Statistical research and modeling network traffic

Tatiana Tatarnikova[1], *Igor* Sikarev[1], *Vladimir* Karetnikov[2], and *Artem* Butsanets[2,*]

[1]Russian State Hydrometeorological University, ul. Voronezhskaya, 79, 192007 St. Petersburg, Russia
[2]Admiral Makarov State University of Maritime and Inland Shipping, 5/7, Dvinskaya str, Saint-Petersburg, 198035, Russia

Abstract. The self-similarity properties of the considered traffic were checked on different time scales obtained on the available daily traffic data. An estimate of the tail severity of the distribution self-similar traffic was obtained by constructing a regression line for the additional distribution function on a logarithmic scale**.** The self-similarity parameter value, determined by the severity of the distribution "tail", made it possible to confirm the assumption of traffic self-similarity. A review of models simulating real network traffic with a self-similar structure was made. Implemented tools for generating artificial traffic in accordance with the considered models. Made comparison of artificial network traffic generators according to the least squares method criterion for approximating the artificial traffic point values by the approximation function of traffic. Qualitative assessments traffic generators in the form of the software implementation complexity were taken into account, which, however, can be a subjective assessment. Comparative characteristics allow you to choose some generators that most faithfully simulate real network traffic. The proposed sequence of methods to study the network traffic properties is necessary to understand its nature and to develop appropriate models that simulate real network traffic.

1 Introduction

Models for estimating the network traffic servicing characteristics remain at present actual scientific tasks. Reliable network traffic estimates are necessary in the planning of the telecommunications networks development, the differentiated service policies choice and computing resources characteristics that guarantee the required quality of service with the appropriate network load [1,2].

The inflamed interest in studying the network traffic nature is explained by the results of studies showing the long-term dependencies presence in the traffic or the self-similarity process. These changes in the traffics structure are associated with the implementation of the single multiservice network concept, involving the voice, data and multimedia integration [3,4].

* Corresponding author: butsanetsaa@gumrf.ru

To date, the self-similar stochastic processes theory is not as well developed as the Poisson processes theory. Given the known conclusions about the network traffic self-similarity, the actual tasks are the methods of its study and the tools development for generating artificial traffic that adequately reflects the real heterogeneous network traffic [5].

2 Properties and Characteristics Self-Similar Processes

Self-similarity describes a phenomenon in which some statistical characteristics of the process are preserved when the time is scaled. When averaging over the time scale in a self-similar process, there is no rapid "smoothing", that is, a tendency to bursts persists.

Properties that characterize the self-similarity of the process are such as slowly damped dispersion, long-term dependence, the presence of a distribution with heavy "tails" [6].

The property of a slowly decaying dispersion is that the variance of the sample mean decays more slowly than the inverse of the sample size, that is

$$D\left(X^{(n)}(t)\right) = \sigma^2 n^{2H-2},\ n \to \infty, \tag{1}$$

σ^2 –variance of the process $X(t)$;
n – sample size;
H – Hurst parameter (self-similarity parameter), $0.5 < H < 1$.

Note that for traditional random processes, the variance of the sample mean decreases inversely with the sample size: $D\left(X^{(n)}(t)\right) = \sigma^2 / n.$

The presence of a long-term dependence lies in the fact that the self-similar process has a hyperbolically damped correlation function.

The Pareto-distribution is determined by the distribution function in the form

$$R(k) \cong k^{(2H-2)} L(k),\ \forall k \geq 1,\ k \to \infty, \tag{2}$$

$L(k)$ – slowly varying function at infinity, for which

$$\lim_{k\to\infty} \frac{L(kx)}{L(k)} = 1. \tag{3}$$

The property of having a distribution with a heavy "tail" is that the random variable X has a distribution with a heavy "tail", if

$$P(X > x) \sim cx^{-\alpha},\ x \to \infty, \tag{4}$$

$0 < \alpha < 2$ – parameter of the distribution form;
c – positive constant.

3 Methods for Investigating the Self-Similar Process

There are a techniques number that allow us to verify the self-similarity property of the process under investigation.

