Approximate waiting times for queuing systems with variable cross-correlated arrival rates

Mikhail I. Bogachev, Nikita S. Pyko, Nikita Tymchenko, Svetlana A. Pyko, Oleg A. Markelov

PII: S0378-4371(24)00661-7
DOI: https://doi.org/10.1016/j.physa.2024.130152
Reference: PHYSA 130152

To appear in: *Physica A*

Received date : 15 July 2024
Revised date : 14 September 2024

Highlights

Approximate waiting times for queuing systems with variable cross-correlated arrival rates

Mikhail I. Bogachev, Nikita S. Pyko, Nikita Tymchenko, Svetlana A. Pyko, Oleg A. Markelov

- Abrupt bursts in traffic dynamics are governed by simultaneous multiple nodes access
- Inter-arrival times distribution is derived from cross-correlations between nodes
- Analytical correction for an approximate evalutaion of waiting times is proposed
- Results are supported by computer simulations and large-batch traffic data analysis
- Corrections are applicable to complex networks governed by autonomous agents dynamics

Approximate waiting times for queuing systems with variable cross-correlated arrival rates

Mikhail I. Bogachev[a,*], Nikita S. Pyko[a], Nikita Tymchenko[a], Svetlana A. Pyko[a], Oleg A. Markelov[a]

[a] *Centre for Digital Telecommunication Technologies, St. Petersburg Electrotechnical University "LETI", 5-F Professor Popov street, St. Petersburg, 197022, Russia*

Abstract

Modern information and telecommunication, transportation and logistic, economic and financial systems are represented by complex networks exhibiting traffic flows with spatio-temporal long-term persistence. Conventional queuing theory relies largely upon stationary models where traffic flows are assumed independent and are typically characterized by the first two moments of inter-arrival and service time distributions, leading to drastic underestimations of traffic flow delays. Here we extend a recent superstatistical approach focusing on traffic models with variable arrival rates by accounting for interdependent activity patterns on multiple network nodes. We suggest an analytical correction to the conventional stationary queue model given by the Kingman's formula based on the calculation of aggregated inter-arrival times variability from the variabilities of arrival rates at individual nodes and cross-correlations between them. We confirm our analytical approximations by comparing with computer simulation results and large-batch empirical traffic analysis from the backbone of a major academic network. We believe that our results, in combination with recent data on the effects of long-term temporal persistence in network traffic flow, are applicable to various complex networks not limited to information and telecommunication, transportation, and logistics but also to economics and finance, rainfall and river flow dynamics, water accumulation in reservoirs, and many other research domains

[*] Corresponding author 1: Centre for Digital Telecommunication Technologies, St. Petersburg Electrotechnical University "LETI", 5-F Professor Popov street, 197022 St. Petersburg, Russia.

Email address: rogex@yandex.com (Mikhail I. Bogachev)

exhibiting spatio-temporal interdependence patterns.

Keywords: non-stationary teletraffic, queuing systems, superstatistics, cross-correlated arrival rates, waiting times

1. Introduction

Understanding traffic flow dynamics is an essential part of the analysis, modeling, and optimization of processes in complex networks relevant for multiple research domains. Although queuing theory based approaches have been widely adopted in operations research with applications ranging from electronic communications, transportation, and public services to the optimization strategies in industrial manufacturing, distribution, logistics, and supply chains, similar formulations are often applicable to various natural and social sciences. River runoff and water accumulation in reservoirs driven by complex rainfall dynamics in the upstream basin, as well as price fluctuations governed by complex interactions between trading agents on financial markets, are among prominent examples of problems commonly addressed by queuing-type models.

Simplified mathematical models of queuing systems, including analytical approximations for the keynote quantities of interest such as queue length, waiting periods, and sojourn times, have been known and successfully utilized for over a century since the groundbreaking works of Erlang [1]. Although further generalizations, for example, by Pollaczek and Khinchine [2, 3], Kingman [4], Marchal [5], Krämer and Langenbach-Belz [6] considered various aspects of queuing systems complexity, one fundamental aspect that hasn't changed since the early years of queuing theory was the assumption of asymptotic independence.

In marked contrast, multiple evidence from research performed in the recent three decades suggests that the independence assumption appears oversimplified for modern complex networks [7, 8, 9, 10], leading to considerable inaccuracies in the evaluation of their performance characteristics [11, 12, 13, 14]. Therefore, there is an increasing interest in finding reasonable analytical and simulational alternatives to the conventional results mainly focusing on stationary queues, that would take into account the real-world variability in both arrival and service rates, as well as cross-correlations in the traffic patterns generated by different network nodes [15, 16, 17, 18, 19].

Deviations from the basic assumption of independence are generally twofold. First, cooperative access patterns on different network nodes that could be quantified by cross-correlations between their traffic intensities lead to clustering of arrivals in the aggregated traffic, in turn manifested in the increased variability of the aggregated arrival rates. Second, there is also temporal persistence in the activity patterns, reflected in the long-term autocorrelations of the arrival rates. The resulting queuing system performance is thus determined by a complex interplay of the above two aspects, leading to seemingly erratic nonlinear dynamics and bursty patterns hardly quantifiable in terms of conventional queuing theory models [20, 21].

Recent advancements in the statistical methodology make the contributions of these two aspects generally distinguishable. Detrended cross-correlation and partial cross-correlation analysis methods [22, 23, 24] are capable of evaluating cross-correlations in the fluctuations around (arbitrary) background trends that could be applied to obtain the parameters of the respective two-compound statistical model at least to a certain approximation, while comparisons with various computer simulation scenarios and partial random shuffling techniques help to evaluate the contributions of the respective correlation components [25].

Recently, we proposed a simplified framework for the numerical simulation and quantitative evaluation of the queuing system's performance based on the superstatistical approach. In this framework, arrivals at individual nodes were assumed to be stationary with fixed rates over short time intervals, and thus short-term dynamics could be described by conventional queuing models based on Poisson flows. In contrast, long-range and long-term dynamics were fully determined by the variability of the arrival rates between different nodes and over long time intervals. Thus, the overall queuing system characteristics could be obtained according to the law of total probability [11, 12, 13, 14]. Based on this approach, in our most recent work we evaluated the effects of the (long-term) autocorrelations and suggested a correction for the queue length and waiting time estimations based on the analysis of the interval distributions between consecutive crossings of the throughput level by the (variable) arrival rate [26]. In turn, the focus of this paper is the extension of the above approach to explicitly account for the effects of cross-correlations between activity patterns at different nodes and show how the variations in the arrival rates translate into the parameters of the waiting times in the aggregated traffic distributions.

This paper is organized as follows: In sec. 2 we provide a brief retrospective over several widely adopted conventional queuing systems models, discussing their underlying assumptions and inherent limitations, and a short outlook on the evolution of modern networks due to the increasing role of AI-based autonomous agents such as IoT (Internet of Things) devices or trading robots and their complex interplay with human network activity. In section 3 we focus on the formalization of the problem and show explicitly that the origin of the variability, either induced by erratic arrival dynamics or variable throughput due to service disruptions, as long as these sources of variability remain independent of each other, does not matter for the queuing system performance, which is determined by the total variance. In section 4, we focus on the extension of the above model that takes into account cross-correlations between traffic dynamics at different nodes and provides an analytical correction to the conventional Kingman's formula. In sections 5 and 6, we validate the analytical treatment by comparing with the same quantities obtained by computer simulations of queuing systems with the same correlation patterns in the arrival rate dynamics, as well as with similar analysis of large-batch empirical traffic patterns collected from the backbone of a major academic network connecting multiple universities and research centers in Japan. Finally, in sections 7 and 8 we discuss the results of this study and outline the key points in a short summary.

2. Retrospective and outlook

From an operations management perspective, the most trivial queuing model scenario implies deterministic inter-arrival times τ and service durations T_S. Once intervals τ between consecutive arrivals are fixed, keeping service durations T_S just below this interval $T_S \lesssim \tau$ for a single service line or $T_S \lesssim k\tau$ for k parallel service lines is sufficient to avoid waiting for service. However, this trivial scenario, denoted as $D/D/k$ in Kendall's notation [27], where k is the number of service lines, is adequate only for very simple processes such as manufacturing identical parts, transmitting standardized messages, or calculating similar non-iterative jobs, using well-established and thoroughly tested methods, algorithms, and procedures.

In practical scenarios, especially when dealing with human-driven dynamics, inter-arrival times τ are variable, and service delays are often imminent, leading to variable T_S. Since the early works of Erlang [1], it has been assumed that both arrival and service processes are governed by stationary Poisson flows, and thus inter-arrival times τ and service durations T_S could be approximated by exponential distributions, each with a single free parameter, the arrival and the service rates, respectively. Of note, for multiple service lines k operating in a cycle, times between arrivals directed to a specific line follow Erlang distribution of order k, that is directly generalized into Γ-distribution, in turn converging to Gaussian at large k.

Following advancements mainly focused on non-exponential (generalized, denoted as $G/G/k$) distributions of inter-arrival times and service durations, leading to the introduction of variances of τ and T_S as additional model parameters [2, 3, 4]. Since for a node or link with a fixed throughput c the amount of traffic flow v is linearly proportional to the service time $T_S = v/c$, variances of T_S are often substituted by the variances of v. To do so without further loss of generality, variances are typically normalized to the respective averaged, and thus coefficients of variation $\rho_t au$ and ρ_v are used as model parameters.

While further generalizations focused on extended models of inter-arrival and service time distributions [5, 6], they either ignored the corresponding temporal and spatial memory patterns or reduced them to simplified models focusing mainly on short-range and short-term correlations [15, 16, 17, 18, 19].

