сетевых пакетов, в котором учитывались только TCP пакеты, полученные в дуплексном режиме.

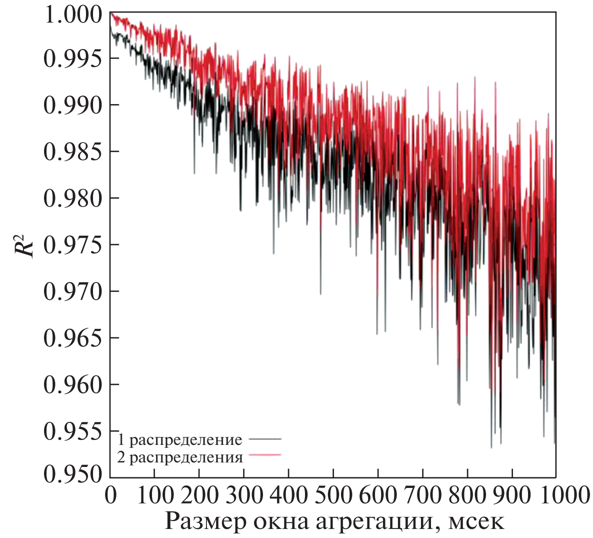

Рис. 19. Зависимость коэффициента R^2 от размера окна агрегации для ПВР скорости передачи сетевых пакетов, в котором учитывались только TCP пакеты, полученные в симплексном режиме.

трафика, то дополнительное (третье) слагаемое позволяет повысить точность аппроксимации, достигнутую (в первом приближении) суммой двух первых распределений.

• На рис. 17 показаны графики для случая аппроксимации ПВР интенсивности сетевого трафика одним логнормальным распределением и суммой двух логнормальных распределений. При этом, поскольку ПВР интенсивности сетевого трафика, полученного в симплексном режиме, содержит только один ярко выраженный пик интенсивности (см. рис. 3), то одно логнормальное распределение аппроксимирует ПВР интенсивности сетевого трафика в первом приближении, а второе распределение уточняет аппроксимацию первого логнормального распределения.

Видно, что, чем больше логнормальных распределений используется для аппроксимации ПВР интенсивности сетевого трафика, тем сильнее коэффициент детерминации R^2 стремится к 1. Такая зависимость очевидна, т.к. чем больше логнормальных распределений используется для аппроксимации ПВР интенсивности сетевого трафика, тем меньше должно быть отклонение случайной величины R_s (3) от 1. Исключения составляют начало и конец указанных распределений.

Из приведенных на рис. 16 и рис. 17 графиков видно, что в области малых значений окна агрегации (от 0 мсек до 50 мсек) коэффициент детерминации R^2, начиная с величины ≈ 0.95, быстро растет, достигая максимального значения ~ 1, после чего следует медленный спад. При этом величина коэффициента R^2 длительное время остается в допустимой области, отвечающей высокой точности аппроксимации анализирумых данных логнормальным распределением.

Уменьшение коэффициента детерминации R^2 с ростом размера окна агрегации вызвано падением точности аппроксимации данных логнормальным законом, а рост ширины коридора разброса обусловлен увеличением погрешности при вычислении ПВР интенсивности сетевого трафика из-за уменьшения числа регистрируемых сетевых пакетов.

Для анализа данных сетевого трафика, накопленных за некоторый промежуток времени на сайте [10], с помощью пакета Wireshark [11] извлекалась и использовалась следующая информация: время прихода сетевого пакета t_i, размер сетевого пакета s_i (заголовок сетевого пакета и данные, переданные сетевым пакетом) о совокупности сетевых пакетов $\{p_i(t, s)\}$.

На рис. 18 и 19 приведены графики зависимости коэффициента детерминации R^2 от размера окна агрегации для ПВР скорости передачи TCP-пакетов (без учета содержащейся в них информации), полученные, соответственно, в дуплексном и симплексном режимах.