The self-similarity effect can be observed on the graphs illustrating the change in the time scale, in which the structure of the series obtained by averaging the groups of elements remains the same as the structure of the original one. This fact is a prerequisite for the assumption of the process self-similar structure under consideration and the basis for further analysis.

Consider the daily traffic data from August 20, 2018, provided by the mobile operator MTS in St. Petersburg (Fig. 1). Time series consists of 1440 observations, each of which is the service volume X_t, GByte, min.

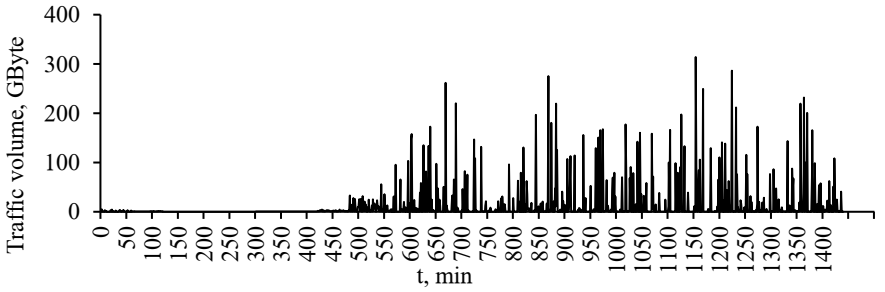

Fig. 1. 2G traffic period 1440 min.

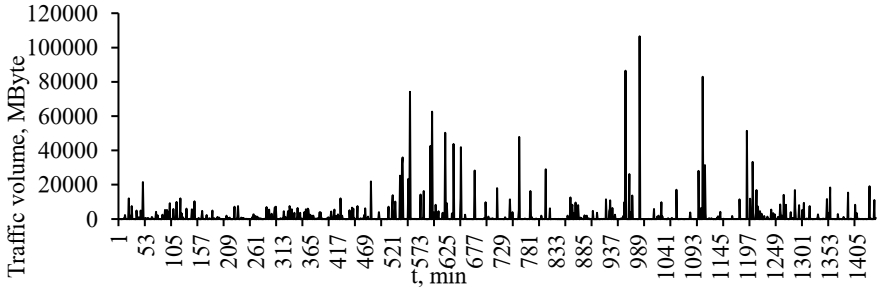

Fig. 2. CS Voice traffic period 1440 min.

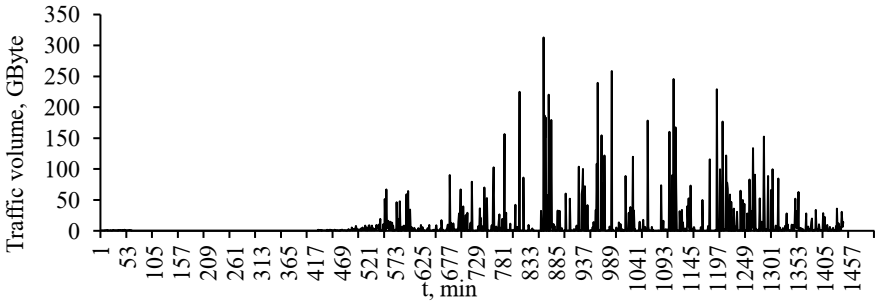

Fig. 3. HSDPA traffic period 1440 min.

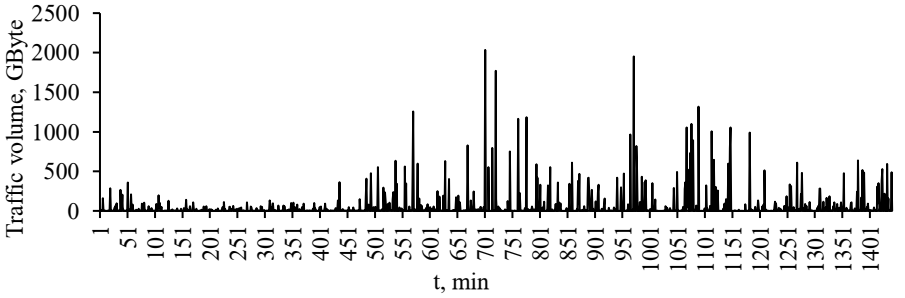

Fig. 4. Mix initial traffic period 1440 min.