While multi-lane service models may still be applicable to certain scenarios, e.g., toll road checkpoints, in modern information networks that are largely heterogeneous in both sources of traffic and means for its transmission, the concept of the overall effective transmission throughput under different scenarios is becoming more relevant. Modern networks based on 5G and industrial IoT technologies are associated with an increasing number of nodes, with their dynamics determined at the application level of the TCP/IP stack, and thus their activity patterns are largely software-defined. Moreover, specific transmission routes are also selected at the software level using intelligent routing algorithms, and thus, from the teletraffic analysis point of view, modern networks are largely software defined networks (SDN) [28].

Accordingly, when modeling modern SDNs, instead of defining a specific topology with a fixed number of links, it is often plausible to reduce the model to the analysis of aggregated traffic flow with variable arrival rates over an aggregated link with variable throughput, both of them generally described by random processes. Moreover, as long as the above two sources of variability can be considered independent, its particular origin does not matter for the queuing system performance, since only the total variance plays the role, which allows for a further model simplification (for deeper details, we refer to Eq. 10 and its explanation at the end of sec. 3).

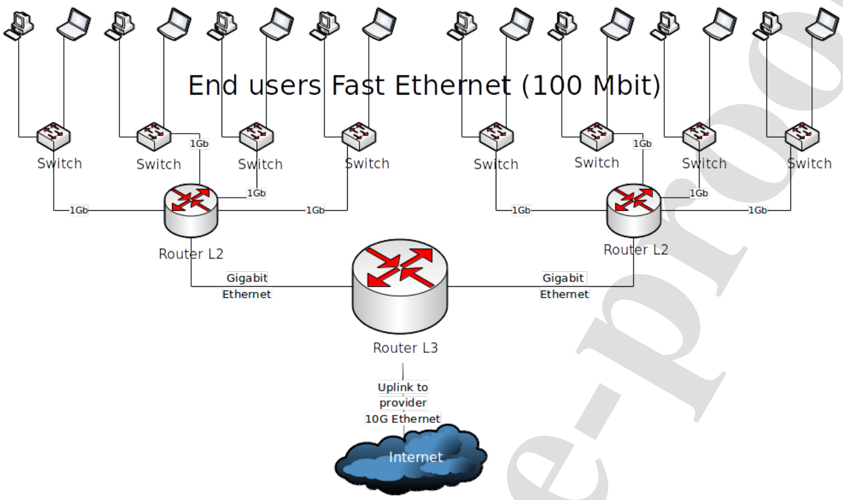


Figure 1: A typical example of the star local area network topology.

3. Theoretical postulation

Let us assume a simple network consisting of N end nodes that generate traffic and are connected to a single aggregated link with limited throughput c. Among prominent examples are local area networks (LAN), where multiple end nodes (user- and/or IoT-devices) are connected to the upstream internet service provider (ISP) via a single external link (for a typical topology, see Fig. 1).

Once all arrivals from N nodes are aggregated in a single link as indicated in Fig. 2, the formulation is reduced to a model with random inter-arrival and service times, although following a generalized distribution, typically denoted as $G/G/1$. Figure 2b illustrates the queuing and service diagrams exemplified for the first ten arrivals, where the sojourn times T_C are generally composed of waiting times in the queue T_W and service times T_S, respectively. In a simple case when the link is empty at the time of arrival (exemplified by the first and the fifth arrivals in Fig. 2), $T_W = 0$, and thus $T_C = T_S$.

Waiting time for the k-th customer from arrival to the beginning of service, under the assumption that arrivals occur at random times (Poisson

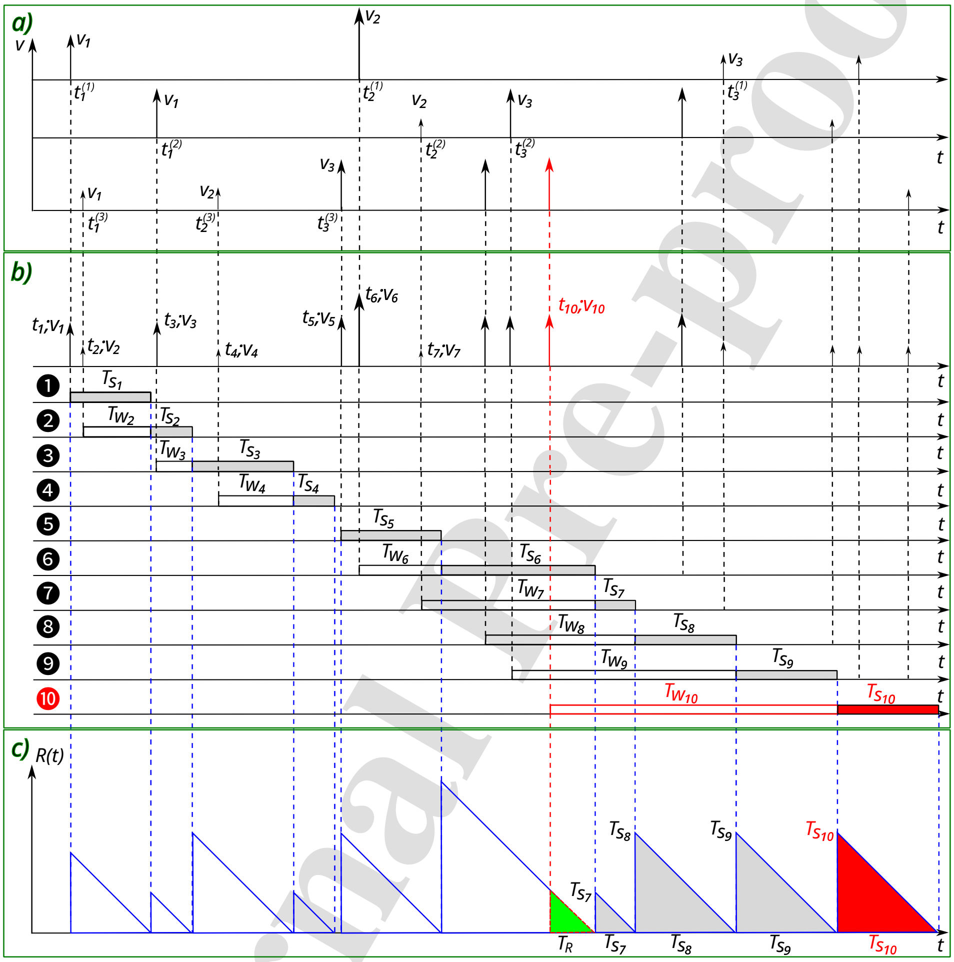


Figure 2: a) Arrivals from three different nodes each characterized by their arrival times t_i and size v_i, the latter proportional to the service time as $T_{S_i} = v_i/c$. b) Aggregated arrivals are served either immediately (exemplified by the first and the fifth one), or queue for service over time T_{W_i}, resulting in sojourn time $T_{C_i} = T_{W_i} + T_{S_i}$ as shown for the first ten arrivals. c) Illustration of the waiting time evaluation procedure assuming arrival at a random time exemplified for the $k = 10$-th arrival as $T_{W_k} = T_R + \sum_{l=1}^{L} T_{S_{k-l}}$, where T_R is the remaining time required to complete the current service performed at the time of arrival, while L is the queue length at the time of arrival.

process), could be evaluated as

$$T_{W_k} = T_R + \sum_{l=1}^{L} T_{S_{k-l}}, \tag{1}$$

where T_R is the remaining time required to complete the current service performed at the time of arrival, while L is the queue length at the time of arrival. Similarly, the sojourn time from arrival to service completion

$$T_{C_k} = T_{W_k} + T_{S_k} = T_R + \sum_{l=0}^{L} T_{S_{k-l}}. \tag{2}$$

According to Little's law [29], the average waiting time $\langle T_W \rangle$ is proportional to the average queue length $\langle L \rangle$, and thus on the average

$$\langle T_W \rangle = \langle T_R \rangle + \langle T_S \rangle \cdot \langle L \rangle = \langle T_R \rangle + \langle T_W \rangle \cdot U, \tag{3}$$

yielding

$$\langle T_W \rangle = \frac{\langle T_R \rangle}{1 - U}, \tag{4}$$

where $U = c/c_0$ is the link utilization, assuming that $c_0 = \left(\sum_i v_i\right)/T_\Sigma$ is the minimum throughput required to perform the entire service over the given analysis time.

In turn, assuming arrivals at random times, T_R can be determined by calculating the expectation value of the residual times $R(t)$ that follow a renewal process represented by a sequence of linear functions decaying as $1/t$ from T_{S_i} at the beginning of the service of i-th arrival until it reaches zero at its completion, as depicted in Fig. 2c, as follows:

$$\langle T_R \rangle = \frac{1}{t} \int_0^t R(t')dt' = \frac{1}{t} \sum_{i=1}^{L(t)} \frac{T_{S_i}^2}{2} = \frac{L(t)}{2t} \sum_{i=1}^{L(t)} \frac{T_{S_i}^2}{L(t)}. \tag{5}$$

Since at large $t \to \infty$ $L(t)/t \to Y$, where Y is the mean arrival rate, while

$$\frac{L(t)}{t} \sum_{i=1}^{L(t)} \frac{T_{S_i}^2}{L(t)} = E(T_S^2), \tag{6}$$

where $E(\ldots)$ is the expectation, this finally yields $\langle T_R \rangle = Y E(T_S^2)/2$, and thus [30]

$$\langle T_W \rangle = \frac{Y E(T_S^2)}{2(1-U)}. \tag{7}$$

Equation 7 widely known as the Pollaczek-Khinchine formula for waiting times implies stationary random arrivals, although variable service times (typically denoted as $M/G/1$ queue) [2, 3]. It could also be rewritten as

$$\langle T_W \rangle = \frac{1}{\langle T_S \rangle} \frac{U}{1-U} \frac{E(T_S^2)}{2}. \tag{8}$$

Finally, one can avoid direct calculation of the second moment $E(T_S^2)$ by replacing it with the squared coefficient of variation $\rho_v^2 = E(T_S^2)/E^2(T_S) = E(v^2)/E^2(v)$, altogether yielding

$$\langle T_W \rangle = \langle T_S \rangle \frac{U}{1-U} \frac{\rho_v}{2} = \frac{\langle v \rangle}{c} \left[\frac{U}{1-U} \right] \left[\frac{\rho_v^2}{2} \right]. \tag{9}$$

In case arrivals are no longer Poissonian, with inter-arrival times following a generalized two-parameter distribution, under the assumption that variabilities of inter-arrival and service times are independent, and thus the total variance is the sum of variances, which also holds for the squared coefficients of variation, similar considerations lead to the well-known Kingman's formula [4]

$$\langle T_W \rangle \approx \frac{\langle v \rangle}{c} \left[\frac{U}{1-U} \right] \left[\frac{\rho_\tau^2 + \rho_v^2}{2} \right] = \frac{\langle v \rangle}{c_0} \left[\frac{U^2}{1-U} \right] \left[\frac{\rho_\tau^2 + \rho_v^2}{2} \right], \tag{10}$$

where ρ_τ and ρ_v are the coefficients of variation for the inter-arrival and service times, respectively.