Из приведенных на рис. 18, 19 графиков можно сделать вывод о том, что их поведение в целом повторяет то, что мы видели на рис. 16, 17. Видно, что коэффициент R^2 в случае аппроксимации ПВР скорости передачи сетевых пакетов (рис. 18, 19)

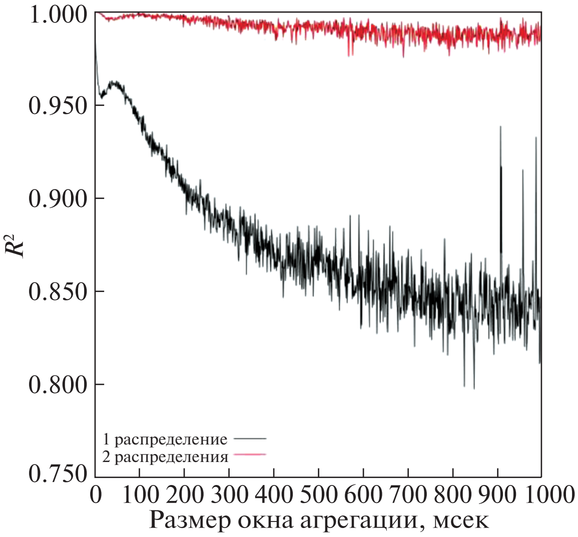

Рис. 20. Зависимость коэффициента R^2 от размера окна агрегации для ПВР интенсивности сетевого трафика, в котором учитывались только служебные пакеты, полученные в дуплексном режиме.

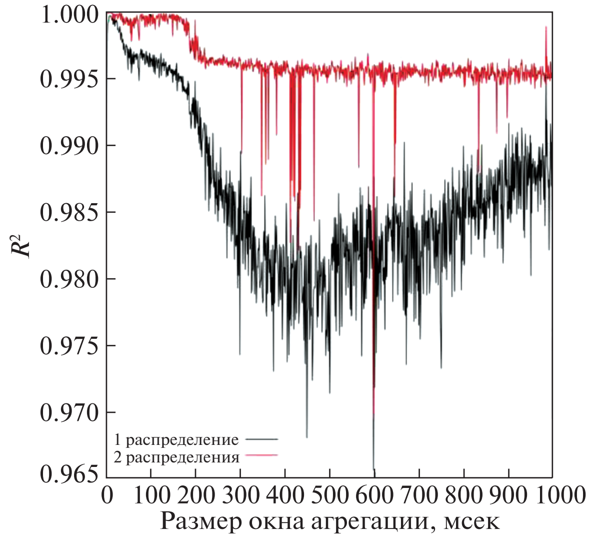

Рис. 21. Зависимость коэффициента R^2 от размера окна агрегации для ПВР скорости передачи сетевых пакетов, в котором учитывались только служебные пакеты, полученные в дуплексном режиме.

логнормальным законом заметно ближе к 1, чем для случая ПВР сетевого трафика (рис. 16, 17). Т.е. аппроксимация логнормальным распределением ПВР скорости передачи сетевых пакетов оказывается точнее в сравнении с ПВР интенсивностью сетевого трафика. Это связано с тем, что в отличие от интенсивности сетевого трафика, который зависит как от времени передачи сетевых пакетов, так и от объема данных, передаваемых в сетевых пакетах, скорость передачи сетевых пакетов зависит только от времени их передачи. Поэтому на ПВР скорости передачи сетевых пакетов влияют меньше случайных параметров.

С учетом вышеизложенного, можно сделать вывод о том, что ПВР скорости передачи сетевых пакетов может служить простым и удобным критерием для оценки работы сетевого оборудования без использования дополнительного, ресурсозатратного сбора и анализа сетевых данных.

Для определения значений коэффициента детерминации R^2, начиная с которых можно говорить о хорошей точности аппроксимации экспериментальных данных логнормальным законом, нами был проведен детальный анализ зависимостей коэффициента детерминации R^2 от размера окна агрегации для:

1. ПВР интенсивности сетевого трафика, измеренных в дуплексном и симплексном режимах, в которых анализировались:

– все сетевые пакеты,

– только TCP-пакеты,

– только служебные сетевые пакеты.