We will aggregate the time series by reducing the size of the observation scale by 5 times. The values of the new time series are obtained in accordance with the following expression

$$X_t = \frac{\sum_{i=(t-1)m+1}^{tm} X_i}{m}, \tag{5}$$

m – number averaged consecutive terms of the series.

The new series includes 288 events and is shown in Fig. 5.

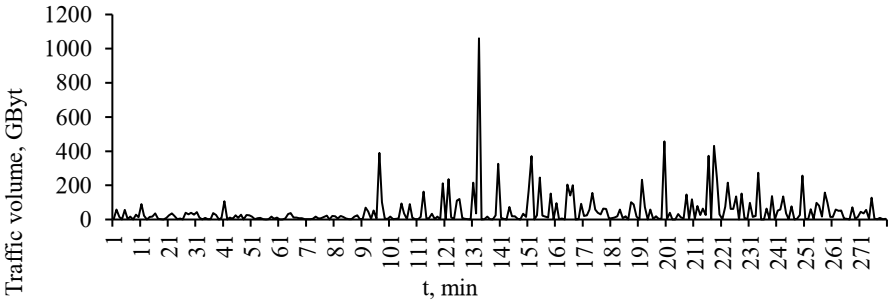

Fig. 5. Aggregation traffic over 5 min. for the period 1440 min.

The same procedure with a decrease in the time scale size of the observations initial series in 10 leads to the result presented in Fig. 6.

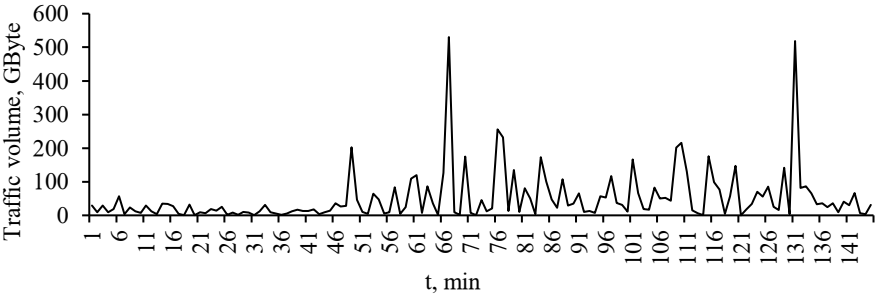

Fig. 6. Aggregation traffic over 10 min. for the period 1440 min.

Reducing the size of the original series of observations time scale 20 times leads to the result shown in Fig. 7.

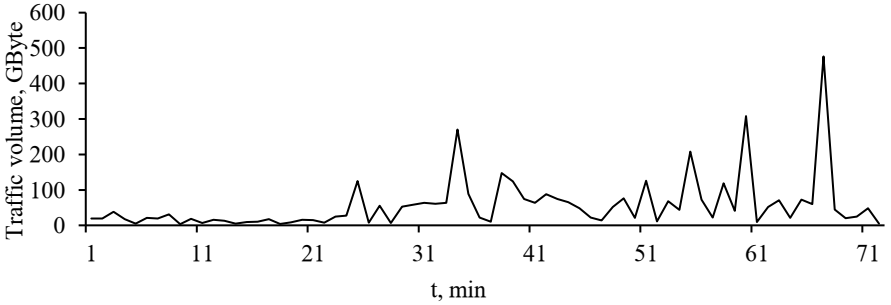

Fig. 7. Aggregation traffic over 20 min. for the period 1440 min.

A visual observation of the aggregated processes in Fig. 1-7 it can be concluded on the preservation process structure.

Further, it is necessary to estimate the distribution "tail" gravity $F(x) = P(X > x)$ (Fig.8)

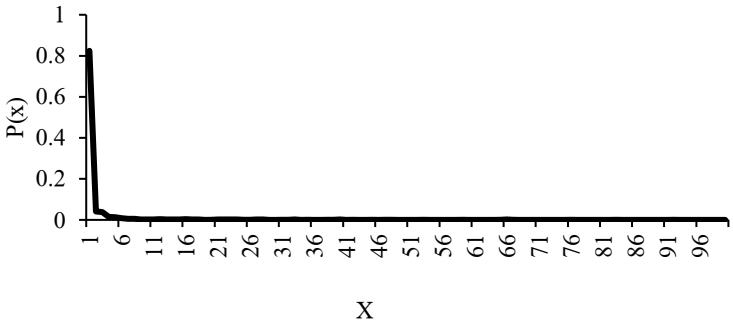

Fig. 8. Distribution "tail" gravity $F(x)$.

Next, you need to assess the gravity "tail" of the distribution - α parameter. To assess the α, it is necessary to construct a graph additional distribution $\bar{F}(x) = 1 - F(x) = P(X > x)$. The angle inclination tangent of the regression line for $\bar{F}(x)$ the horizontal axis is the parameter value α.

The properties of the heavy-tailed distributions are as follows:

- If $\alpha \leq 2$, then the distribution has infinite variance;
- If $\alpha \leq 1$, then the distribution has an infinite average.
- As α decreases, an arbitrary large portion of the density can be represented in the tail of the distribution.

In fact, a heavy tail means the presence of infinite variance, in other words, a random variable can take very large values, but with a very small probability.

Regression equation derived y=-1.29x+0.5067 shows that the α takes a value equal to 1.29 and $\alpha \in [0; 2]$, from which it follows that the traffic distribution data has the property of a heavy tail.

Knowing the α you can find the parameter of self-similarity. We calculate the self-similarity parameter value H=(3-1.29)/2=0.855, which also confirms the process self-similarity properties under consideration, since $H \in [0.5, 1]$.

It is known that the Hurst parameter is a measure of persistence - the tendency of the process to trends. Using the example of filling the tank with incoming and outgoing flow, it can be shown how the Hurst parameter was originally calculated. And so that the water in the tank is stationary, it is necessary that the output flow is equal to the average input, so that the tank never empties or overflows (Fig. 9).

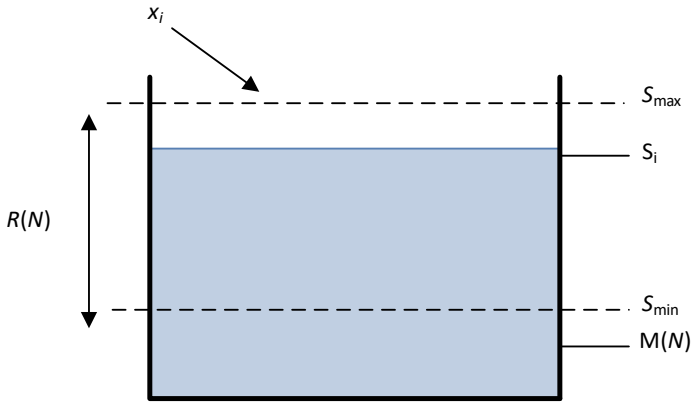

Fig. 9. Explanations for the evaluation of the Hurst parameter.

$$M(N) = \frac{\sum_{i=1}^{N} x_i}{N}, \tag{6}$$

where M – the average input stream over N time units.

$$S(N) = \frac{\sum_{i=1}^{N} x_i - iM(N)}{N}, \tag{7}$$

where S – the difference of the total input stream for N time units and the total output stream for the same time.

$$R(N) = \max_i S - \min_i S, \tag{8}$$

where R – the difference between the maximum and minimum values of S for N time units.

Thus, R is the value that best describes the x variability.

The H is related to the coefficient of the normalized scope R/S, where R is the scope of traffic on the entire time series, and S is the standard deviation

$$R / S = \frac{N}{2^H}. \tag{9}$$

In Table. 1 shows the estimating results variance D, mean M, and parameter H for traffic demonstrating the properties of distributions with heavy tails.

Table 1. Traffic properties evaluation results.

Traffic	***D*(x)**	***M*(x)**	***H***
2G	2506	57.1	0.91
CS Voice	1000	22.7	0.82
HSDPA	3226	43.6	0.89
Mix	3002	5124	0.83

4 Simulations of self-similar traffic

The traditional analysis of telecommunication systems, which is based on the assumption of the Poisson flow, cannot accurately estimate the amount of computing resources and system performance in terms of pulsating traffic [7].