Equation 10 has a clear physical interpretation indicating that the waiting times are determined by (i) the fixed term that represents the average service time $\langle T_S \rangle = \langle v \rangle / c$, (ii) the utilization term $U/(1-U)$ that depends solely on the throughput U, and (iii) the variability term $(\rho_\tau^2 + \rho_v^2)/2$ that depends explicitly on the coefficients of variations of inter-arrival and service times, respectively. Another important consequence of Kingman's formula is that the origin of the variance generally does not matter. Whether it arises from users' erratic behavior resulting in bursty access patterns and non-exponential inter-arrival times, or from unstable service with frequent disruptions leading to variable total aggregated link thoughput, assuming that these two sources of variance are independent, only the total variance matters for the service delays.

4. Analytical treatment

Now let us assume that the traffic generated by N nodes connected to a single aggregated link with limited throughput c exhibits cross-correlations characterized by the correlation matrix $\mathbf{R} = [R_{ij}]$.

The total traffic denoted as Y could be considered a sum of the constant (flat rate) component Y_0 and the (zero mean) fluctuating component Y_f. In this postulation, Y_0 is represented by a fixed-intensity (stationary) Poisson flow, with its contribution following simple laws of the conventional queuing theory. In turn, Y_f characterizes random fluctuations around the fixed level Y_0. Since the stationary component is characterized by a single fixed intensity parameter $Y_0 = C$, it is reasonable to express the standard deviation of the fluctuating component Y_f in the units of C as σ/C.

However, a Gaussian distribution with any non-zero variance implies a non-zero probability of negative values, which makes it not the best choice for modeling traffic intensity, which is non-negative by definition. To overcome this issue, one commonly replaces Gaussian by another distribution with non-negative support, such as Γ-distribution which has the same number of free parameters that could be easily expressed via the mean C and the coefficient of variation σ/C [26].

Let us assume that the total traffic intensity $Y = Y_0 + Y_f$ is Γ-distributed

$$P(Y) = \frac{\lambda^\alpha}{\Gamma(\alpha)} Y^{\alpha-1} \exp(-\lambda Y), \tag{11}$$

where $\Gamma(\alpha)$ is the Γ-function, α is the shape parameter and λ is the rate parameter, the average arrival rate is then given by $\langle Y \rangle = \alpha/\lambda$ and its variance equals $\sigma_Y^2 = \alpha/\lambda^2$. Accordingly, the shape parameter $\alpha = 1/\left(\sigma_Y/\langle Y \rangle\right)^2 = 1/\rho^2$ that determines the asymptotic decay of the distribution depends solely on the coefficient of variation $\rho = \sigma_Y/\langle Y \rangle$, and the rate parameter $\lambda = \alpha/\langle Y \rangle$ is responsible for the rescaling of the distribution to the given average arrival rate $\langle Y \rangle$.

According to the law of total probability

$$P(\tau) = \int_0^\infty P(Y) P(\tau|Y) \mathrm{d}Y = \int_0^\infty P(Y) Y \mathrm{e}^{-Y\tau} \, \mathrm{d}Y. \tag{12}$$

where $Y = 1/\langle\tau\rangle$ is the local intensity for each time fragment that is sufficiently short to be considered stationary, $\langle\tau\rangle$ is the local average inter-access time for each fragment. Here $P(\tau|Y) = Y\mathrm{e}^{-Y\tau}$ denotes the conditional probability to observe inter-arrival time τ within a fragment with local access rate Y.

Then the marginal inter-arrival time distribution is given by [12], see also [31, 32]

$$P(\tau) = C[1 + b(q-1)\tau]^{-1/(q-1)}, \tag{13}$$

where

$$q = 1 + \frac{1}{\alpha+2} = \frac{\alpha+3}{\alpha+2}\ ,$$

$$b = \frac{\overline{\beta}}{3-2q} = \frac{\alpha+2}{\lambda}\ ,$$

$$C = \frac{\alpha(\alpha+1)}{\lambda^2}.$$

The above is known as the q-exponential distribution and represents a subclass of generalized Pareto distributions. This class of distributions has been associated with the maximization of generalized entropy [33] and found numerous applications in the dynamical and structural characterization of various complex systems [12, 34, 35, 36, 37, 38, 39, 40]. Among other prominent applications, a similar result has been obtained in earlier works in the context of rainfall dynamics to characterize intervals between rainy days [31]. More recently, the same method was applied to characterize waiting time statistics between periods of low wind [32]. For $q = 1$ q-exponential distribution reduces to a simple exponential distribution.

The expectation and the variance of the q-exponential distribution equals

$$E(\tau) = \frac{1}{b(3-2q)}\ ,$$

and

$$D(\tau) = \frac{q-2}{(2q-3)^2(3q-4)b^2}\ .$$

defined for $q < 3/2$ and $q < 4/3$, respectively. Accordingly, the squared coefficient of variation equals

$$\rho_\tau^2 = \frac{D(\tau)}{E^2(\tau)} = \frac{(q-2)b^2(3-2q)^2}{(2q-3)^2(3q-4)b^2} = \frac{q-2}{3q-4}.$$

Taking into account that $\alpha = 1/\rho^2$ we can rewrite the shape parameter as

$$q = \frac{\alpha+3}{\alpha+2} = \frac{\frac{1}{\rho^2}+3}{\frac{1}{\rho^2}+2} = \frac{1+3\rho^2}{1+2\rho^2}.$$

Then

$$q-2 = \frac{1+3\rho^2}{1+2\rho^2} - 2 = \frac{-\rho^2-1}{1+2\rho^2}$$

and

$$3q-4 = \frac{3+9\rho^2}{1+2\rho^2} - 4 = \frac{\rho^2-1}{1+2\rho^2}$$

finally yielding

$$\rho_\tau^2 = \frac{-\rho^2-1}{1+2\rho^2} \cdot \frac{1+2\rho^2}{\rho^2-1} = \frac{-\rho^2-1}{\rho^2-1} \tag{14}$$

The above expression is valid for

$$q = \frac{1+3\rho^2}{1+2\rho^2} < \frac{4}{3} => \rho^2 < 1.$$

Next, let us assume that the aggregated intensity Y is obtained by adding up traffic contributions from N nodes. For simplicity, let us first assume that the contribution of all nodes to the overall traffic intensity is the same, and traffic variations from individual nodes are equicorrelated. Then, according to [41] one could express the expectation and the variance via the moment generation function, which for an arbitrary $f(N)$ is defined as

$$E_N\left(f\left(N\right)\right) = \sum_{n=0}^{\infty} f\left(n\right) P\left(N=n\right)$$

yielding

$$\varphi''\left(0\right) = E\left(N^2\right)$$

and

$$D\left(N\right) = E\left(N^2\right) - E^2\left(N\right) = \varphi''\left(0\right) - \left(\varphi'(0)\right)^2.$$

Assuming that

$$E\left(X_i|N\right) = e^{g+hN},$$
$$D\left(X_i|N\right) = e^{a+tN}.$$

one arrives at

$$E\left(Y\right)=e^{g}\varphi'\left(h\right),$$

$$D\left(Y\right)=e^{a}\left\{\varphi'\left(t\right)+E_N\left(R\left(N^2-N\right)e^{tN}\right)\right\}+e^{2g}\varphi''\left(2h\right)-e^{2g}(\varphi'(h))^2.$$

For the equicorrelated scenario $R=[r_{ij}]=R_0, i\neq j$ one obtains [41]

$$D\left(S\right)=e^{a}\left\{(1-R_0)\varphi'\left(t\right)+R_0\varphi''(t)\right\}+e^{2g}\varphi''\left(2h\right)-e^{2g}(\varphi'(h))^2$$

and thus

$$\rho^2=\frac{D(S)}{E^2(S)}=e^{a-2g}\frac{(1-R_0)\varphi'\left(t\right)+R_0\varphi''(t)}{(\varphi'(h))^2}+\frac{\varphi''(2h)}{(\varphi'(h))^2}-1.$$