2. ПВР скорости передачи сетевых пакетов, измеренных в дуплексном и симплексном режимах, в которых анализировались:

– все сетевые пакеты,

– только TCP-пакеты,

– только служебные сетевые пакеты.

В результате проведенного анализа было установлено, что при $R^2 > 0.75$ точность аппроксимации может считаться допустимой, для $R^2 > 0.90$ – высокой, а в случае $R^2 > 0.95$ – близкой к идеальной. При этом графики для ПВР интенсивности сетевого трафика и ПВР скорости передачи сетевых пакетов, полученные для TCP-пакетов и для всех сетевых пакетов, зарегистрированных в дуплексном и симплексном режимах, повторяют поведение графиков, наблюденное нами ранее: смотри, в частности, рис. 16, 17.

Несколько иное поведение графиков наблюдается в случае обработки только служебных сетевых пакетов. На рис. 20, 21 приведены зависимости коэффициента R^2 от размера окна агрегации для ПВР интенсивности сетевого трафика и ПВР скорости передачи сетевых пакетов, полученные при обработке только служебных пакетов, зарегистрированных в дуплексном режиме.

На рис. 22, 23 показано поведение соответствующих графиков для служебных пакетов, полученных в симплексном режиме.

Из приведенных графиков видно, что для обоих типов передачи данных с ростом размера окна агрегации заметно растет разброс ошибок, при-

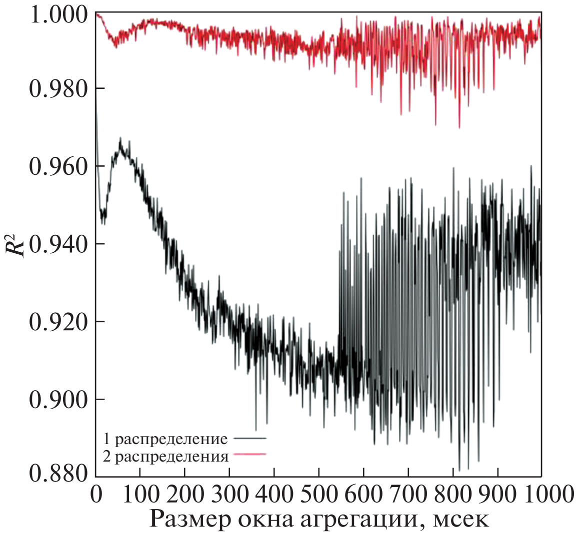

Рис. 22. Зависимость коэффициента R^2 от размера окна агрегации для ПВР интен-сивности сетевого трафика, в котором учитывались только служебные пакеты, полученные в симплексном режиме (mac: ”00 0e 39 e3 34 00”).

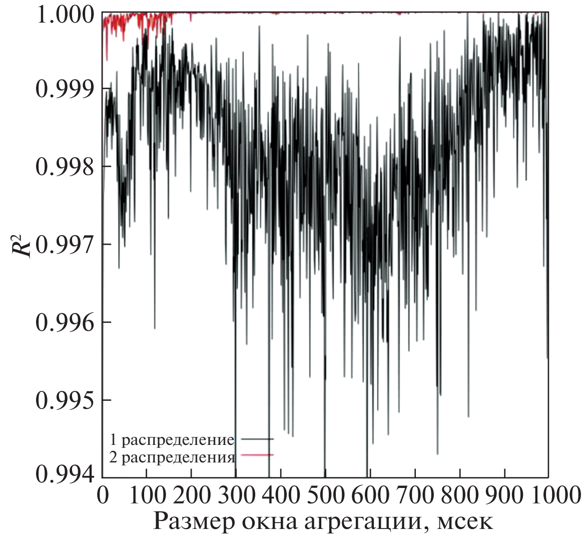

Рис. 23. Зависимость коэффициента R^2 от размера окна агрегации для ПВР скорости передачи сетевых пакетов, в котором учитывались только служебные пакеты, полученные в симплексном режиме (mac: ”00 0e 39 e3 34 00”).