The necessary tools for generating artificial traffic that corresponds to the properties of real network traffic that can be used when modeling the processes of transmission, storage and processing of network traffic.

There are only a few models that are designed to simulate self-similar traffic.

The work implements the tools for generating artificial traffic on the models listed in Table. 2 [8]. Comparative characteristics allow you to choose generators that mimic the real network traffic as plausibly as possible [9]. When comparing, the criterion of the least squares method Y is approximated by the point values of the artificial traffic by the approximating function of the real traffic

$$Y = \sum_{i=1}^{N} \left(F(x_i) - y_i \right)^2, \tag{10}$$

where $F(x_i)$ – values of the approximating function at the points x_i of artificial traffic;
y_i – specified array of source traffic at points x_i.

Every 60th minute is taken as a point, for a total of 24 hours - 24 points.

Table 2. Models of self-similar traffic generators.

Model	**Mathematical model**	**The number of tunable parameters / training**	***Y***
Fractional Brown Motion (FBM) – the movement process of particle performing a chaotic movement with a step given by the history of movement	$B_H(t) = \frac{1}{\Gamma\left(H - 1/2\right)} \int_{-\infty}^{t} K\left(t - t'\right) dB(t')$, $\Gamma(z)$ – gamma function; H –Hurst parameter; $dB(t')$ – independent random displacements of the Brownian particle at time t'; $K(t\text{-}t')$ – memory function of the system: $K\left(t-t'\right) = \begin{cases} \left(t-t'\right)^{H-1/2}, & 0 \le t' \le t \\ \left(t-t'\right) - \left(-t'\right)^{H-1/2}, & t' \le 0. \end{cases}$	Selection and adjustment function K (t - t')	0.32
Fractal Gaussian noise (FGN) – an iterative process of successively dividing a unit length segment in half.	$y_c = \frac{\left(y_1 + y_2\right)}{2} + h,$ h – random variable distributed according to the normal Gaussian law with zero mean and variance $\sigma = r^H$, where $r=(x_r\text{-}x_l)/2$ is the distance from the midpoint of the working segment x_c; H – Hurst parameter.	Setting generation parameters h, selecting the number of iterations.	0.35
Chaotic Map (CMAP) – logistic equation	$X_{n+1} = CX_n - C\left(X_n\right)^2 = CX_n\left(1 - X_n\right),$ where C is a parameter of the propagation velocity of a random variable X, self-similarity manifests itself at $3<C<3.57$	Setting parameter C and basic generation X parameters	0,29

Dynamic Markov Modeling (DMM) – automata with a finite number of states in which a probabilistic transition from one state to another is realized	$$P_{A\to B}=\frac{C_{A\to B}}{\sum_i C_i},$$ where i – the number of the counter; C_i – the value of the i-th counter; $P_{A\to B}$ probability of transition from state A to state B.	Setting transition probabilities	0.24
Fuzzy Logic Modeling (FLM) – burst levels are represented as a step utility function f.	$$U=f(\lambda,P),$$ where U – the traffic burst level λ – average traffic intensity in the self-similarity interval t_i, $i=\overline{1,n}$; $P\sim p_i$, $i=\overline{1,n}$ – probabilities of playing out value.	Setting the step function utility	0.38
Neural Network Modeling (NNM) – functions approximation of several variables by a training sample of a time series	$$Y(t)=\hat{F}(Z(t)),$$ where $\hat{F}(.)$ – the neural network operator; $Z(t)=\{t, X(t), X(t-1),\ldots, X(t-n)\}$ – time series; $X(t)$ – traffic intensity at time t; n – size of the packet.	Neural network training requires tuning $N=x^2y$ scales, where x is the number of neurons in a layer; y – the number of inner layers	0,02
Autoregressive Models (AR) time series, in which any values of the time series linearly depend on the previous values of the same series.	$$X_t=c+\sum_{i=1}^{p}a_iX_t-1+\varepsilon_t,$$ where a - autoregression coefficients; c – constant; p – size of the traffic pack; ε_t – white noise.	Setting a, c, white noise generator parameters ε_t	0,09
ON/OFF-models – source of packets in ON-periods during the time T_0 generates packets, in OFF-periods during the time T_1 the source is passive	The periods T_0, T_1 are random variables with a probability density function $w_0(t)$ and $w_1(t)$, respectively. Distributions $w_0(t)$, $w_1(t)$ are distributions with heavy "tails".	No training required	0.05

In addition to the quantitative assessment, Table 2 also provides qualitative assessments in the form of the laboriousness of implementing a software generator (the number of tunable parameters or the need for training). This is a subjective assessment, which is

difficult to estimate, for example, as the time spent on generator programming or the complexity of the algorithm, since everything depends on the size of the pack, the time spent setting up one parameter and a set of parameters, programming knowledge, and others. For example, despite the fact that the neural network model showed the best result according to the Y criterion, most of its time was spent on choosing the neural network architecture and then setting up the model (3 days), while the model of fractal Gaussian noise was implemented in 40 min, but the Y criterion is 17.5 times greater than that of the neural network model. Moreover, to simulate traffic with a different Hurst parameter, the procedure for choosing a neural network architecture and its training will be required again.

Analysis of the above models allows you to concentrate on the last three presented in Table. 2 and use them in solving problems of modeling telecommunication systems and networks with the resulting global problems – planning the development of telecommunication networks, implementing differentiated services, evaluating the characteristics of computing resources that guarantee the required quality of service of the corresponding traffic [10-12].

5 Conclusion

The article presents the traffic research results in order to identify its self-similarity property. The assumption about traffic self-similar structure is based on the consideration of available data for a different timeline. Using the method of additional distribution function constructing for a logarithmic scale, the gravity of the tail distribution and the self-similarity parameter are estimated. The results obtained allowed us to verify the traffic self-similarity properties in question according to the definition and thus confirm the traffic self-similarity assumption.

Such studies are necessary for understanding the network traffic behavior and developing models that simulate the process real traffic entering the network.

A review existing simulations of self-similar traffic was performed. It is assumed that the model adjustment can be performed according to the Hurst parameter if there are recorded real network traffic traces.

References

1. V.A. Bogatyrev, S.A. Parshutina, DCCN 2015 CCIS **601**, 199-207 (2015) doi: 10.1007/978-3-319-30843-2_21
2. O.I. Kutuzov, T.M. Tatarnikova, *Mathematical Schemes and Communication Systems Simulation Algorithms* (SUAI Publ, St. Petersburg, 2013)
3. A. Tanenbaum, D. Wetherall, *Computer Networks* (Prentice Hall, 2010)
4. O.I. Kutuzov, T.M. Tatarnikova, Moscow Workshop on Electronic and Networking Technologies, MWENT 2018 - Proceedings **1**, 1-3 (2018)
5. W.E. Leland, M.S. Taqqu, W. Willinger, D.V. Wilson, *Proc. ACM SIGCOMM'93* (San-Fransisco, 1993)
6. A.P. Zwart, *Queueing Systems with Heavy Tails* (Eindhoven University of Technology Publ., 2001)
7. O. Kutuzov, T. Tatarnikova, *XV International Symposium "Problems of Redundancy in Information and Control Systems"* (St. Petersburg, Russia, 2016)
8. V.A. Bogatyrev, Engineering Simulation **16(4)**, 463-469 (1999)
9. T. Tatarnikova, M. Kolbanev, *IEEE EUROCON 2009* (St. Petersburg, 2009)

10. S.A. Parshutina, V.A Bogatyrev, *International Conference "Quality Management, Transport and Information Security, Information Technologies" (IT&QM&IS), IET - 2017*, 96-99 (2017)
11. S. Ageev, V. Karetnikov, E. Olkhovik, P. Andrey, E3S Web of Conferences **203**, 05017 (2020) DOI: 10.1051/e3sconf/202015704027
12. S. Ageev, V. Karetnikov, E. Ol'khovik, A. Privalov, E3S Web of Conferences **157**, 04027 (2020) DOI: 10.1051/e3sconf/202020305017
13. G. Kozlov, et al. E3S Web of Conferences, **215**, 03002 (2020) doi:10.1051/e3sconf/202021503002