Finally, by substituting

$$E\left(S\right)=E\left(Y_k\right)E_N\left(N\right);$$

$$D\left(S\right)=D\left(Y_k\right)\left(1-R_0\right)E_N\left(N\right)+D\left(Y_k\right)R_0E_N\left(N^2\right)+(E\left(Y_k\right))^2D_N\left(N\right)$$

one obtains

$$\rho^2=\frac{D(S)}{E^2(S)}=\frac{D\left(Y_k\right)\left(1-R_0\right)}{E^2(Y_k)E_N(N)}+\frac{D\left(Y_k\right)R_0E_N\left(N^2\right)}{E^2(Y_k)(E_N(N))^2}+\frac{D_N(N)}{(E_N(N))^2}$$

and thus finally

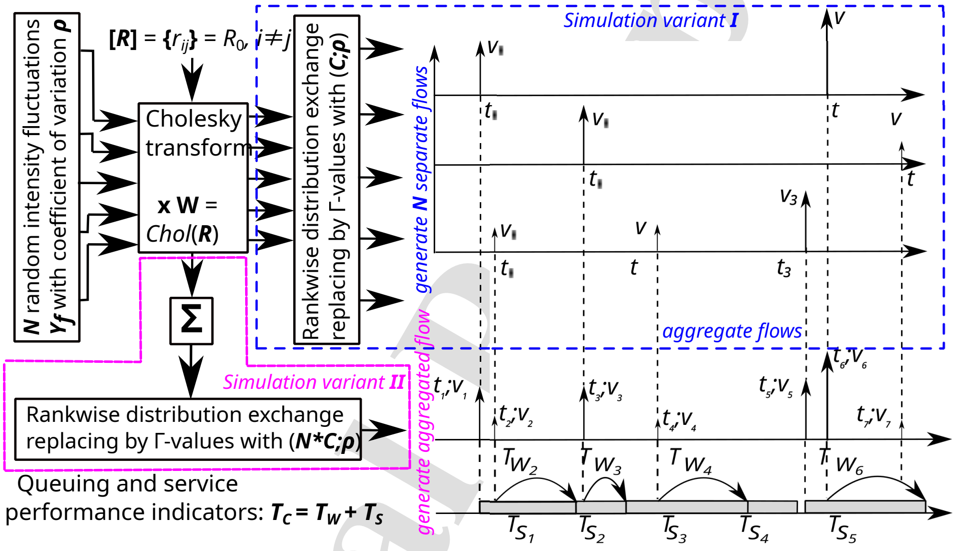
$$\rho^2=\frac{D(S)}{E^2(S)}=\frac{ND\left(Y_k\right)\left(1-R_0\right)}{N^2E^2(Y_k)}+\frac{D\left(Y_k\right)R_0N^2}{N^2E^2(Y_k)}=\frac{D\left(Y_k\right)\left(NR_0+1-R_0\right)}{NE^2(Y_k)}. \tag{15}$$

The above expression is valid for

$$\rho^2<1=>\frac{D\left(X_1\right)\left(NR_0+1-R_0\right)}{NE^2(X_1)}<1=>\frac{D\left(X_1\right)}{E^2(X_1)}<\frac{N}{NR_0+1-R_0}.$$

To apply the above result, one first obtains ρ^2 from Eq. 15 and substitutes ρ^2 in Eq. 14 for ρ_τ^2. Finally, ρ_τ^2 obtained from Eq. 14 could be used to estimate the waiting times according to Eq. 10.

Simulation model parameters: ***N* - number of aggregated nodes; *L*** - record duration; ***V*** - number of configurations; ***C*** - average intensity; ***ρ*** - coefficient of intensity variations; **[*R*] = {*r*$_{ij}$}** - cross-correlation matrix of intensities at individual nodes



N random intensity fluctuations
Y_f with coefficient of variation ρ
[R] = {r_ij} = R_0, i≠j
Cholesky transform
x W =
Chol(R)
Σ
Rankwise distribution exchange replacing by Γ-values with (C;ρ)
Simulation variant I
generate N separate flows
aggregate flows
Simulation variant II
Rankwise distribution exchange replacing by Γ-values with (N*C;ρ)
generate aggregated flow
Queuing and service performance indicators: T_C = T_W + T_S
t_1;v_1
t_2;v_2
t_3;v_3
t_4;v_4
t_5;v_5
t_6;v_6
t_7;v_7
T_W2
T_W3
T_W4
T_W6
T_S1
T_S2
T_S3
T_S4
T_S5

Figure 3: Two variants of the correlated traffic flows simulation procedure: (I) separate Poissonian flows with variable intensities are generated for each of N nodes independently, followed by their aggregation and arrangement in accordance with their arrival times; (II) traffic intensities from N nodes are added to generate the total intensity records, followed by the simulation of a single aggregated Poissonian flow. Next, the aggregated traffic flow is subjected to queuing system simulations as indicated in Fig. 2. to obtain the performance quantities T_S, T_W, and T_C that are finally being compared with the results obtained by analytical treatment.

5. Computer simulations

To validate the above results, we have generated surrogate arrival rate series for both uncorrelated and long-term correlated traffic intensity records as indicated in Fig. 3.

First, we generated the intensity series $Y = Y_0 + Y_f$, where $Y_0 = C$ is a simple constant, and Y_f is a fluctuating time series with zero mean. For each combination of the model parameters, in particular, the number of nodes $N = 4$, 16, and 64, the mean intensity $C = 10$ and 100, the coefficients of variation of the intensity fluctuations $\rho = 0.1 \ldots 0.9$, and cross-correlation coefficients $R_0 = 0.1 \ldots 0.9$, we have generated random configurations of for different segment durations L. For a more relevant comparison between simulated data and observational records, we have chosen the parameter L such that the overall statistics remain the same as for a complete daily record from a network with the same number of N active nodes.

For further queuing system simulation, in each of the intensity segments we generated local Poissonian flows consisting of Y_k elements with average interval $1/Y_k$, and concatenated them for all $k = 1 \ldots L$. The resulting arrival series was used as an input to the standard queuing system simulation procedure as described, for example, in [12, 26]. We tested two simulation options, (I) with the Poissonian flows generated for each of N nodes independently followed by their aggregation and arrangement in accordance with their arrival times, as well as (II) with the aggregation of the local traffic intensities from N nodes followed by the simulation of a single aggregated Poissonian flow (for a detailed illustration of both simulation scenarios, we refer to Fig. 3).

Figure 4 shows distributions $P(\tau)$ of inter-arrival times τ in a network with $N = 4$ active nodes with average intensity $C = 100$ per time unit, duration $L = 860$ time units, with statistics collected over $V = 100$ random configurations, with boxplots indicating results of computer simulations, in comparison with analytical approximations according to Eq. 13. The figure shows that, while all discrete simulation variants are affected by both discreteness and finite size effects, with variant I representing the upper and variant II the lower bounds, respectively, the analytical curve typically lies between them, with more pronounced deviation from a single exponential with increasing ρ and R_0, respectively. Figures 5 and 6 show similar results for the simulated networks with $N = 16$ and 64 nodes, respectively.

Figures 7, 8 and 9 show representative examples of waiting times T_W

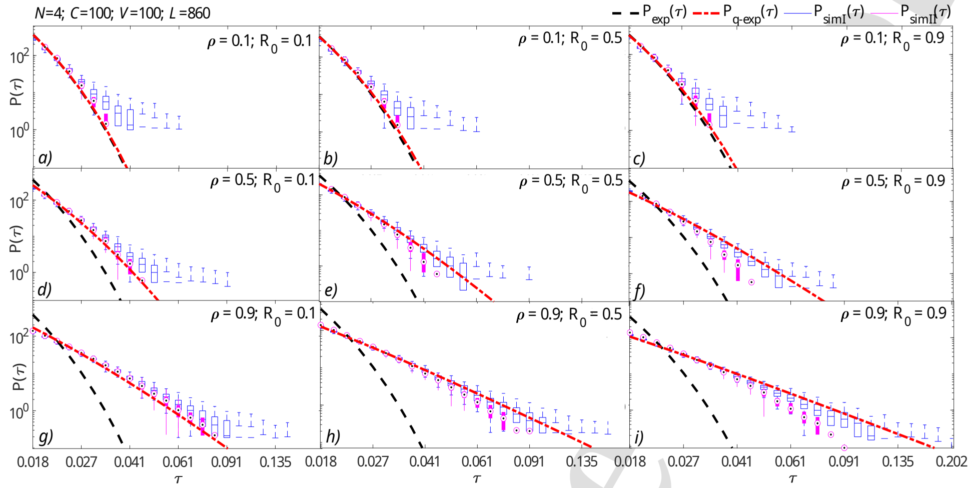


Figure 4: Inter-arrival time distributions $P(\tau)$ for the simulated traffic in a network with $N = 4$ active nodes with average intensity $C = 100$ per time unit, duration $L = 860$ time units, and statistics collected over $V = 100$ random configurations. Open and full boxplots correspond to the I and II simulation variants. The dashdot red curves correspond to the analytical approximations for ρ_τ according to Eq. 13, while the dashed black curves show a simple exponential with the same average inter-arrival interval.

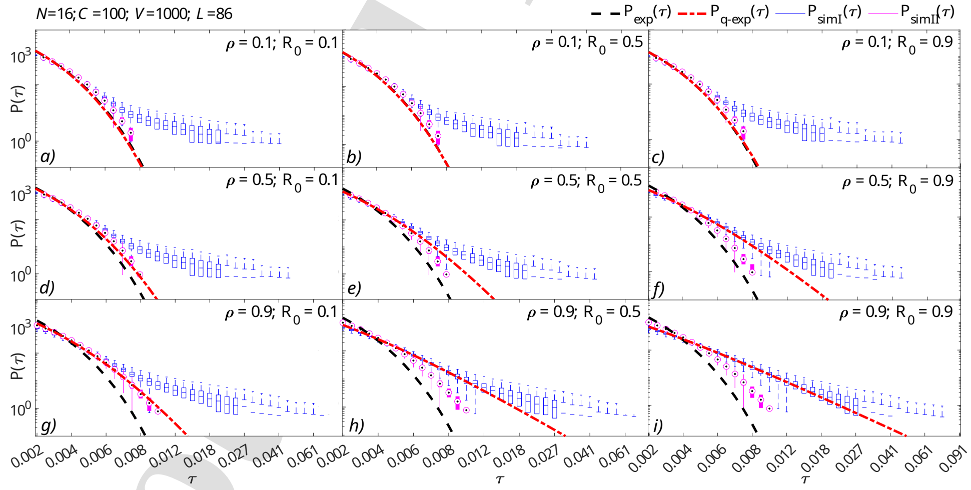


Figure 5: Similar to Fig. 4 but for a network with $N = 16$ active nodes with average intensity $C = 100$ per time unit, duration $L = 86$ time units, and statistics collected over $V = 1000$ random configurations.