обретая порой хаотический характер. Такое поведение объясняется тем, что количество регистрируемых служебных пакетов существенно меньше общего количества всех сетевых и TCP-пакетов. Для того, чтобы исправить указанную ситуацию, можно при анализе служебных пакетов увеличить размер окна при агрегации измерений.

Более простое и интересное решение обсуждаемой проблемы представлено прямо на самих графиках. Видно, что при аппроксимации анализируемых данных суммой двух логнормальных распределений коэффициент детерминации R^2 выходит на уровень, близкий к 1 практически на всем диапазоне значений окна агрегации.

С учетом того, что в отличие от рассмотренных нами других сетевых пакетов, на служебные пакеты не оказывают воздействия различные случайные параметры, то можно предположить, что распределения служебных пакетов строго подчиняются логнормальному закону.

5. АНАЛИЗ СЕТЕВОГО ТРАФИКА ПО ДИНАМИКЕ МАКСИМУМА ЕГО ИНТЕНСИВНОСТИ

Для дополнительной оценки точности аппроксимации анализируемых данных логнормальным распределением анализировалась динамика зависимости максимума интенсивности сетевого трафика от размера окна агрегации.

Если гипотеза о логнормальном распределении сетевых данных верна, то следует ожидать, что с ростом размера окна агрегации максимум интенсивности будет стремиться (при этом довольно интенсивно) к некой постоянной величине. В противном случае, положение пика интенсивности должно носить случайный характер.

Максимум интенсивности сетевого трафика, аппроксимируемого одним логнормальным распределением, вычисляется как:

$$f(x_{pick}) = 0,$$

где $f(x)$ определена равенством (1).

Откуда получаем выражение для вычисления пика максимума интенсивности (x_{pick}):

$$x_{pick} = e^{\mu-\sigma^2}.$$

На рис. 24, 25 приведены зависимости максимума (x_{pick}) интенсивности от размера окна агрегации для ПВР сетевого трафика в дуплексном и симплексном режимах, соответственно. Видно, что величина пика интенсивности x_{pick} при увеличении размера окна агрегации вначале резко возрастет, после чего медленно и плавно выходит на плато. Зависимости максимума интенсивности от размера окна агрегации для ПВР скорости передачи сетевых пакетов в указанных режимах ведут себя аналогичным образом.

Таким образом, зависимость максимума интенсивности x_{pick} от размера окна агрегации хорошо согласуется с предсказанным поведением, а минимальный разброс данных наблюдения указывает на правильность выбора логнормального распределения для аппроксимации экспериментальных данных и высокой ее точности.

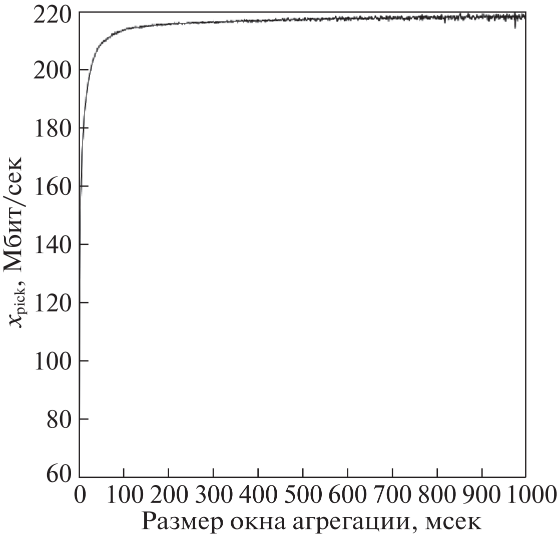

Рис. 24. Зависимость пика интенсивности x_{pick} от размера окна агрегации для ПВР интенсивности сетевого трафика, в котором учитывались только TCP пакеты, полученные в дуплексном режиме.