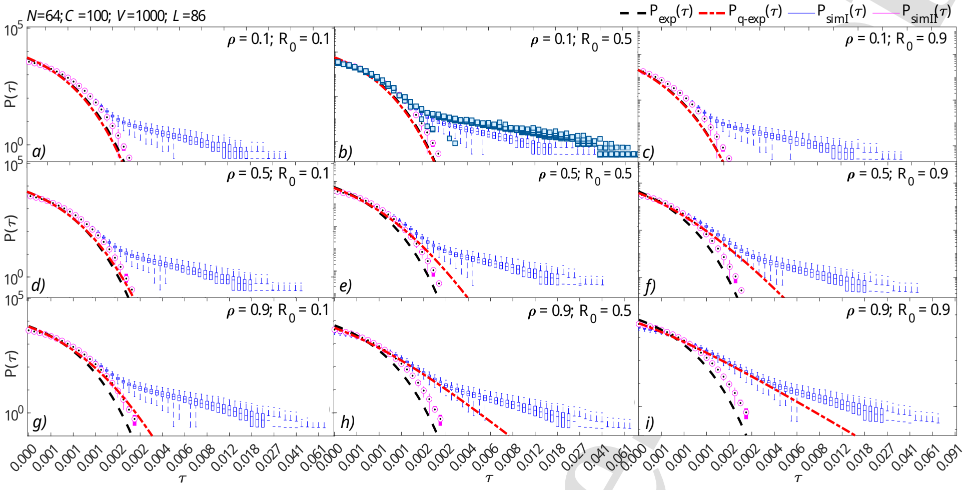


Figure 6: Similar to Fig. 4 but for a network with $N = 64$ active nodes with average intensity $C = 100$ per time unit, duration $L = 86$ time units, and statistics collected over $V = 1000$ random configurations.

as functions of the throughput utilization U for the simulated aggregated traffic in networks of $N = 4$, 16, and 64 active nodes, respectively. Open and full boxplots provide results of computer simulations according to variants I and II, respectively, while dashed lines correspond to the approximations using the Kingman's formula (10) [4] with respective analytical corrections to the coefficient of variation of the inter-arrival times ρ_τ^2 according to Eq. 14 depending on the cross-correlations of the intensities from different nodes, as determined by the analytical treatment.

6. Observational traffic data analysis

The above simulations corresponding to equicorrelated traffic patterns with equal arrival rates and variability patterns may seem rather hypothetical, especially in comparison with common real-world environments, where distribution of intensity over multiple nodes tend to follow heavy-tailed functional forms [13], and their correlation patterns exhibit complex evolutions over time [26]. In order to evaluate whether the above analytical approximations are relevant for traffic analysis in real-world networks, in the following we analyzed the same statistics for the empirical traffic data collected from a sample point on the backbone of the MAWI academic network connecting

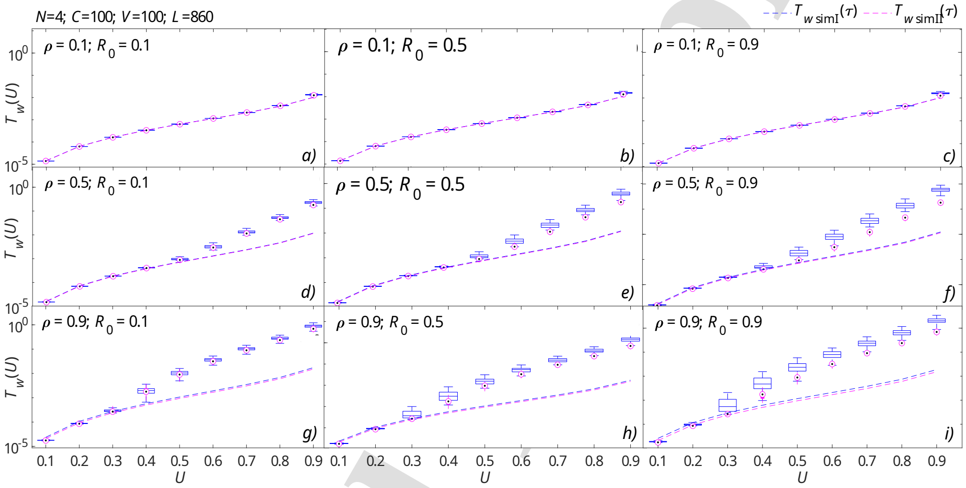


Figure 7: Representative examples of waiting times T_W as functions of the throughput utilization U for the simulated aggregated traffic in networks of $N = 4$ active nodes with average intensity $C = 100$ per time unit, duration $L = 860$ time units, and statistics collected over $V = 100$ random configurations. Open and full boxplots provide results of computer simulations according to variants I and II, respectively, while dashed lines correspond to the approximations using the Kingman's formula (10) [4] with respective analytical corrections to the coefficient of variation of the inter-arrival times ρ_τ^2 according to Eq. 14 depending on the cross-correlations of the intensities from different nodes, as determined by the analytical treatment.

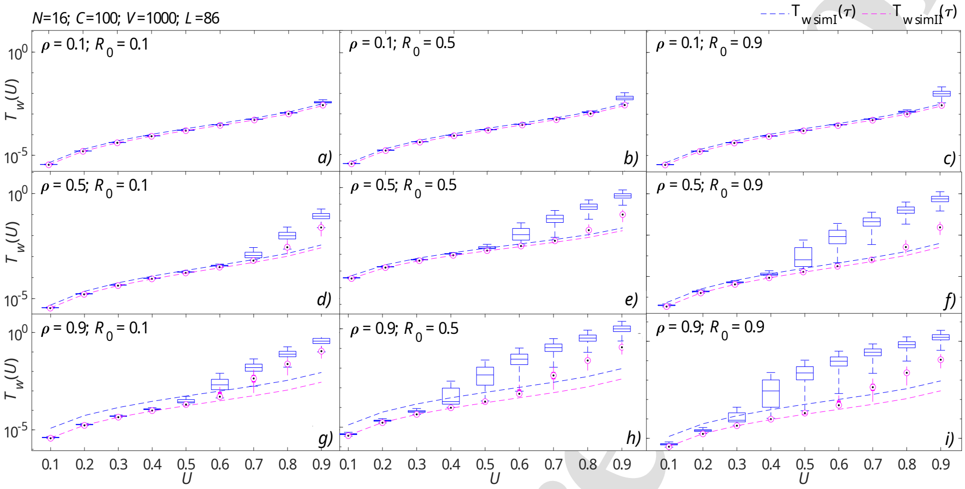


Figure 8: Similar to Fig. 7 but for a network with $N = 16$ active nodes with average intensity $C = 100$ per time unit, duration $L = 86$ time units, and statistics collected over $V = 1000$ random configurations.

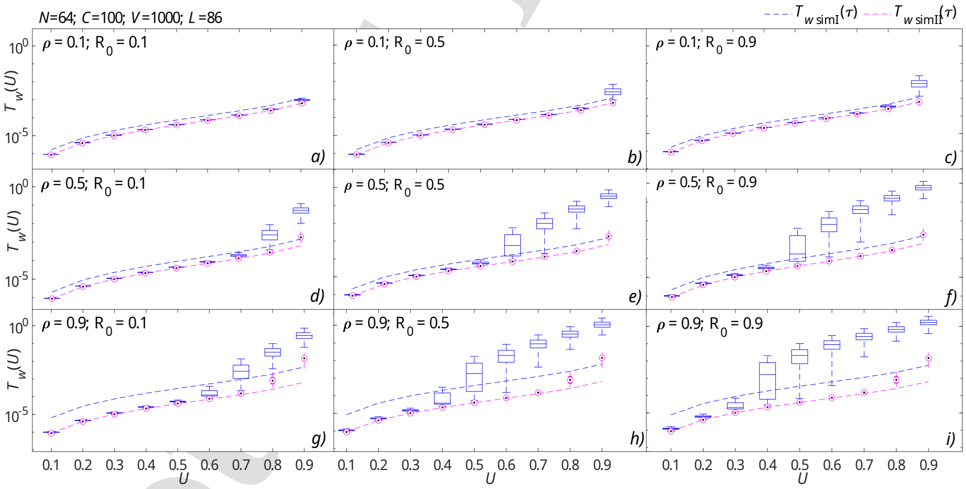


Figure 9: Similar to Fig. 7 but for a network with $N = 64$ active nodes with average intensity $C = 100$ per time unit, duration $L = 86$ time units, and statistics collected over $V = 1000$ random configurations.

multiple universities and research centers in Japan. We analyzed representative daily traffic records with second-resolution intensity and split it into segments of different time durations.

It is important to note that, in most practical scenarios, traffic variations are characterized by both auto- and cross-correlations. The impact of long-term autocorrelations has been investigated in deeper detail in our recent work [26]. In this work, the goal was rather to characterize the effect of cross-correlations in the mutual access patterns, and thus autocorrelations have been eliminated from the real-world observational record by random shuffling of the analyzed traffic intensities in time, while preserving their relative positions. For that, V random series of length L have been created and each of them sorted rankwisely to generate V independent index vectors. The respective vectors were next used as indices in the re-arrangement procedure that is being applied simultaneously to *all* N traffic intensity series in order to preserve cross-correlations between them, while a separate index vector was applied to each of the V independent configurations.