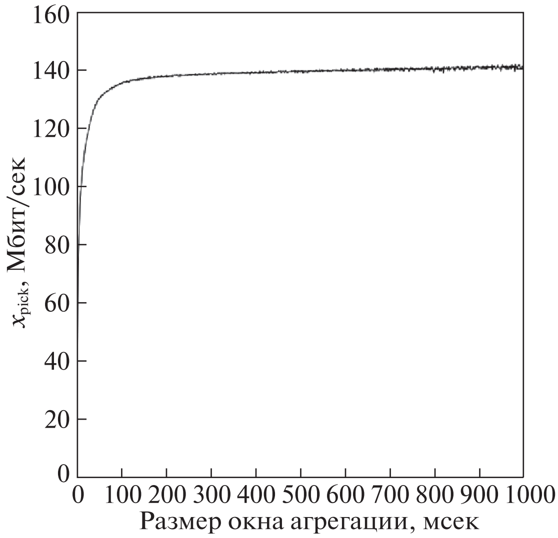

Рис. 25. Зависимость пика интенсивности x_{pick} от размера окна агрегации для ПВР интенсивности сетевого трафика, в котором учитывались только TCP пакеты, полученные в симплексном режиме.

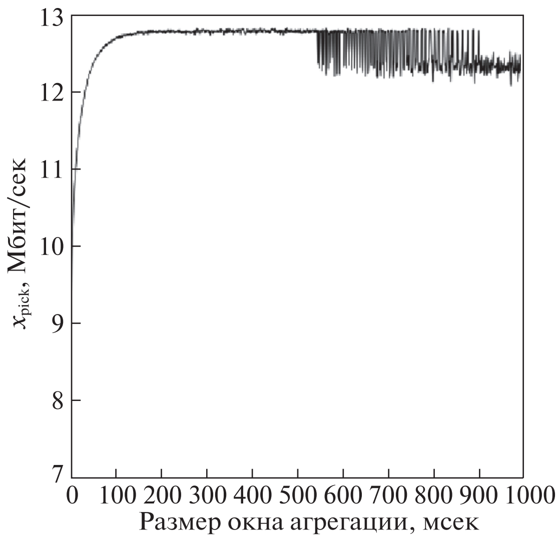

Рис. 26. Зависимость пика интенсивности x_{pick} от размера окна агрегации для ПВР интенсивности сетевого трафика, в котором учитывались только служебные пакеты, полученные в симплексном режиме (mac: “00 0e 39 e3 34 00”).

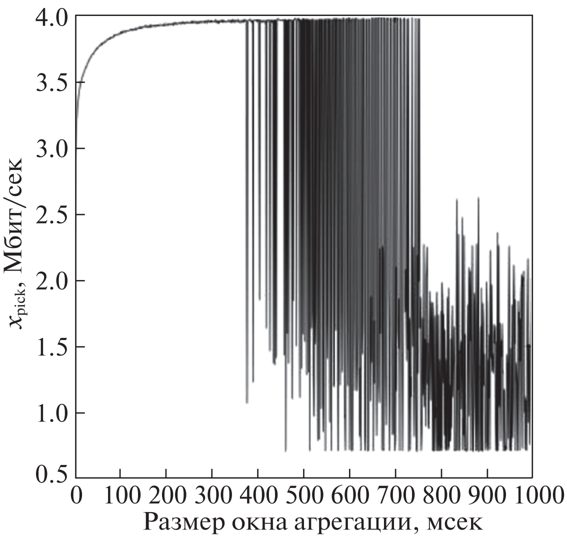

Рис. 27. Зависимость пика интенсивности x_{pick} от размера окна агрегации для ПВР интенсивности сетевого трафика, в котором учитывались только служебные пакеты, полученные в симплексном режиме (mac: “6c 9c ed 7a 49 ca”).