Figure 10 shows typical cross-correlation coefficients R_0 (upper panel) and coefficients of variation ρ (lower panel) of the intensities of the bidirectional (i.e., both originating from and/or addressed to) traffic patterns of several clusters of $N = 4$, 16, and 64 representative network nodes with high, middle, and low traffic intensities, respectively, for the empirical traffic data collected from a sample point on a backbone of the MAWI academic network. The figure shows that the correlation coefficients R_0 typically range between 0.3 and 0.9, while the coefficient of variation ρ ranges between 0.1 and 1.5, with short-term dynamics on scales up to several hours represented by $\rho < 1$, indicating that the above model parameters are relevant for a realistic network example at least *on the average.* In the following, we compared direct evaluations and queue simulations for these sub-networks with the above analytical solutions.

Figure 11 shows that the functional forms of the distributions are in most cases preserved when comparing simulated and empirical records, despite the fact that the respective coefficients of variation ρ and cross-correlation coefficients R_0 in the empirical traffic patterns agree with the model parameters and with the simulated data *on the average* only. Notable deviations from the simulation scenarios can be observed at rather large ρ and R_0 close to one, and only for selected simulation scenarios, that could be possibly attributed to the effects of heavy-tailed empirical distributions and nonlinear interactions which are not captured by linear cross-correlation analysis.

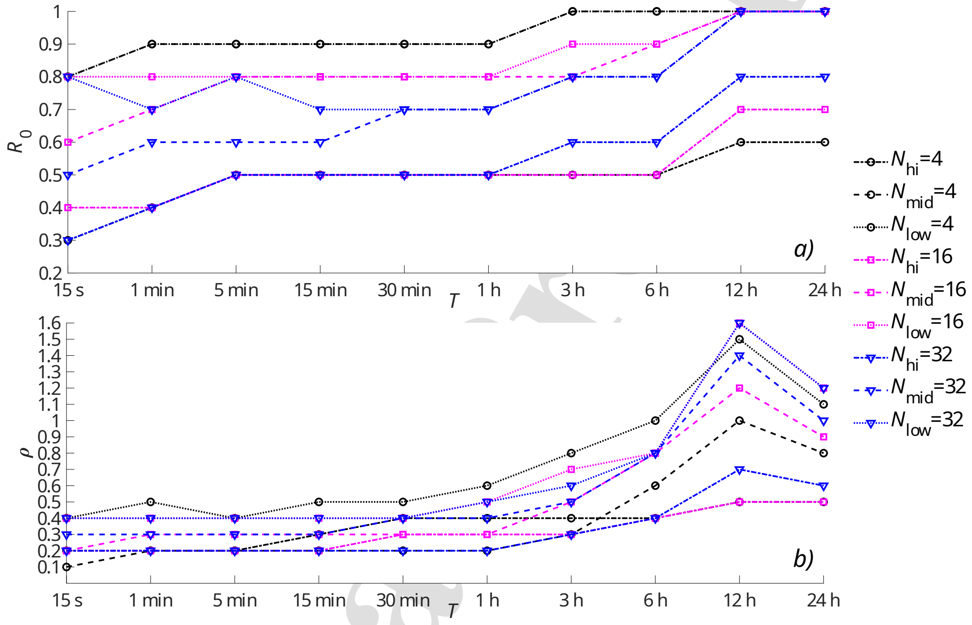


Figure 10: Cross-correlation coefficients R_0 (upper panel) and coefficients of variation ρ (lower panel) of the intensities of the bidirectional (i.e., both originating from and/or addressed to) traffic patterns of several clusters of $N = 4$, 16, and 64 representative network nodes with high, middle, and low traffic intensities, respectively, for the empirical traffic data evaluated over time segments of duration T. The values have been averaged over non-overlapping nodes and rounded to single-digit precision.

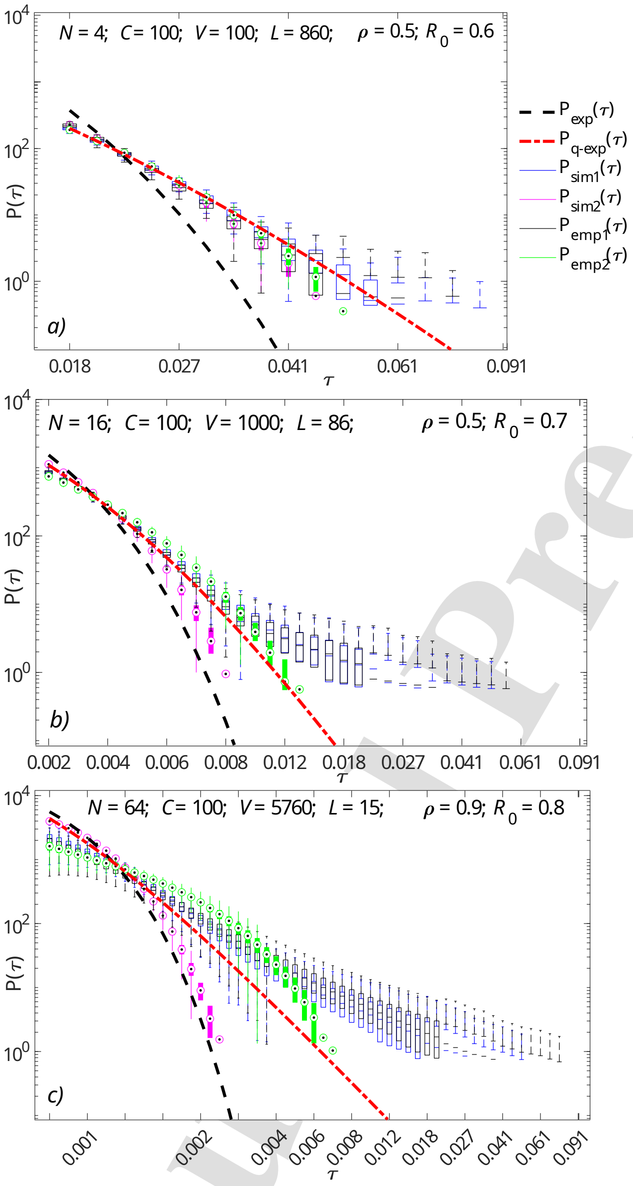


Figure 11: Inter-arrival time distributions $P(\tau)$ for the simulated and empirical traffic in a network with $N = 4$, 16, and 64 active nodes, respectively, based on the intensity patterns collected from a MAWI backbone network, after random re-arrangement of the intensity series to eliminate autocorrelations, while preserving cross-correlations in the traffic intensities. Open and full boxplots correspond to the two simulation variants, with intensity patterns either simulated (sim) or obtained from empirical records of the MAWI network (emp). The dashdot red curves correspond to the analytical approximations for ρ_τ according to Eq. 13, while the dashed black curves show a simple exponential with the same average inter-arrival interval.

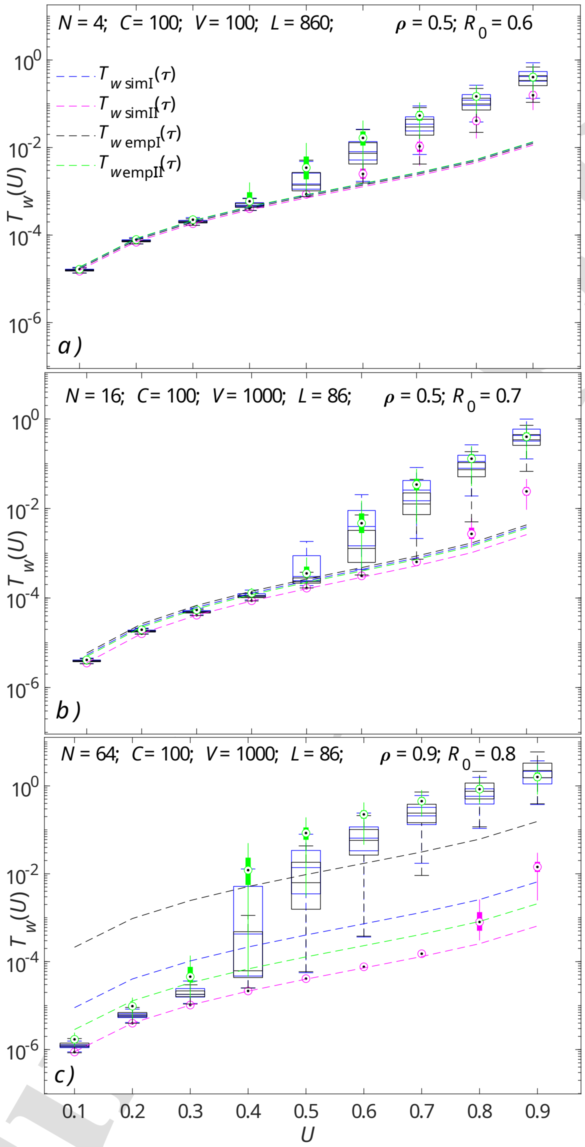


Figure 12: Representative examples of waiting times T_W as functions of the throughput utilization U for the simulated and empirical aggregated traffic in networks of $N = 4$, 16, and 64 active nodes, respectively, with statistics collected over $V = 100$ random intensity series re-arrangements. Open and full boxplots correspond to the results of two simulation variants, with intensity patterns either simulated (sim) or obtained from empirical records of the MAWI network (emp). Dashed lines correspond to the approximations using the Kingman's formula (10) [4] with respective analytical corrections to the coefficient of variation of the inter-arrival times ρ_τ^2 according to Eq. 14 depending on the cross-correlations of the intensities from different nodes, as determined by the analytical treatment.

Figure 12 shows that for the same representative combination of model parameters considered above, the proposed analytical approximations to the coefficients of variation of the inter-arrival times inserted into the Kingman's formula (10) [4] provide reasonable approximations to the waiting times T_W. Similarly as for the distributions, analytical approximations appear between the results obtained by respective queuing system simulation variants I and II affected by discreteness and finite size effect, respectively, indicating the validity and reasonable accuracy of the proposed analytical corrections that take into account cross-correlations in the dynamics of multiple nodes in the network.

7. Discussion

Public information and telecommunication, transportation and logistic, economic and financial systems are represented by complex networks, reflecting the complex hierarchy and interconnections of modern society. With multi-path connections available, public networks from specialized public services to the entire World Wide Web can no longer be described by simple Poisson-based models [8, 42, 43]. In addition to bursty patterns attributable to the erratic human dynamics [44, 45, 46], an increasing role of IoT devices with autonomous sensors and/or sensor networks generating their own traffic patterns leads to a complex superposition of multiple layers at a wide range of scales both in time and space. Accordingly, it is no surprise that two-compound models based on mixtures on convoluted distributions, retrospectively known as stochastic volatility [47], or as superstatistical in the theoretical postulations by C. Beck and E. Cohen [48, 49], also reminiscent of the so-called diffusing diffusivity in recent works [50], that represent traffic by conventional models only at very short scales, while exhibiting variable intensities over long scales [11, 12, 13], appeared more adequate in the above context. Our recent results indicate that evaluations based on superstatistical models drastically reduced waiting time underestimations from 3-4 decades to less than one decade [14, 26].

However, while the superstatistical models provide accurate evaluations of waiting times (and thus also of sojourn times or simply delays characterizing the time elapsed from request arrival to service completion), it rather provides an overall picture without digging deeper into the details of the corresponding network structure. While this could be sufficient for general evaluations, a better understanding of the laws governing the respective traffic intensity variations is of immense interest for the improvement of routing and other resource re-distribution algorithms.

In terms of perspective networks, which do not yet exist, and thus no direct measurements could be performed on them, it would be nice to have some rules to evaluate the delays, or vice versa, the throughput required to keep the majority of delays below a certain threshold, from a limited set of simple quantities, such as the number of active nodes (users and/or IoT devices), as well as some general quantities characterizing their activity and interactions. From a statistical perspective, traffic patterns from/to individual nodes are characterized by coefficients of variation ρ and persistence, which, given the long-term correlated nature of the traffic intensity, could be characterized by its Hurst exponent H. In turn, mutual dynamics are typically characterized by cross-correlation matrices $R = [r_{ij}]$, where r_{ij} are the pairwise cross-correlation coefficients between traffic intensities at nodes i and j, respectively.

In large networks, bottlenecks commonly occur at certain aggregation levels, with traffic intensity bursts contributed by simultaneous activity of multiple, in many cases hundreds or thousands of individual nodes. To reveal the exact locations of these bottlenecks, as well as gain a better understanding of the delays emerging at these bottlenecks, it is often sufficient to connect them with the above quantities at least *on the average*. With this simplification, the model could be easily parameterized even for large-scale networks: the (weighted) average coefficient of variation ρ for *all* contributing to the aggregated traffic in the samplepoint of interest, the average Hurst exponent H, and the average r_{ij} over all $i \neq j$ (i.e., except the diagonal, where by definition $r_{ij} = 1$ for $i = j$).

In this work, we have shown that an analytical approximation originally derived under the assumption of the equicorrelated traffic intensity patterns, also demonstrated reasonable accuracy when the exact correlations between multiple nodes are substituted by the averages, supported by both computer simulations and large-batch analysis of real-world empirical teletraffic collected from the backbone of a large academic network. We have shown that the proposed analytical models provide with approximations that typically occupy an intermediate position between two simulation variants, none of which is accurate due to inevitable discreteness and finite-size effects both in the simulation and real-world data analysis scenarios.

The role of computational model discreteness and finite-size effects deserves a separate note here. When dealing with various natural science problems at macroscales, we often assume that the system under investigation is rather continuous, or possible discreteness arises at the quantum level, which is many decades beyond the scale range of interest. In the context of teletraffic analysis, the network under study is a discrete system *per se*, with all traffic elements adjusted to its own timing grid, resulting in *de facto* minimal inter-arrival times. The discrete nature of network traffic is most prominent at the transport level, where all information is transmitted in the form of short packets. In many cases, discreteness expands also to higher levels: for the end user interested in downloading a specific file, only the entire downloaded file means a closed request, while there are ubiquitous exceptions, such as audio and video broadcast services.

In our approach, we took into account a single and most universal transport-level discreteness by limiting the inter-arrival time distributions from below at a fixed threshold and re-normalizing them accordingly. However, we like to note, that there is another source of discreteness arising from the traffic intensity taken with some fixed, in our case one-second time resolution. When a superstatistical model is employed in the traffic simulations given its intensity variations, 1 s fragments are assumed to be stationary, which may be in most cases sufficient for human-driven traffic patterns, while far beyond the response times of autonomous IoT devices and/or online robots. Moreover, this modeling strategy results in spurious short-term persistence (autocorrelations) at scales up to 1 s, which may be important at busy nodes (characterized by large numbers of transmitted packets per second). In contrast, reducing the time interval of stationary traffic may lead to additional discreteness effects at less busy nodes, since the number of packets is always an integer. Finally, at the opposite end of the distribution, finite size effects are dominating in the model limitations. Notably, like the discreteness effects, finite size is similarly common for real-world networks, where delays beyond a certain bar typically result in the so-called exits, when the user decides that the service is likely down at the moment, and thus no longer waits for the response. Given the inherently discrete and finite nature of both computer-simulated and observational traffic records described above, the idea behind the two simulation variants suggested in this work was to employ them as the lower and upper boundaries for the analytical solution. In turn, long-term dynamics is also characterized by periodic (e.g., daily and/or weekly) patterns generating additional long-term persistence effects that could be taken into account in the framework of the superstatistical concept, as shown in deeper detail in our recent work [26].

8. Conclusion

To summarize, we have analyzed waiting times in queuing systems where traffic patterns are initiated by multiple nodes with variable intensities that are additionally characterized by cross-correlated access dynamics. We have shown explicitly how the variability of inter-arrival times in the aggregated traffic from multiple nodes could be obtained from the variability of the traffic intensity at separate nodes and the cross-correlation coefficient between them. Based on these calculations, we have suggested an analytical correction to the conventional stationary queue model given by the Kingman's formula. Using the proposed correction, one can evaluate the coefficients of variation that characterize the aggregated traffic pattern, and thus also estimate the queuing system performance, even in the absence of empirical traffic data. We believe that the latter is useful for evaluating the characteristics and optimizing the design of perspective networks that have not been yet physically implemented from a very basic set of initial parameters that are also scalable. Moreover, since modern networks are largely software defined, including real-time control over resource sharing and traffic routing, the above estimates could be applied for the selection of the most preferential network designs and routing algorithms by comparing several potential implementation scenarios, as well as their optimization aiming at the minimization of delays based solely on analytical evaluations, without running large-batch computer simulations.

We have shown explicitly that the proposed analytical approximation for the waiting times remains in good agreement with the corresponding statistics obtained by direct computer simulations of queuing systems, as well as through the analysis of large-batch empirical traffic intensity data from a backbone of a major academic network, further supporting the validity of the analytical treatment. Moreover, using observational traffic analysis, we have shown explicitly that the proposed analytical results lead to reasonable accuracy approximations also when the respective model parameters (such as traffic intensities and their variances at individual nodes, as well as their cross-correlation coefficients) are valid only *on the average*, indicating their applicability to a considerably broader combination of initial conditions that are generally more relevant to real-world scenarios.

Of note, although the effects of long-term persistence typically appear considerably more pronounced than the effects of cross-correlations, in real-world scenarios these properties are typically combined. More specifically, when long-term persistent traffic patterns at individual nodes are strongly cross-correlated, further superimposed by heavy-tailed intensity distributions, their interplay would often further magnify the bursts in the aggregated traffic. While all possible scenarios could hardly be described in a single, compact, and universal analytical treatment, which appears to be an obvious limitation of the current approach, reasonable approximations are nevertheless useful to localize potential variants of interest at the network design stage when no real observational data could be obtained, and thus decisions have to be made based on a very limited number of *a priory* assumptions.

Altogether, we believe that the proposed approach, in combination with recently characterized effects of long-term autocorrelations [26], would be useful for the quantification of queuing system performance in various applications arising in connection with information and telecommunication, transportation and logistic, as well as economic, financial and other real-world complex systems where bursty intensity patterns are governed by a complex interplay of human erratic activity with an increasing contribution of autonomous agents such as IoT devices, online sensors, trading robots, etc. As an outlook, for complex systems with heavy-tailed traffic intensity variations represented by geometric Brownian motion [51, 52], the proposed approximations could be applied to the logarithms of the intensity patterns, followed by an exponentiation in the last step of the analysis to obtain the resulting queue length and waiting times, respectively. Moreover, due to the striking similarities between statistical properties of various complex systems [53, 54, 31, 55], we believe that a similar analytical approach is also applicable to many other engineering and natural systems, for example, for the evaluation of river flows and water accumulation in reservoirs based on the rainfall dynamics and flow data from multiple tributaries in the upstream basin.

Acknowledgements

This research has been supported by the Ministry of Science and Higher Education (assignment No. FSEE-2020-0002).

References

[1] A. K. Erlang, Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges, Elektrotkeknikeren 13

(1917) 5–13.

[2] F. Pollaczek, Über Eine Aufgabe der Wahrscheinlichkeitstheorie. I, Mathematische Zeitschrift 32 (1) (1930) 64–100.

[3] A. Y. Khintchine, Mathematical theory of stationary queues, Matem. Sbornik 39 (1932) 73–84.

[4] J. Kingman, The single server queue in heavy traffic, in: Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 57, Cambridge Univ Press, 1961, pp. 902–904.