На рис. 26–28 можно наблюдать хаотичное поведение пика интенсивности после некоторого критического значения размера окна агрегации (400+ мсек), в результате которого практически постоянное значение пика интенсивности x_{pick} от размера окна агрегации “разрушается”, что приводит к случайным, резким скачкам пика интенсивности x_{pick} при разных значениях размера окна агрегации.

При этом хаотичное поведение наблюдается только для зависимостей пика интенсивности для ПВР интенсивности сетевого трафика и ПВР скорости передачи сетевых пакетов, в которых учитывались только служебные сетевые пакеты. Такое поведение связано с небольшим количеством служебных сетевых пакетов среди всех собранных сетевых пакетов и, как следствие, высокой погрешностью при вычислении ПВР интенсивно-

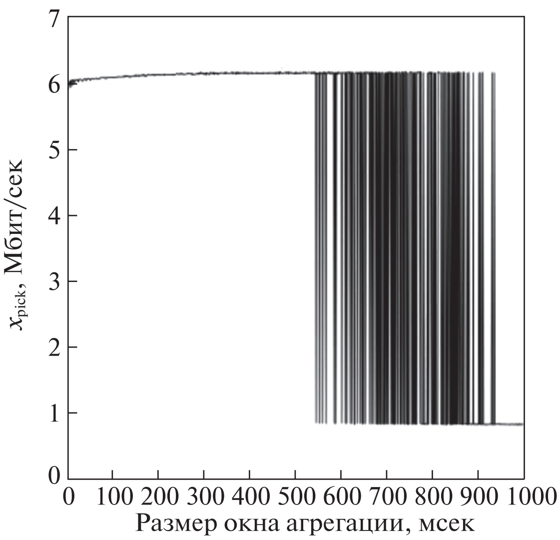

Рис. 28. Зависимость пика интенсивности x_{pick} от размера окна агрегации для ПВР скорости передачи сетевых пакетов, в котором учитывались только служебные пакеты, полученные в симплексном режиме (mac: “6c 9c ed 7a 49 ca”).

сти сетевого трафика и ПВР скорости передачи сетевых пакетов (аналогично ситуации, описанной выше, см. раздел 4).

То, что хаотичное поведение зависимости пика интенсивности x_{pick} и коэффициента детерминации R^2 от размера окна агрегации имеют одинаковую природу, можно также предположить исходя из того, что хаотичное поведение возникает после одного и того же критического значения размера окна агрегации, как для зависимости пика интенсивности x_{pick}, так и для коэффициента детерминации R^2. Например, на рис. 22 и 26 (зависимость коэффициента R^2 пика интенсивности x_{pick} от размера окна агрегации для ПВР интенсивности сетевого трафика, в котором учитывались только служебные пакеты, полученные в симплексном режиме для mac: “00 0e 39 e3 34 00”) хаотичное поведение начинается с момента ≈600 мсек. На рис. 22 и 27 (зависимость коэффициента R^2 пика интенсивности x_{pick} от размера окна агрегации для ПВР интенсивности сетевого трафика, в котором учитывались только служебные пакеты, полученные в симплексном режиме для mac: “6c 9c ed 7a 49 ca”) хаотичное поведение начинается с момента ≈400 мсек. А на рис. 21 и 28 (зависимости коэффициента R^2 пика интенсивности x_{pick} от размера окна агрегации для ПВР скорости передачи сетевых пакетов, в котором учитывались только служебные пакеты, полученные в симплексном режиме для mac: “6c 9c ed 7a 49 ca”), хаотичное поведение начинается с момента времени ≈550 мсек.

ЗАКЛЮЧЕНИЕ

В настоящей работе представлены результаты применения двух численных подходов для аппроксимации суммой логнормальных распределений измерений сетевого трафика, зарегистрированных в магистральном канале, на основе: 1) метода наименьших квадратов, 2) коэффициента детерминации R^2.

В первом и втором подходах был достигнут высокий уровень соответствия данных наблюдения логнормальному закону. При этом, после включения в аппроксимирующую функцию дополнительных слагаемых, точность аппроксимации заметно возрастает.

Показано, что зависимость коэффициента детерминации R^2 от размера окна агрегации для анализируемых сетевых пакетов позволяет контролировать точность аппроксимации данных наблюдения логнормальным законом.

В результате проведенного анализа было установлено, что при $R^2 > 0.75$ точность аппроксимации может считаться допустимой, для $R^2 > 0.90$ – высокой, а в случае $R^2 > 0.95$ – близкой к идеальной.

Интересное явление было обнаружено в случае анализа только служебных пакетов. При аппроксимации указанных измерений суммой двух логнормальных распределений коэффициент детерминации R^2 выходит на уровень близкий к 1 практически на всем диапазоне значений окна агрегации, что указывает на то, что распределения служебных пакетов строго подчиняются логнормальному закону. Отмеченное наблюдение требует отдельного, детального анализа.

Таким образом, проанализирована динамика максимума интенсивности измерений трафика в зависимости от размера окна агрегации и показано, что ее поведение указывает на правильность выбора логнормального распределения для аппроксимации данных наблюдения, а минимальный их разброс подчеркивает высокую ее точность.

СПИСОК ЛИТЕРАТУРЫ

1. *Zrelov P.V., Ivanov Valery V., Ivanov Victor V., Kryukov Yu.A., Tatarinov I.I.* Study of Internet-Traffic Features in the Trunk Channel, Physics of Particles and Nuclei Letters. 2019. V. 16. № 3. P. 289–299.
2. *Antoninou I., Ivanov V.V., Ivanov Valery V., Zrelov P.V.* On the Log-Normal Distribution of Network Traffic, Physica D. 2002. V. 167. № 7. P. 2–85.
3. *Antoninou I., Ivanov V.V., Ivanov Valery V., Zrelov P.V.* Statistical Model of Network Traffic, “Физика элементарных частиц и атомного ядра” (ЭЧАЯ). 2004. Т. 35. Вып. 4. С. 984–1019.
4. *Бахрушин В.Е.* Методы оценивания характеристик нелинейных статистических связей. Системные технологии, 2011. Т. 73. № 2. С. 9–14.

5. *Ершов Э.Б.* Выбор регрессии, максимизирующий несмещенную оценку коэффициента детерминации (рус./англ.) // Айвазян С.А. Прикладная эконометрика. М.: Маркет ДС, 2008. Т. 12. Вып. 4. С. 71–83.
6. *Березин И.С., Жидков Н.П.* Методы вычислений. Т. 1, 2. М.: Физматгиз, 1962, 464 с.
7. *Бахвалов Н.С., Жидков Н.П., Кобельков Г.М.* Численные методы. Под общ. ред. И.Н. Тихонова, 2-е изд. М.: Физматлит: Лаб. базовых данных; Спб.: Нев. диалект, 2002, 630 с.
8. *Гельфанд И.М., Цейтлин М.Л.* ДАН СССР. 1961. Т. 137. № 2. С. 295–98.
9. *Ершов Э.Б.* Распространение коэффициента детерминации на общий случай линейной регрессии, оцениваемой с помощью различных версий метода наименьших квадратов (рус./ англ.) // ЦЭМИ РАН Экономика и математические методы. М.: ЦЭМИ РАН, 2002. Т. 38. Вып. 3. С. 107–120.
10. MAWI Working Group Traffic Archive. URL: http://mawi.wide.ad.jp/mawi/
11. Wireshark, May 2014, http://www.wireshark.org

Vestnik Natsional'nogo issledovatel'skogo yadernogo universiteta "MIFI", 2019, vol. 8, no. 4, pp. 382–396

Approximation of Internet Traffic Measurements in the Trunk Channel by the Sum of Lognormal Distributions

Valery V. Ivanov[a,b], Victor V. Ivanov[b,c], A. V. Kryanev[c,b], and I. I. Tatarinov[d]

[a] *CJSC "MPOTC "TECHNOKOMPLEKT", Dubna, Moscow oblast, 141981 Russia*
[b] *United Institute for Nuclear Research, Dubna, Moscow oblast, 141980 Russia*
[c] *National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, 115409 Russia*
[d] *Kaspersky Lab, Moscow, 123060 Russia*

Received June 20, 2019; revised June 20, 2019; accepted June 25, 2019

Abstract—Two numerical approaches for approximating measurements of the network traffic recorded in a trunk channel based on the traditional least squares method and the coefficient of determination R^2. To additionally estimate the accuracy of the approximation of the analyzed data by a lognormal distribution, the dynamics of the dependence of the maximum intensity of the network traffic on the size of the aggregation window has been analyzed. A high level of compliance of the observation data with a lognormal law has been achieved in both approaches. At the same time, the accuracy of approximation increases noticeably when additional terms are included in the approximating function. It has been shown that the dependence of the determination coefficient on the size of the aggregation window for the analyzed network packets allows one to control the accuracy of the observation data by a lognormal law.

Keywords: network traffic, intensity, approximation, lognormal distribution, aggregation window, coefficient of determination

DOI: 10.1134/S2304487X19040047

REFERENCES

1. Zrelov P.V., Ivanov Valery V., Ivanov Victor V., Kryukov Yu.A., Tatarinov I.I., Study of Internet-Traffic Features in the Trunk Channel, *Physics of Particles and Nuclei Letters*, 2019, vol. 16, no. 3, pp. 289–299.
2. Antoninou I., Ivanov V.V., Ivanov Valery V., and Zrelov P.V., On the Log-Normal Distribution of Network Traffic, *Physica D*, 2002, vol. 167, no. 7, pp. 2–85.
3. Antoninou I., Ivanov V.V., Ivanov Valery V., and Zrelov P.V., Statistical Model of Network Traffic, "Fizika elementarnykh chastits i atomnogo yadra" [Physics of elementary particles and nucleus], 2004, vol. 35, no. 4, pp. 984–1019.
4. Bakhrushin V.E., Metody ocenivaniya kharakteristik nelineynykh statisticheskikh svyazey [Estimation methods of characteristics of nonlinear statistic connections]. *Sistemnye technologii*, 2011, vol. 73, № 2, pp. 9–14.
5. Ershov E.B., Vybor regressii maksimiziruyuschiy nesneschyonnuyu ocenku koefficienta determinacii [A choice of regression that maximizes unbiased estimate of determination coefficient] // Ayvazyan S.A. *Prikladnaya ecometrica*. M.: Market DS, 2008, vol. 12, no. 4, pp. 71–83.
6. Berezin I.S., Zhidkov N.P., Metody vychisleniy [Methods of calculations]. Vol. 1, 2. M.: *Fizmatgiz*, 1962, 464 p.
7. Bakhvalov N.S., Zhidkov N.P., Kobel'kov G.M., Chislennye metody [Numerical methods]. Red. by Tikhonov I.N., 2 ed. M.: Fizmatlit: Lab. bazovykh dannykh; SPb.: Nev. dialekt, 2002, 630 p.
8. Gel'fand I.M., Ceytlin M.L., DAN USSR, 1961, vol. 137, № 2, pp. 295–98.
9. Ershov E.B., Rasprostranenie koefficienta determinacii na obschiy sluchay lineynoy regressii, ocenivaemoy s pomosch'yu razlichnykh versiy metoda naimen'schikh kvadratov // CEMI RAN Ekonomika i mat. metody. M.: CEMI RAN, 2002, vol. 38, no. 3, pp. 107–120.
10. MAWI Working Group Traffic Archive. URL: http://mawi.wide.ad.jp/mawi/
11. Wireshark, May 2014, http://www.wireshark.org