[5] W. G. Marchal, An approximate formula for waiting time in single server queues, AIIE transactions 8 (4) (1976) 473–474.

[6] W. Krämer, M. Langenbach-Belz, Approximate Formulae for the Delay in the Queueing System GI/G/l, Congressbook, 8th ITC, Melbourne 235 (1) (1976) 1–8.

[7] W. E. Leland, M. S. Taqqu, W. Willinger, D. V. Wilson, On the self-similar nature of Ethernet traffic (extended version), IEEE/ACM Transactions on networking 2 (1) (1994) 1–15.

[8] V. Paxson, S. Floyd, Wide area traffic: the failure of Poisson modeling, IEEE/ACM Transactions on Networking (ToN) 3 (3) (1995) 226–244.

[9] A. Feldmann, A. C. Gilbert, W. Willinger, T. G. Kurtz, The changing nature of network traffic: Scaling phenomena, ACM SIGCOMM Computer Communication Review 28 (2) (1998) 5–29.

[10] K. Park, W. Willinger, Self-similar network traffic and performance evaluation, Wiley Online Library, 2000.

[11] A. Tamazian, V. Nguyen, O. Markelov, M. Bogachev, Universal model for collective access patterns in the Internet traffic dynamics: A superstatistical approach, EPL (Europhysics Letters) 115 (1) (2016) 10008.

[12] O. Markelov, V. N. Duc, M. Bogachev, Statistical modeling of the Internet traffic dynamics: To which extent do we need long-term correlations?, Physica A: Statistical Mechanics and its Applications 485 (2017) 48–60.

[13] V. Nguyen, O. Markelov, A. Serdyuk, A. Vasenev, M. Bogachev, Universal rank-size statistics in network traffic: Modeling collective access patterns by Zipf's law with long-term correlations, EPL (Europhysics Letters) 123 (5) (2018) 50001.

[14] M. Bogachev, N. Pyko, S. Pyko, A. Vasenev, Service delays in strongly linked network communities, in: Journal of physics: Conference series, Vol. 1352, IOP Publishing, 2019, p. 012006.

[15] Y. Liu, W. Whitt, Stabilizing performance in networks of queues with time-varying arrival rates, Probability in the Engineering and Informational Sciences 28 (4) (2014) 419–449.

[16] J. Pender, R. H. Rand, E. Wesson, An analysis of queues with delayed information and time-varying arrival rates, Nonlinear Dynamics 91 (4) (2018) 2411–2427.

[17] W. Whitt, Time-varying queues, Queueing models and service management 1 (2) (2018).

[18] A. Dudin, V. I. Klimenok, V. M. Vishnevsky, The theory of queuing systems with correlated flows, Vol. 430, Springer, 2020.

[19] J. Zhang, T. T. Lee, T. Ye, L. Huang, An approximate mean queue length formula for queueing systems with varying service rate, Journal of Industrial & Management Optimization 17 (1) (2021) 185.

[20] M. Bogachev, J. Eichner, A. Bunde, The effects of multifractality on the statistics of return intervals, The European Physical Journal Special Topics 161 (1) (2008) 181–193.

[21] M. I. Bogachev, J. F. Eichner, A. Bunde, On the occurence of extreme events in long-term correlated and multifractal data sets, in: Earth Sciences and Mathematics, Springer, 2008, pp. 1195–1207.

[22] B. Podobnik, H. E. Stanley, Detrended cross-correlation analysis: A new method for analyzing two nonstationary time series, Physical review letters 100 (8) (2008) 084102.

[23] X.-Y. Qian, Y.-M. Liu, Z.-Q. Jiang, B. Podobnik, W.-X. Zhou, H. E. Stanley, Detrended partial cross-correlation analysis of two nonstationary time series influenced by common external forces, Physical Review E 91 (6) (2015) 062816.

[24] N. Yuan, Z. Fu, H. Zhang, L. Piao, E. Xoplaki, J. Luterbacher, Detrended partial-cross-correlation analysis: A new method for analyzing correlations in complex system, Scientific reports 5 (1) (2015) 1–7.

[25] M. Bogachev, A. Bunde, On the occurrence and predictability of overloads in telecommunication networks, EPL (Europhysics Letters) 86 (6) (2009) 66002.

[26] M. I. Bogachev, A. V. Kuzmenko, O. A. Markelov, N. S. Pyko, S. A. Pyko, Approximate waiting times for queuing systems with variable long-term correlated arrival rates, Physica A: Statistical Mechanics and its Applications 614 (2023) 128513.

[27] D. G. Kendall, Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain, The Annals of Mathematical Statistics (1953) 338–354.

[28] D. Vorobyova, A. Muthanna, A. Paramonov, O. A. Markelov, A. Koucheryavy, G. Ali, M. ElAffendi, A. A. Abd El-Latif, Iot network model with multimodal node distribution and data-collecting mechanism using mobile clustering nodes, Electronics 12 (6) (2023) 1410.

[29] J. D. Little, A proof for the queuing formula: L= λ w, Operations research 9 (3) (1961) 383–387.

[30] R. M. Oliver, An alternate derivation of the pollaczek-khintchine formula, Operations Research 12 (1) (1964) 158–159.

[31] G. C. Yalcin, P. Rabassa, C. Beck, Extreme event statistics of daily rainfall: Dynamical systems approach, Journal of Physics A: Mathematical and Theoretical 49 (15) (2016) 154001.

[32] J. Weber, M. Reyers, C. Beck, M. Timme, J. G. Pinto, D. Witthaut, B. Schäfer, Wind power persistence characterized by superstatistics, Scientific reports 9 (1) (2019) 1–15.

[33] C. Tsallis, Possible generalization of Boltzmann-Gibbs statistics, Journal of statistical physics 52 (1-2) (1988) 479–487.

[34] C. Tsallis, et al., Introduction to nonextensive statistical mechanics, Springer, 2009.

[35] J. Ludescher, C. Tsallis, A. Bunde, Universal behaviour of interoccurrence times between losses in financial markets: An analytical description, EPL (Europhysics Letters) 95 (6) (2011) 68002.

[36] C. G. Antonopoulos, G. Michas, F. Vallianatos, T. Bountis, Evidence of q-exponential statistics in Greek seismicity, Physica A: Statistical Mechanics and its Applications 409 (2014) 71–77.

[37] M. I. Bogachev, A. R. Kayumov, A. Bunde, Universal internucleotide statistics in full genomes: a footprint of the DNA structure and packaging?, PloS one 9 (12) (2014) e112534.

[38] J. Ludescher, A. Bunde, Universal behavior of the interoccurrence times between losses in financial markets: Independence of the time resolution, Physical Review E 90 (6) (2014) 062809.

[39] C. Tsallis, Inter-occurrence times and universal laws in finance, earthquakes and genomes, Chaos, Solitons & Fractals 88 (2016) 254–266.

[40] M. I. Bogachev, O. A. Markelov, A. R. Kayumov, A. Bunde, Superstatistical model of bacterial DNA architecture, Scientific reports 7 (1) (2017) 1–12.

[41] J. E. Cohen, Sum of a random number of correlated random variables that depend on the number of summands, The American Statistician 73 (1) (2019) 56–60.

[42] G. Faÿ, B. González-Arévalo, T. Mikosch, G. Samorodnitsky, Modeling teletraffic arrivals by a poisson cluster process, Queueing Systems 54 (2006) 121–140.

[43] I. D. Moscholios, M. D. Logothetis, Efficient multirate teletraffic loss models beyond Erlang, John Wiley & Sons, 2019.

[44] A. Erramilli, M. Roughan, D. Veitch, W. Willinger, Self-similar traffic and network dynamics, Proceedings of the IEEE 90 (5) (2002) 800–819.

[45] J. Mathiesen, L. Angheluta, P. T. Ahlgren, M. H. Jensen, Excitable human dynamics driven by extrinsic events in massive communities, Proceedings of the National Academy of Sciences 110 (43) (2013) 17259–17262.

[46] S. Chen, J. Wu, Y. Pan, J. Ge, Z. Huang, Simulation and case study on residential stochastic energy use behaviors based on human dynamics, Energy and Buildings 223 (2020) 110182.

[47] S. J. Taylor, Modeling stochastic volatility: A review and comparative study, Mathematical finance 4 (2) (1994) 183–204.

[48] C. Beck, Dynamical foundations of nonextensive statistical mechanics, Physical Review Letters 87 (18) (2001) 180601.

[49] C. Beck, E. Cohen, Superstatistics, Physica A: Statistical mechanics and its applications 322 (2003) 267–275.

[50] W. Wang, A. G. Cherstvy, A. V. Chechkin, S. Thapa, F. Seno, X. Liu, R. Metzler, Fractional brownian motion with random diffusivity: emerging residual nonergodicity below the correlation time, Journal of Physics A: Mathematical and Theoretical 53 (47) (2020) 474001.

[51] D. Vinod, A. G. Cherstvy, W. Wang, R. Metzler, I. M. Sokolov, Nonergodicity of reset geometric brownian motion, Physical Review E 105 (1) (2022) L012106.

[52] D. Vinod, A. G. Cherstvy, R. Metzler, I. M. Sokolov, Time-averaging and nonergodicity of reset geometric brownian motion with drift, Physical Review E 106 (3) (2022) 034137.

[53] M. Bogachev, A. Bunde, Universality in the precipitation and river runoff, EPL (Europhysics Letters) 97 (4) (2012) 48011.

[54] A. Bunde, M. I. Bogachev, S. Lennartz, Precipitation and river flow: Long-term memory and predictability of extreme events (2012).

[55] F. Serinaldi, F. Lombardo, C. G. Kilsby, All in order: Distribution of serially correlated order statistics with applications to hydrological extremes, Advances in Water Resources 144 (2020) 103686.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: