mathematics

MDPI

Article

Queueing Network with Moving Servers as a Model of Car Sharing Systems

Chesoong Kim [1,*], Sergei Dudin [2,3] and Olga Dudina [2,3]

[1] Department of Industrial Engineering, Sangji University, Wonju, Kangwon 26339, Korea
[2] Department of Applied Mathematics and Computer Science, Belarusian State University, 4 Nezavisimosti Ave., Minsk 220030, Belarus
[3] Applied Mathematics and Communications Technology Institute, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow 117198, Russia
* Correspondence: dowoo@sangji.ac.kr

Received: 13 August 2019; Accepted: 3 September 2019; Published: 6 September 2019

Abstract: We consider a queueing network with a finite number of nodes and servers moving between the nodes as a model of car sharing. The arrival process of customers to various nodes is defined by a marked Markovian arrival process. The customer that arrives at a certain node when there is no idle server (car) is lost. Otherwise, he/she is able to start the service. With known probability, which depends on the node and the number of available cars, this customer can balk the service and leave the system. The service time of a customer has an exponential distribution. Location of the server in the network after service completion is random with the known probability distribution. The behaviour of the network is described by a multi-dimensional continuous-time Markov chain. The generator of this chain is derived which allows us to compute the stationary distribution of the network states. The formulas for computing the key performance indicators of the system are given. Numerical results are presented. They characterize the dependence of some performance measures of the network and the nodes on the total number of cars (fleet size of the car sharing system) and correlation in the arrival process.

Keywords: queueing network; moving servers; car sharing; marked Markovian arrival process

1. Introduction

During the past few years, car sharing services have quickly developed in many countries for client transportation, especially in urban areas where there are many potential clients who are ready to pay for the private mobility via the short-term use of a vehicle on a per trip basis. The use of well managed car sharing services is profitable for individuals. They have enough mobility without buying or leasing of expensive cars, its maintenance, possible repair, refueling, parking, paying taxes and insurance, etc. The social profit is provided via more efficient use of parking places, the increase in the throughput of the roads and a decrease in the probability of traffic jam, reduction of carbon emission, etc. Car sharing services may be a profitable business if they are well built and managed. One of the most important problems, which has to be resolved while starting or developing car sharing service is to determine the number of required cars (fleet size). The problem of fleet management is quite challenging and complicated. On the one hand, it is too costly to use (buy or lease) many cars, due to relatively high cost, wearout, maintenance, etc. On the other hand, if the number of cars is insufficient, the probability that a customer will not obtain access to the service on demand will be high. Customers can cancel membership in the company providing this service and use service of another car sharing or car-renting company. This, in turn, can essentially reduce the potential income of the car sharing provider. The importance and profitability of car sharing services made it popular both in real life and scientific literature. A recent survey and the detailed classification of the research in the field of car sharing are given, e.g., in the paper [1]. It may be noted that, following classification of [1], we consider in the present paper "free-floating mode", which is now a popular model in the literature. This means that the cars are freely parked in public spaces within the operational area. The journey can start and finish at any point in this area. As recent papers in the field of car sharing, we can also mention [2,3].

The theory of queueing networks seems to be an adequate tool for analysis of car sharing systems, see [1]. Application of this theory is described, e.g., in [4,5] and references therein. In [4,5], the dynamics of the car sharing system are described by a closed queueing network where the cars are interpreted as the customers which are served by the servers (arriving clients). The analysis of the networks implemented in [4,5] is based on mean value analysis or on the assumption about the existence of the product form solution. This analysis is approximate.

Aiming to create a simple analytical model of a car sharing system, we consider this system as an open queueing network. Customers are associated with clients. Servers correspond to the cars. Each idle car can be located in a corresponding node (zone of the operational area). If the car is busy, its location in the network is temporarily undefined for customers because they cannot observe it. In this paper, we present an exact analysis of the constructed model of a car sharing system.

The study of queueing systems and networks with moving servers has not received proper attention due to the mathematical complexity of such systems. As the simplest examples of such systems (with the exception of the trivial examples of systems with servers vacations, unreliable systems, and polling systems, in which a server sequentially connects to the existing queues according to a certain schedule), we can note the following two systems. One of them is the tandem system with moving servers that are dynamically redistributed between stations. A part of the available servers is permanently assigned to tandem stations, and then the remaining part is dynamically redistributed between stations depending on the ratio of the number of users present at the corresponding station. For references to the results of the study of systems of this type, see, for example, [6,7]. Another example is the system considered in [8]. This system is in fact a natural generalization of the classic polling queueing systems, which are considered in a large number of papers and books, for example, a recent survey [9], due to their wide applicability in the analysis of various multiple access systems, including urban wireless networks, to the case of several servers. In [8], it is assumed that the customers arrive at service systems located at the vertices of a graph. Service of the users in this system is carried out by a finite number of servers. Each of the servers provides service to one user at the vertex, in which it is located, and then moves to another vertex arbitrarily, at which currently there is no server. We can also mention the works [10,11]. The servers are distributed over the nodes of the network and from time to time they interchange the nodes, taking the existing queues of users with them.

Therefore, the main contribution of our paper is the construction of the model of operation of car-sharing systems as the semi-open queueing network with the random transition of the **servers** between the nodes of the network. Such queueing networks are not considered in the existing literature and our results represent significant contribution to the theory of queueing networks with arbitrary topology and moving servers. Essential technical difficulty in analysis of such networks consists of the complexity of the multidimensional Markov chain describing the behavior of the network. We successfully overcome this difficulty only due to the recent experience in analysis of the semi-open queueing networks without the movement of the servers via the analysis of Markov chains with structured generator, see [12,13].

An essential advantage of the queueing network considered in our paper, compared to the overwhelming majority of existing queueing networks literature, is the consideration of a quite general, marked Markovian arrival process ($MMAP$) of customers that allows us to take into account random fluctuations of the intensity of customers arrival in various nodes. Such fluctuations (e.g., due to the increase in activity of customers at some periods of the day: morning, the time before and after lunch, time after the end of the working day, late evening, etc.) take place in the overwhelming majority of real car sharing systems. At the same time, the majority of the research is implemented under the assumption that the arrivals occur according to a stationary Poisson process that has a constant arrival rate. Quite a short list of papers dealing with queueing networks with Markovian arrival process (not relating with car sharing) can be found in recent papers by [12,14] and references therein. A possibility of transition of servers between the nodes is not considered in these papers. An essential novelty of the analysis in this paper is that the queueing network takes into account servers mobility and their unavailability at any node during the service process. It is also worth noting that we allow the scenario when the arriving customer meets idle servers in the target node, but he/she balks the network with a certain probability, which depends on the node and the number of available cars. Such a situation is typical in real car sharing systems because the client can refuse to go, e.g., because his/her walking distance to the nearest car seems to be too long or the available car is not suitable for him/her.

The goal of the analysis implemented in this paper is the computation of the performance measures of the system under the fixed number of cars N and estimation of possible variation of these measures when the number N is changed. The results of this analysis can be used for the choice of the optimal value of N. In addition, these results can be helpful in answering the question: whether or not the existing distribution of the nodes of staring and finishing journey is satisfactory or it has to be somehow varied. Such variation seems to be possible, e.g., via differentiation of the tariffs for the use of cars depending on the time and the nodes of staring and finishing the journey. To implement this analysis, we model the operation of the car sharing system in terms of a queueing network as described in the next section.

In the following section, we build a mathematical model of a car sharing system as an open queueing network with servers moving between the nodes of the network.

2. Mathematical Model

Let us assume that the operational area of a car sharing system is divided into K zones. There exists statistics (or expert estimation) about:

- the pattern of the flow of potential users of the car sharing system, including the average arrival rates to each zone at different periods of a day and night;
- the number N of cars in the car sharing system;
- the average duration of a trip (journey);
- the proportion of zones at which the journeys finish;
- the proportion of users, which had the intention to use a car at some zone, but then balk, despite availability of a definite number of cars.

The client of a car sharing system is considered as a customer which receives service in a queueing network. The queueing network consists of K nodes. The total number of servers (cars) in the network is equal to N. The servers can move and change their location. The location of a busy server is not monitored. The number of idle servers in each node is observed. It is random and can vary in the interval $[0, N]$. This number can be changed at any instant of the start or finish of the trip of a client. The sum of the numbers of idle servers in the nodes also can vary in the interval $[0, N]$.

We distinguish the arriving customers by the type according to the node at which a customer appears. Namely, a type-k customer arrives at the k-th node of the network, $k = \overline{1, K}$. The notations like $k = \overline{1, K}$ mean that the integer parameter k admits the values in the range $\{1, \dots, K\}$. The arrival process is assumed to be defined by the $MMAP$ (Marked Markovian Arrival Process), see [15,16]. The possible customer's arrival moments in the $MMAP$ coincide with the moments of the jumps of an irreducible continuous-time Markov chain ν_t, $t \geq 0$, with a finite state space $\{1, 2, ..., W\}$. This chain is called as the underlying process of the $MMAP$. The $MMAP$ is defined by the set of square matrices D_0, D_k, $k = \overline{1, K}$, of size W. The transition intensities of the chain ν_t, which are accompanied by an arrival of a type-k customer, $k = \overline{1,K}$, are defined by the entries of the matrix D_k. The matrix $D(1) = \sum\limits_{k=0}^{K} D_k$ is the infinitesimal generator of the Markov chain ν_t.

Formulas for computing such characteristics of the $MMAP$ as the average intensity of customer arrival (fundamental rate) λ, the average intensity λ_k of the arrival of type-k customers, the squared coefficient of variation c_{var} and the coefficient of correlation c_{cor} of two successive intervals between arrivals can be found, e.g., in [15,17]. The main advantages of the $MMAP$ over the stationary Poisson arrival process are its abilities to catch, besides the average arrival rate, possible variation of instantaneous arrival rates, different variance of inter-arrival times and their dependence. In application of the theoretical results obtained for the systems with the $MMAP$ to analysis of concrete systems, it is necessary to construct the set of matrices D_0, D_k, $k = \overline{1,K}$, which define the $MMAP$, based on observation of the traces of the flow in the concrete system. In particular, coincidence of the mean arrival rates of the real flow with the rates of the constructed $MMAP$ as well as of coefficients of variation and correlation is required. The problem of constructing the matrices D_0, D_k, $k = \overline{1,K}$, is not easy. However, this problem is already well addressed in the existing literature, see, e.g., [18,19].

An illustration of the distribution of arriving customers among the nodes is given in Figure 1. The circles within the nodes show the currently available servers at the nodes of the network. In this particular figure, node 2 does not have free servers at the given moment.

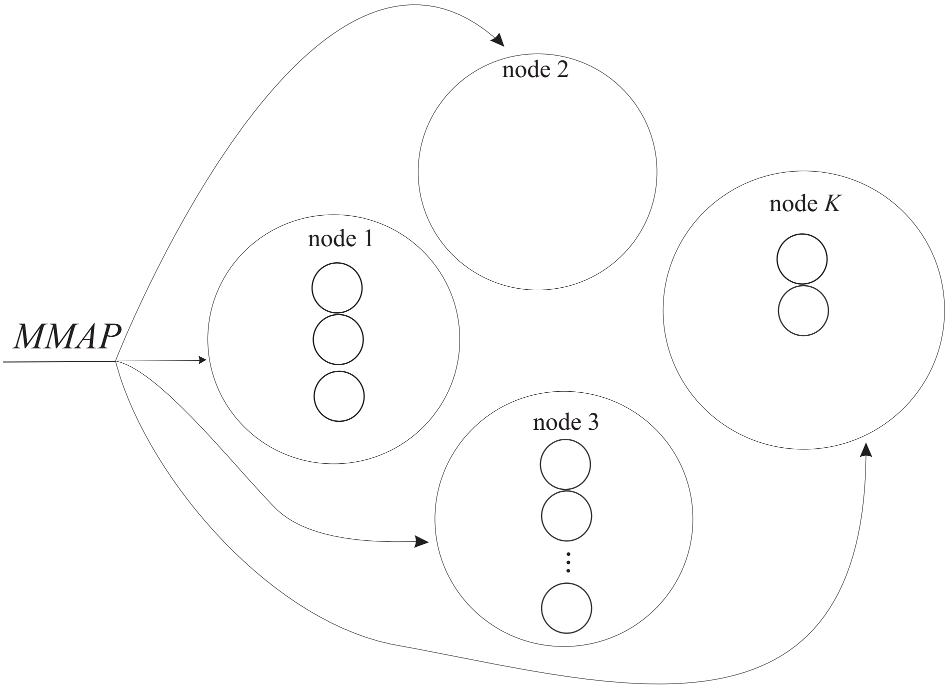

Figure 1. Scheme of arriving customers distribution in the queueing network under study.

If an arriving type-k customer does not meet idle servers at node k, the customer leaves the network permanently. If a customer arrives at node k when there are n free servers in this node, he/she starts service at one of the free servers with probability $h_{n,k}$ or balks with the complementary probability (e.g., due to the long walking distance to the nearest car). The service time of a customer has an exponential distribution with the parameter μ. During the service of a customer, the server is temporarily cancelled from the list of the available servers in the network. After the service time expires, the serviced customer leaves the network permanently while the server becomes available in the k-th node, $k = \overline{1,K}$, with the probability q_k, $\sum\limits_{k=1}^{K} q_k = 1$.

For reader convenience, we summarise the main notation that characterizes the system in Table 1.

Table 1. Notation.

K	the number of nodes (zones)
N	the total number of servers (cars)
W	the number of states of the underlying process of the $MMAP$ (marked Markovian arrival process) arrival flow of customers
$D_k, k = \overline{0,K},$	the square matrices of size W that characterize the $MMAP$ arrival flow of customers
λ	the average arrival intensity of customers
$\lambda_k, k = \overline{1,K},$	the average arrival intensity of type-k customers
μ	the parameter of exponential distribution of the service time
$h_{n,k}$	the probability that an arriving customer starts service at one of the free servers if he/she arrives at node k when there are n free servers in this node
H	the matrix $H = (h_{n,k})_{n=\overline{1,N},k=\overline{1,K}}$, of size $N \times K$ consisting of the probabilities $h_{n,k}$
q_k	the probability that after service completion a server becomes available in the k-th node, $k = \overline{1,K}$
$\mathbf{q}$	the vector consisting of the probabilities $(q_1, \ldots, q_K)$
I	the identity matrix
O	a zero matrix
$\otimes$	the symbol of the Kronecker product of matrices, see [20]
$\mathrm{diag}\{\ldots\}$	the diagonal matrix with the diagonal entries defined by the entries of the vector given in the brackets
$\mathbf{e}$	the column vector $(1, \ldots, 1)^T$ of an appropriate size
$\mathbf{0}$	a zero row vector of an appropriate size
$\tilde{H}_k$	the matrix of size $N \times K$ with all zero entries except the k-th column, whose entries are equal to the corresponding entries of the k-th column of the matrix H.

In the next sections, we analyse the described queueing network. In Section 3, we construct a multi-dimensional continuous-time Markov chain that describes the operation of this queueing network. Usually, there exist many ways for construction of such a chain and it is necessary to find a way leading to the good structure of the generator of the chain and the minimally possible size of the blocks of the generator. The generator of the constructed Markov chain is presented and a brief explanation of the intuitive meaning of its blocks is given. After derivation of the explicit expressions for the generator and its blocks, we briefly touch the problem of computation of the stationary distribution of the Markov chain. In Section 4, we present the expressions for the key performance indicators of the network, including the availability of the servers at each node. Then, illustrative numerical examples are presented in Section 5.

3. Process of the Network States

It is easy to see that the operation of the queueing network under study can be described in terms of the following multi-dimensional continuous-time Markov chain

$$\xi_t = \{n_t, \nu_t, n_t^{(1)}, \ldots, n_t^{(K)}\},\ t \geq 0,$$

where, at the moment $t,\ t \geq 0$,

- n_t is the number of free servers in the network, $n_t = \overline{0,N}$;
- ν_t is the state of the underlying process of the $MMAP$, $\nu_t = \overline{1,W}$;

- $n_t^{(k)}$ is the number of free servers in the k-th node, $k=\overline{1,K}$, $n_t^{(k)}=\overline{0,n_t}$, $\sum\limits_{k=1}^{K} n_t^{(k)}=n_t$,

which is regular and irreducible.

To analyse the Markov chain ξ_t, we will combine the states of the process ξ_t having the value n of the component n_t into the set of states called as the level n. Inside the level, we will enumerate the states of the process ξ_t in the direct lexicographic order of the component ν_t and the reverse lexicographic order of the components $\{n_t^{(1)},\ldots,n_t^{(K)}\}$. We denote $\mathbf{G}_{n,n'}$ as the matrix consisting of the intensities of transitions from the level n to the level n' and $\mathbf{G}$ as the block matrix having the blocks $\mathbf{G}_{n,n'}$.

In derivation of the expressions for the blocks $\mathbf{G}_{n,n'}$, of the generator of the process $\xi_t=\{n_t,\nu_t,n_t^{(1)},\ldots,n_t^{(K)}\}$ defining the intensities of transition from the states, which belong to the level n, to the states, which belong to the level n', the most technically difficult step is the computation of the matrices that define the transition probabilities of the components $\{n_t^{(1)},\ldots,n_t^{(K)}\}$ at the moments of service finishing and beginning. It can be verified that at the moments when $\sum\limits_{k=1}^{K} n_t^{(k)}=n$ the state space of these components consists of $T_n=\binom{n+K-1}{K-1}=\frac{(n+K-1)!}{n!(K-1)!}$ elements, $n=\overline{1,N}$. To implement this step, we have proven the following two Lemmas.

Lemma 1. *Let the matrices $C_n=C_n(\mathbf{q})$, $n=\overline{0,N-1}$ define the transition probabilities of the process $\{n_t^{(1)},\ldots,n_t^{(K)}\}$, $t\geq 0$, at the moment when service of a customer is finished and the number of free servers was equal to n. These matrices can be found as $C_0=\mathbf{q}$, $C_n=C_n^{(K-2)}$, $n=\overline{1,N-1}$, where the matrices $C_n^{(k)}$ of block size $(n+1)\times(n+2)$ define the transition probabilities of the components $n_t^{(K)},\ldots,n_t^{(K-k-1)}$ at the moment of service completion when there are n free servers conditional on the fact that the server that finished service becomes available in the node with the number from the set $K,K-1,\ldots,K-k-1$, $k=\overline{0,K-2}$. These matrices can be computed by the recursion*

$$C_n^{(0)}=\begin{pmatrix} q_{K-1} & q_K & 0 & \cdots & 0 & 0\\ 0 & q_{K-1} & q_K & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & q_{K-1} & q_K \end{pmatrix},$$

$$C_n^{(k)}=\begin{pmatrix} q_{K-k-1} & \tilde{\mathbf{q}}^{(k)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0}\\ \mathbf{0}^T & q_{K-k-1}I & C_1^{(k-1)} & O & \cdots & O & O\\ \mathbf{0}^T & O & q_{K-k-1}I & C_2^{(k-1)} & \cdots & O & O\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ \mathbf{0}^T & O & O & O & \cdots & q_{K-k-1}I & C_n^{(k-1)} \end{pmatrix}, k=\overline{1,K-2}, n=\overline{1,N-1},$$

where $\tilde{\mathbf{q}}^{(k)}=(q_{K-k},q_{K-k+1},\ldots,q_K)$, $k=\overline{1,K-2}$.

Lemma 2. *Let the matrices $S_n=S_n(H)$, $n=\overline{1,N}$, define the transition probabilities of the process $\{n_t^{(1)},\ldots,n_t^{(K)}\}$ at the moment when a new customer arrives at the network when the number of free servers is n and starts service. These matrices can be found as $S_1=\mathbf{h}_1^T$ where $\mathbf{h}_1$ is the first row of the matrix H and $S_n=S_n^{(K-2)}$, $n=\overline{2,N}$, where the matrices $S_n^{(k)}$, $n=\overline{2,N}$, of block size $(n+1)\times n$ define the transition probabilities of the components $n_t^{(K)},\ldots,n_t^{(K-k-1)}$ at the moment of customer acceptance to the network when* *there are n free servers conditional on the fact that the service starts at some node of the network from the set* $K, K-1, \dots, K-k-1$, $k = \overline{0, K-2}$. *These matrices can be computed by the recursion*

$$S_n^{(0)} = \begin{pmatrix} h_{n,K-1} & 0 & 0 & \cdots & 0 & 0 \\ h_{1,K} & h_{n-1,K-1} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & h_{n-1,K} & h_{1,K-1} \\ 0 & 0 & 0 & \cdots & 0 & h_{n,K} \end{pmatrix},$$

$$S_n^{(k)} = \begin{pmatrix} h_{n,K-k-1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \tilde{\mathbf{h}}^{(k)} & h_{n-1,K-k-1}I & O & O & \cdots & O & O \\ \mathbf{0}^T & S_2^{(k-1)} & h_{n-2,K-k-1}I & O & \cdots & O & O \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}^T & O & O & O & \cdots & S_{n-1}^{(k-1)} & h_{1,K-k-1}I \\ \mathbf{0}^T & O & O & O & \cdots & O & S_n^{(k-1)} \end{pmatrix}, \ k = \overline{1, K-2},$$

where $\tilde{\mathbf{h}}^{(k)} = (h_{1,K-k}, h_{1,K-k+1}, \dots, h_{1,K})^T$, $k = \overline{1, K-2}$.

Proof of Lemmas 1 and 2 is implemented by induction taking into account the reverse lexicographic order of components of the process $\{n_t^{(1)}, \dots, n_t^{(K)}\}$.

Using the results of these Lemmas, we can prove the following statement.

Theorem 1. *The generator* $\mathbf{G}$ *of the Markov chain* ξ_t, $t \geq 0$, *has the following block-tridiagonal (QBD) structure:*

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_{0,0} & \mathbf{G}_{0,1} & O & \dots & O & O \\ \mathbf{G}_{1,0} & \mathbf{G}_{1,1} & \mathbf{G}_{1,2} & \dots & O & O \\ O & \mathbf{G}_{2,1} & \mathbf{G}_{2,2} & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & \mathbf{G}_{N,N-1} & \mathbf{G}_{N,N} \end{pmatrix} \quad (1),$$

where

$$\mathbf{G}_{0,0} = D(1) - \mu N I_W, \quad (2)$$

$$\mathbf{G}_{n,n} = D_0 \otimes I_{T_n} - \mu(N-n) I_{WT_n} + \sum_{k=1}^{K} D_k \otimes \mathrm{diag}\{\mathbf{e}_{T_n} - S_n(\tilde{H}_k)\mathbf{e}\}, \ n = \overline{1, N}, \quad (3)$$

$$\mathbf{G}_{n,n+1} = \mu(N-n) I_W \otimes C_n, \ \ n = \overline{0, N-1}, \quad (4)$$

$$\mathbf{G}_{n,n-1} = \sum_{k=1}^{K} D_k \otimes S_n(\tilde{H}_k), \ n = \overline{1, N}. \quad (5)$$

Proof of Theorem 1 is implemented by means of the analysis of all possible transitions of the Markov chain during an interval of an infinitesimal length.

The generator $\mathbf{G}$ has a block-tridiagonal structure (1) (i.e., $\mathbf{G}_{n,n'} = O$ if $|n - n'| > 1$) because the probability that more than one customer arrives or departs from the network during an infinitesimally small interval is negligible. The diagonal entries of the block $\mathbf{G}_{0,0}$ are negative. The modulus of the corresponding entry of this block defines the rate of departure of the Markov chain ξ_t from the corresponding state. Because 0 servers are idle, all N servers of the network provide service and the total service completion rate is μN. Correspondingly, all arriving customers are rejected due to the lack of free servers. The corresponding intensities of the departure of the underlying process of the $MMAP$ from its states are given by the diagonal entries of the matrix $D(1)$. The non-diagonal entries of the block $\mathbf{G}_{0,0}$ are non-negative and define the intensities of the transition of the underlying process of the *MMAP* between its states. As the results of these considerations, we obtain formula (2).

The derivation of formula (3) for the block $\mathbf{G}_{n,n},\ n = \overline{1,N}$ is similar to the derivation of formula (2) with accounting for the difference that now in the network there are n idle servers and an arriving customer has a chance to start service. Therefore, because the block $\mathbf{G}_{n,n}$ only accounts for possible transitions of the Markov chain ξ_t without the change of the number n of idle servers in the network, we have to consider only the jumps of the underlying process of the *MMAP* without generation of customers (the intensities of these jumps are given by the matrix D_0) or with generation of customers that are immediately lost due to the lack of available server or balking. The matrix D_k defines the intensities of the jumps that are accompanied by a customer generation in node k. The diagonal matrix $\text{diag}\{\mathbf{e}_{T_n} - S_n(\tilde{H}_k)\mathbf{e}\}$ has the diagonal entries equal to 1 at the rows that correspond to the states $(n^{(1)}, \ldots, n^{(K)})$ of the process $\{n_t^{(1)}, \ldots, n_t^{(K)}\}$ such as $n^{(k)} = 0$ and equal to the probability $1 - h_{n,k}$ in all other rows. Thus, the matrix $D_k \otimes \text{diag}\{\mathbf{e}_{T_n} - S_n(\tilde{H}_k)\mathbf{e}\}$ defines the intensities of transitions without the change of the number of idle servers when the servers are available in the requested node. The symbol $\otimes$ of Kronecker product of matrices is used here (and in the analysis of multi-dimensional Markov chains in general) to describe simultaneous transitions of several independent Markov chains.

The derivation of formula (4) is clear because the block $\mathbf{G}_{n,n+1},\ n = \overline{0, N-1}$, describes the intensities of transitions of the process $\{n_t^{(1)}, \ldots, n_t^{(K)}\}$ from the level n to the level $n+1$. Such transitions occur when service in one of $N - n$ busy servers is finished (the intensity of this event occurrence is equal to $\mu(N-n)$), and the server appears in a certain node. Transition probabilities of the process $\{n_t^{(1)}, \ldots, n_t^{(K)}\},\ t \geq 0$, at the moment when service of a customer is finished and the number of free servers, was equal to n are defined by the matrix C_n described above.

In the derivation of formula (5) for the block $\mathbf{G}_{n,n-1},\ n = \overline{1,N}$, we account that the decrease in the number of idle servers from n to $n-1$ occurs when a new customer arrives in some node k (the intensities of the corresponding transitions are given by the matrix D_k), in which the number of idle servers is not equal to 0, and starts service. The corresponding intensities of transitions of the process $\{n_t^{(1)}, \ldots, n_t^{(K)}\}$ are given by the matrix $S_n(\tilde{H}_k)$. Formula (5) and Theorem 1 are proved.

After computation of transition probabilities of the Markov chain, it is necessary to compute steady-state (stationary) probabilities of the states of the chain.

Because the Markov chain ξ_t has a finite state space and its generator is irreducible, the following limits (stationary probabilities)

$$\pi(n, \nu, n^{(1)}, \ldots, n^{(K)}) = \lim_{t \to \infty} P\{n_t = n, \nu_t = \nu, n_t^{(1)} = n^{(1)}, \ldots, n_t^{(K)} = n^{(K)}\}$$

exist for any set of the parameters of the system.

Denote by $\boldsymbol{\pi}_n$ the row vector consisting of the stationary probabilities of the states of the Markov chain that belong to the level $n,\ n = \overline{0,N}$, and are enumerated in correspondence with definition of the level. The vectors $\boldsymbol{\pi}_n,\ n = \overline{0,N}$ satisfy the following system of linear algebraic equations:

$$(\boldsymbol{\pi}_0, \ldots, \boldsymbol{\pi}_N)\mathbf{G} = \mathbf{0},\ (\boldsymbol{\pi}_0, \ldots, \boldsymbol{\pi}_N)\mathbf{e} = 1.$$

The number of equations of this system may be large. Therefore, to solve this system, it is necessary to use algorithms that effectively use the sparse block structure of the generator $\mathbf{G}$. In particular, the algorithm from [21] can be recommended. Note that often it is possible to skip the computation of the stationary distribution of the Markov chain if only the values of some performance measures of the network are of interest. This can be done via the use of a memory-efficient method developed in [22].

4. Performance Measures of the Network

As soon as the vectors $\boldsymbol{\pi}_n,\ n = \overline{0,N}$ have been computed, we can determine various performance measures of the queueing network under consideration.

The average number of idle servers in the network is

$$N_{idle} = \sum_{n=1}^{N} n\boldsymbol{\pi}_n\mathbf{e}.$$

The average number of busy servers in the network is

$$N_{busy} = N - N_{idle}.$$

The average number of idle servers in the kth node is

$$N_{idle}^{(k)} = \sum_{n=1}^{N} \boldsymbol{\pi}_n(\mathbf{e}_W \otimes S_n(P_k)\mathbf{e}_{T_n-1}),\ k = \overline{1,K},$$

where the matrix P_k of size $N \times K$ has all zero columns except the k-th column, which is equal to $(1,2,\ldots,N)^T$.

The probability $P_{loss}^{(k)}$ that an arbitrary customer arriving to the kth node will be lost is computed by

$$P_{loss}^{(k)} = \frac{1}{\lambda_k}\left[\boldsymbol{\pi}_0 D_k\mathbf{e}_W + \sum_{n=1}^{N}\boldsymbol{\pi}_n\left(D_k\mathbf{e}_W \otimes (\mathbf{e}_{T_n} - S_n(\tilde{H}_k)\mathbf{e}_{T_n-1})\right)\right],\ \ k = \overline{1,K}.$$

The probability $P_{loss-no-car}^{(k)}$ that an arbitrary customer arriving to the kth node will be lost because there are no available cars in this node is computed by

$$P_{loss-no-car}^{(k)} = \frac{1}{\lambda_k}\left[\boldsymbol{\pi}_0 D_k\mathbf{e}_W + \sum_{n=1}^{N}\boldsymbol{\pi}_n(D_k\mathbf{e}_W \otimes \mathbf{a}_{n,k})\right], k = \overline{1,K},$$

where $\mathbf{a}_{n,k}$ is the vector of size T_n, the lth entry of which is equal to 1 if the lth entry of the vector $S_n(\tilde{H}_k)\mathbf{e}_{T_n-1}$ is equal to 0, $l = \overline{1,T_n}$, and is equal to 0, otherwise.

Availability of the kth node defined as the share of time, during which an arbitrary arriving customer will meet an available car, is computed as $1 - P_{loss-no-car}^{(k)},\ \ k = \overline{1,K}$.

The value $P_{loss}^{(k)} - P_{loss-no-car}^{(k)}$ is equal to the probability that an arbitrary customer arriving at the kth zone will be lost in the situation when there are available cars in this node, but the customer decides to cancel his/her journey (e.g., due to long walking distance between his/her current location and the available car in this node).

The probability P_{loss} that an arbitrary customer arriving to the network will be lost is computed by

$$P_{loss} = \frac{1}{\lambda}\left[\boldsymbol{\pi}_0\sum_{k=1}^{K} D_k\mathbf{e}_W + \sum_{n=1}^{N}\boldsymbol{\pi}_n\sum_{k=1}^{K}\left(D_k\mathbf{e}_W \otimes (\mathbf{e}_{T_n} - S_n(\tilde{H}_k)\mathbf{e}_{T_n-1})\right)\right].$$

The probability $P_{loss-no-car}$ that an arbitrary customer arriving to the network will be lost because there are no cars available is computed by

$$P_{loss-no-car} = \frac{1}{\lambda}\left[\boldsymbol{\pi}_0\sum_{k=1}^{K} D_k\mathbf{e}_W + \sum_{n=1}^{N}\boldsymbol{\pi}_n\sum_{k=1}^{K}(D_k\mathbf{e}_W \otimes \mathbf{a}_{n,k})\right].$$

Having analytical expressions for computation of performance characteristics of the system, we can develop software for their computation and give some numerical illustrations.

5. Numerical Results

In this paper, we built a novel model of a car sharing system as a queueing network with moving servers. Analogous models are not known in the literature and, therefore, we have no opportunity to compare our results with the existing in the literature. The single universal tool for testing the obtained results is computer simulation. To check the correctness of our analysis, we made a simulation model, and the results of simulation match our analytical results well.

The goals of the numerical experiment are: (i) to demonstrate the feasibility of the algorithms for computation of the stationary distribution of the Markov chain with the generator defined in Theorem 1 and the performance measures presented in the previous sections; (ii) to evaluate the dependence of the performance measures on the number N of servers (cars); and (iii) to show the necessity of account of correlation in the arrival process for adequate modelling of a car sharing system.

Let us consider the town which can be geographically divided into three zones. The average duration of the trip is 15 min. The information about the arriving flows of clients (customers) to each zone, probabilities of trip completion in a certain zone, and probabilities of a customer balking in the case when there are available cars is given below in the following description of the operation of this car sharing system in terms of the queueing network.

The number of nodes of the network is $K = 3$. The service rate is $\mu = \frac{1}{15} = 0.066$. The vector $\mathbf{q} = (q_1, \dots, q_K)$ defining the probabilities q_k that an arbitrary service finishes in the kth node, $k = \overline{1, K}$, is given by $\mathbf{q} = (0.29, 0.45, 0.26)$. The probabilities $h_{n,k}$ that an arbitrary customer arriving to the kth node when n servers are available in this node will start service (but not balk) are given by $h_{n,k} = \min\{1, \tilde{h}_k + 0.05(n-1)\}$, where $\tilde{h}_1 = 0.7$, $\tilde{h}_2 = 0.68$, $\tilde{h}_3 = 0.75$.

Because one of our goals is to show the importance of the account of correlation in the arrival process, we consider the following two $MMAP$s.

The first arrival process, which we code as $MMAP_0$, is the superposition of three stationary Poisson processes with the intensities $\lambda_1 = 0.170747$, $\lambda_2 = 0.270468, \lambda_3 = 0.158861$. The total rate of customers arrival to the network is $\lambda = 0.600076$. The coefficient of correlation of successive inter-arrival times in this arrival process is equal to 0.

The second arrival process, denoted as $MMAP_{0.15}$, is defined by the matrices

$$D_0 = \begin{pmatrix} -1.8 & 0 \\ 0.0 & -0.4458 \end{pmatrix}, D_1 = \begin{pmatrix} 0.51 & 0.05 \\ 0.006 & 0.1147 \end{pmatrix},$$

$$D_2 = \begin{pmatrix} 0.31 & 0.01 \\ 0 & 0.2641 \end{pmatrix}, D_3 = \begin{pmatrix} 0.91 & 0.01 \\ 0.003 & 0.058 \end{pmatrix}.$$

This arrival flow has the coefficient of correlation $c_{cor} = 0.1485$, therefore we code it as $MMAP_{0.15}$. The total arrival rate and arrival rates to all nodes are the same as those for $MMAP_0$. Thus, two considered $MMAP$s have equal arrival rates to each node but different correlation of successive inter-arrival times.

We evaluate the influence of the number N of servers operating within the network on some of its performance measures. Figures 2–4 show the dependence of the number of busy servers N_{busy}, the probability P_{loss} and the probability $P_{loss-no-car}$ on the number N servers in the network where the parameters N varies from 1 to 100. For computations, we use a PC with an Intel Core i7-8700 CPU (Santa Clara, CA, USA) and 16 GB RAM, Wolfram Mathematica (version 11, Wolfram, Champaign, IL, USA).

The times required for computation of the stationary distribution of the Markov chain and the listed performance measures for some values of N are given in Table 2.

Table 2. Computation time (CT) for different N.

N	10	20	30	40	50
CT	0.064 s	0.51 s	2 s	6.3 s	18 s
N	60	70	80	90	100
CT	46 s m	1.68 m	3.18 m	5.9 m	10.3 m

It can be observed that the computation time essentially increases with the growth of N. For $N = 20$, it is about half of a second. For $N = 50$, it is about 18 seconds. For $N = 100$, it is about 10 min. This fast increase of the computation time is explained by the increase of the maximal size of the blocks of the generator. In particular, the size of the block $\mathbf{G}_{n,n}$ is equal to $W \times T_n$, where $T_n = \binom{n+K-1}{K-1} = \frac{(n+K-1)!}{n!(K-1)!} = \frac{(n+2)(n+1)}{2}$, $n = \overline{1, N}$. For $N = 20$, the size of the maximal block is 462. For $N = 100$, the size is already 10, 302. It is worth noting that the data in Table 1 are given for the $MMAP_{0.15}$ arrival flow. Computations for the $MMAP_0$ arrival flow defined as the superposition of three stationary Poisson processes are faster because the size of the blocks is twice as small ($W = 1$).

It is worth noting that, in this experiment, we computed not only the stationary distribution and N_{busy}, P_{loss} and $P_{loss-no-car}$, but also all other performance measures listed in the previous section. To speed up computations, computation of some performance measures can be cancelled. In addition, it is possible to use more advanced computers and the results of [22].

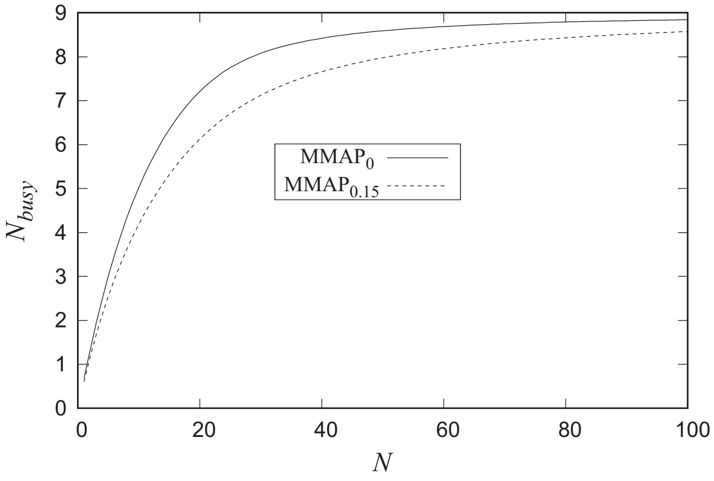

Figure 2. Dependence of the average number of busy servers in the network on the number N.

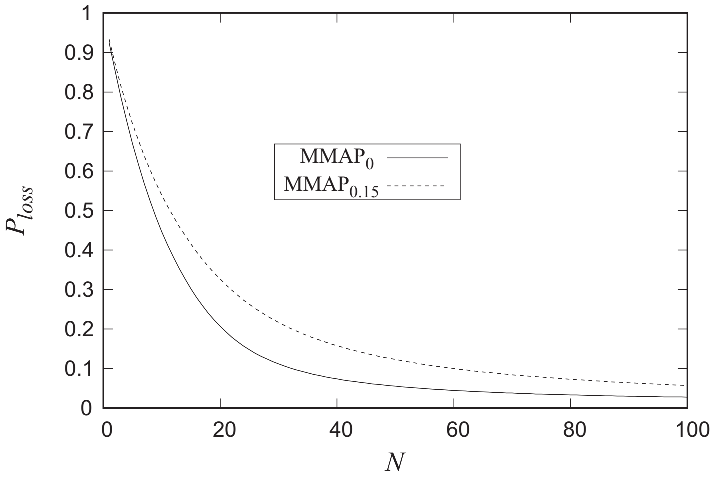

Figure 3. Dependence of the probability P_{loss} on the number N.

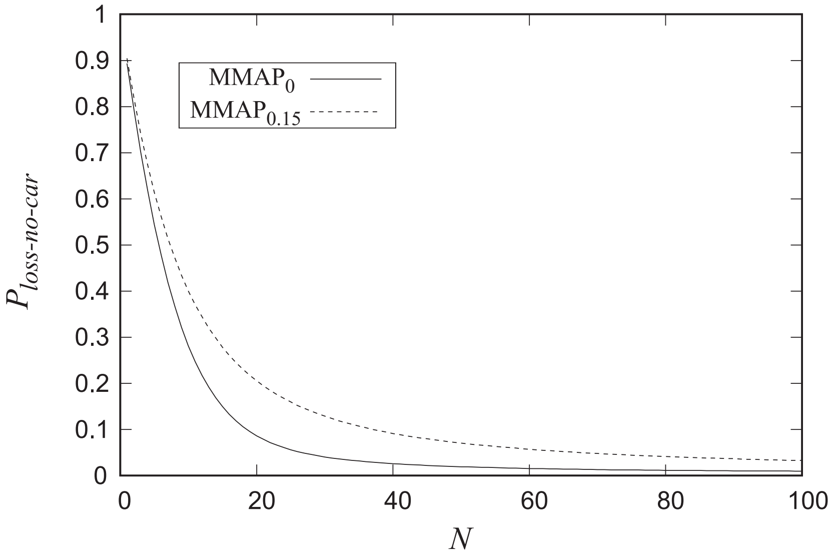

Figure 4. Dependence of the probability $P_{loss-no-car}$ on the number N.

Although we present the results of comparison of the results for the network with arrival flow having a correlation equal to 0 with the results for the network with arrival flow having relatively small correlation equal to 0.15, we can conclude that there is an essential difference in the values of N_{busy}, P_{loss} and $P_{loss-no-car}$. This leads to essential errors in the performance evaluation of the network and choosing the required fleet size N if the real arrival flow is correlated, but someone models it by the stationary Poisson process.

Likely, the most important performance measure of the network is the probability $P_{loss-no-car}$ that an arbitrary customer arriving at the network will be lost because there are no available cars. Let us formulate the problem of designing the car sharing system. To estimate the necessary investments into buying or leasing and maintenance of cars, we have to choose the number N of cars required in the considered network. We have to make this choice taking into account that we would like to obtain that the probability $P_{loss-no-car}$ that an arbitrary arriving customer will see available cars is not less than, say, 0.95. If we compute this required number N of cars in the assumption that the arrival flow is the superposition of three stationary Poisson processes, we obtain that $N = 27$ cars are enough. However, if statistics will show that the arrival flow is correlated with the coefficient of correlation 0.15, we have to account for this correlation and make computations using $MMAP_{0.15}$ as the model of the arrival process. After these computations, we obtain that the required number N is equal to 68. This is 2.5 times larger than 27. Therefore, if we choose $N = 27$ and plan availability of cars in 95 percent of cases, we have a lack of 41 cars that are indeed additionally required. The necessity of these additional investments into the development of the car sharing system can essentially change the estimation of reasonability of starting or holding this business.

Note that, for $N = 100$, the model with the arrival flow given by the superposition of three stationary Poisson processes estimates the loss probability due to unavailability of cars as 0.0098. However, for the correlated flow $MMAP_{0.15}$, this probability is about 3.34 times higher (it is equal to 0.0327917). Therefore, it is necessary to account for even a relatively small correlation of inter-arrival times.

In the previous examples, we considered the integral characteristics of the whole network. Let us consider now the characteristics of each node. Let us first consider the case when the arrival flow is $MMAP_{0.15}$. Figures 5–7 show the dependencies of the number $N_{idle}^{(k)}$ of idle servers, the probability $P_{loss}^{(k)}$ that an arbitrary customer arriving at the kth node will be lost and the probability $P_{loss-no-car}^{(k)}$ that an arbitrary customer arriving at the kth node will be lost because there are no available cars in this node on the number N.

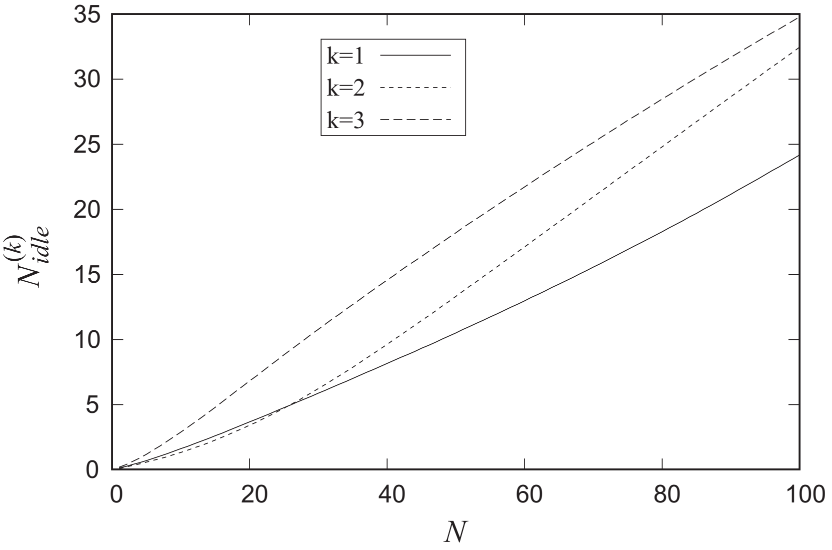

Figure 5. Dependence of the average number $N_{idle}^{(k)}$ of idle servers in the kth node on the number N for the case $MMAP_{0.15}$.

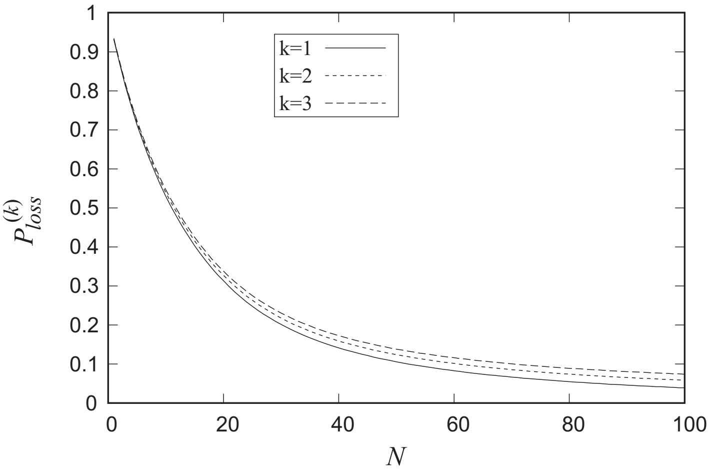

Figure 6. Dependence of the probability $P_{loss}^{(k)}$ that an arbitrary customer arriving to the kth node will be lost on the number N for the case $MMAP_{0.15}$.

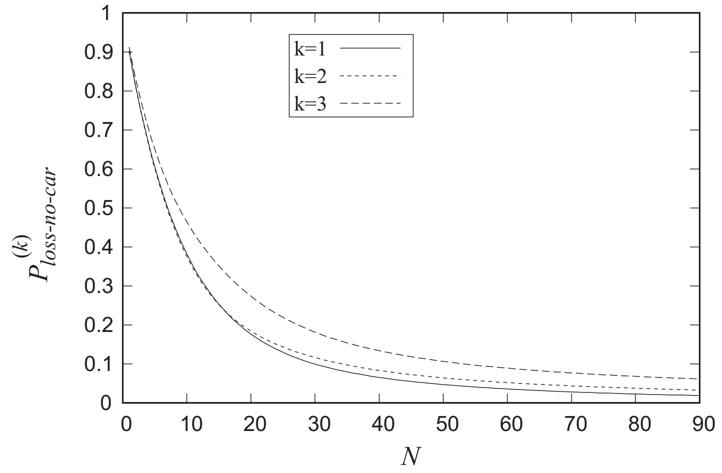

Figure 7. Dependence of the probability $P_{loss-no-car}^{(k)}$ that an arbitrary customer arriving to the kth node will be lost because there are no available cars in this node on the number N for the case $MMAP_{0.15}$.

First of all, it is worth noting the good matching of Figures 2 and 5. For example, for any N, the sum of values $N_{idle}^{(k)}$ given in Figure 5 by $k,\ k = 1,2,3$ is equal to the total number of idle servers N_{idle} in the network, which, in turn, is equal to $1 - N_{busy}$ where N_{busy} is given in Figure 2. Another observation is the following. Nodes 1 and 3 have very similar parameters (arrival rate, probability to finish service in this node). However, it is seen from Figures 6 and 7 that the loss probability in node 1 is essentially smaller. Correspondingly, the number of the idle servers is also essentially smaller. To explain this phenomenon, we additionally compute the coefficients of correlation of inter-arrival times of customers in each node separately. The values of these coefficients of correlation are 0.15388, 0.0019, 0.28229, correspondingly. Therefore, correlation in the flow of customers arriving at node 3 is almost double that of correlation in the flow of customers arriving at node 1. It is already known from the literature that the increase of correlation in the arrival flow to a queueing system leads to a higher value of the loss probability. Because correlation in the arrival process to node 2 is smaller than in the arrival processes to nodes 1 and 3, we could expect that the value of $N_{idle}^{(2)}$ is less than $N_{idle}^{(1)}$ and $N_{idle}^{(3)}$. However, this expectation contradicts Figure 5. Namely, we observe that $N_{idle}^{(2)} > N_{idle}^{(1)}$ for relatively large values of N. This fact becomes clear if we recall that the probability of transition of an arbitrary server at the moment of service completion to node 2 is equal to 0.45 while the corresponding probability for node 1 is only 0.29. Therefore, the number of servers arriving per unit of time to node 2 is essentially larger. This causes a larger number of idle servers in node 2.

The effect of the intersection of the curves for $k = 1$ and $k = 2$ on Figure 5 can be explained by correlation in the arrival process. If we consider the non-correlated arrival process $MMAP_0$, this intersection disappears, see Figure 8 below.

Figures 8–10 illustrate the same dependencies as given in Figures 5–7, but for the $MMAP_0$ arrival process.

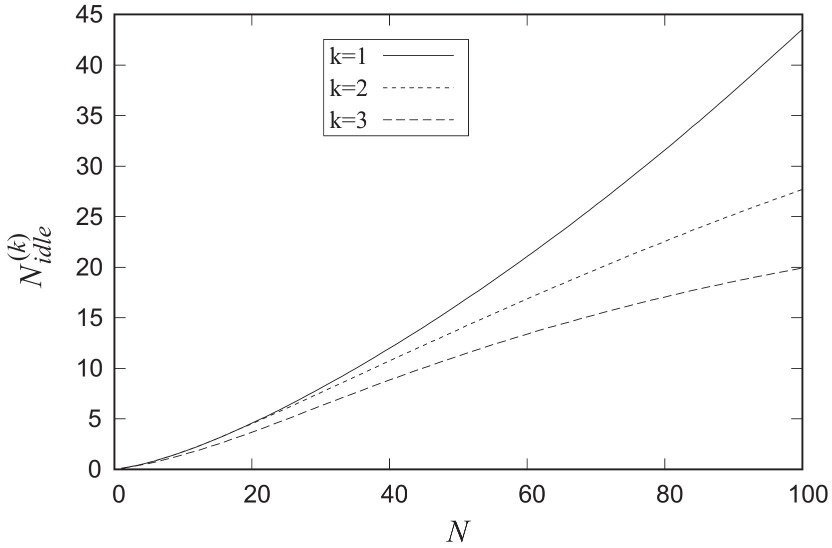

Figure 8. Dependence of the average number $N_{idle}^{(k)}$ of idle servers in the kth node on the number N for the case $MMAP_0$.

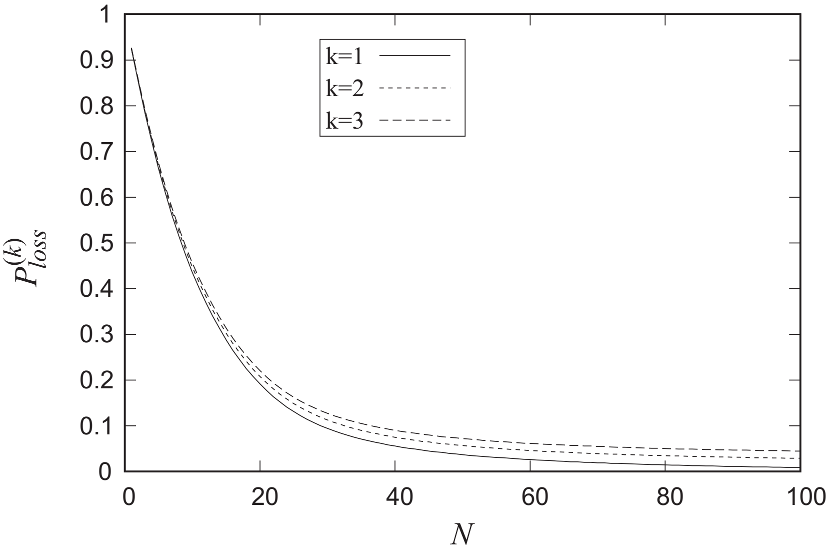

Figure 9. Dependence of the probability $P^{(k)}_{loss}$ that an arbitrary customer arriving to the kth node will be lost on the number N for the case $MMAP_0$.

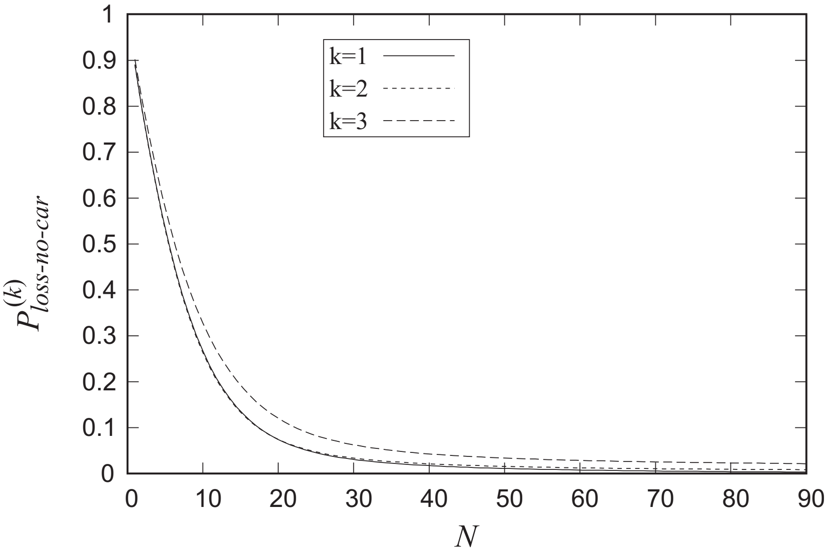

Figure 10. Dependence of the probability $P^{(k)}_{loss-no-car}$ that an arbitrary customer arriving to the kth node will be lost because there are no available cars in this node on the number N for the case $MMAP_0$.

Comparison of Figures 8–10 with Figures 5–7 confirms our conclusion made above that correlation has an essential impact on the performance measures of the network. Larger correlation implies a higher value of the loss probability and a larger number of idle servers. The approximation of a real-world arrival process with positive correlation by the stationary Poisson process leads to underestimation of the loss probability and too optimistic prediction of the network performance measures. Individual performance measures of the nodes also drastically change depending on correlation in the arrival process. For example, for $N = 100$, the maximal value of $N^{(k)}_{idle}$ is achieved in node 1 (about 43) and the minimal value of $N^{(k)}_{idle}$ is achieved in node 3 (about 20) in the case $MMAP_0$. At the same time, the maximal value of $N^{(k)}_{idle}$ is achieved in node 3 (about 35) and the minimal value of $N^{(k)}_{idle}$ is achieved in node 1 (about 24) in the case $MMAP_{0.15}$.

Therefore, correlation in the arrival process has a profound effect and only the use of our results allows for exactly computing performance measures of the network under the fixed values of its parameters, including the pattern of arrival process.

6. Conclusions

We consider a problem of the choice of the number of required cars (fleet management problem) in a car sharing system and performance evaluation of this system under any fixed set of its parameters. We construct the mathematical model of car sharing as an open queueing network. The dynamics of this network are described by a multi-dimensional continuous-time Markov chain. We derive the generator of this chain and expressions for the main performance measures of the network. Presented results of numerical experiments illustrate the dependence of some performance measures of the network and nodes of this network on the number of available servers (cars). The importance of the accounting of possible correlation in the arrival process is numerically shown. Results can be used for the optimal choice of the number of required cars with respect to various criteria, e.g., the loss probability of an arbitrary customer, availability of servers, the idle time of a car, profit from the operation, etc.

Author Contributions: Conceptualization, C.K. and S.D.; methodology, C.K. and S.D.; software, O.D.; validation, O.D.; formal analysis, O.D., C.K., and S.D.; investigation, C.K., O.D.; writing—original draft preparation, C.K. and S.D.; writing—review and editing, C.K. and O.D.; supervision C.K.; project administration O.D.

Funding: The work by Chesoong Kim has been supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grant No. 2018K2A9A1A06072058). The work by Sergei Dudin and Olga Dudina has been supported by "RUDN University Program 5-100".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ferrero, F.; Perboli, G.; Rosano, M.; Vesco, A. Car-sharing services: An annotated review. *Sustain. Cities Soc.* **2018**, *37*, 501–518. [CrossRef]
2. Perboli, G.; Ferrero, F.; Musso, S.; Vesco, A. Business models and tariff simulation in car-sharing services. *Trans. Res. Part A Policy Pract.* **2018**, *115*, 32–48. [CrossRef]
3. Ströhle, P.; Flath, C.M.; Gärttner, J. Leveraging Customer Flexibility for Car-Sharing Fleet Optimization. *Trans. Sci.* **2018**, *53*, 42–61. [CrossRef]
4. Fanti, M.P.; Mangini, A.M.; Pedroncelli, G.; Ukovich, W. Fleet sizing for electric car sharing system via closed queueing networks. In Proceedings of the 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), San Diego, CA, USA, 5–8 October 2014; pp. 1324–1329.
5. George, D.K.; Xia, C.H. Fleet-sizing and service availability for a vehicle rental system via closed queueing networks. *Eur. J. Oper. Res.* **2011**, *211*, 198–207. [CrossRef]
6. Zou, A.A.; Down, D.G. Asymptotically Maximal Throughput in Tandem Systems with Flexible and Dedicated Servers. *Asia-Pac. J. Oper. Res.* **2018**, *35*, 1850038. [CrossRef]
7. Grassmann, W.K.; Tavakoli, J. A tandem queuewith a movable server: An eigenvalue approach. *SIAM J. Matrix Anal. Appl.* **2002**, *24*, 465–474. [CrossRef]
8. Ganguly, A.; Ramanan, K.; Robert, P.; Sun, W. A Large-Scale Network with Moving Servers. *ACM SIGMETRICS Perform. Eval. Rev.* **2017**, *45*, 42–44. [CrossRef]
9. Borst, S.; Boxma, O. Polling: Past, present, and perspective. *TOP* **2018**, *26*, 335–369. [CrossRef]
10. Baccelli, F.; Rybko, A.N.; Shlosman, S.B. Queueing networks with mobile servers: The mean-field approach. *Probl. Inf. Trans.* **2016**, *52*, 178–199. [CrossRef]
11. Baccelli, F.; Rybko, A.; Shlosman, S.; Vladimirov, A. Metastability of Queuing Networks with Mobile Servers. *J. Stat. Phy.* **2018**, *173*, 1227–1251. [CrossRef]
12. Kim, J.; Dudin, A.; Dudin, S.; Kim, C. Analysis of a semi-open queueing network with Markovian arrival process. *Perform. Eval.* **2018**, *120*, 1–19. [CrossRef]
13. Kim, C.; Dudin, S.; Dudin, A.; Samouylov, K. Analysis of a Semi-Open Queuing Network with a State Dependent Marked Markovian Arrival Process, Customers Retrials and Impatience. *Mathematics* **2019**, *7*, 715. [CrossRef]
14. Dhingra, V.; Kumawat, G.L.; Roy, D.; de Koster, R. Solving semi-open queuing networks with time-varying arrivals: An application in container terminal landside operations. *Eur. J. Oper. Res.* **2018**, *267*, 855–876. [CrossRef]

15. He, Q.M. Queues with marked customers. *Adv. Appl. Probab.* **1996**, *28*, 567–587. [CrossRef]
16. Baek, J.; Dudina, O.; Kim, C. A queueing system with heterogeneous impatient customers and consumable additional items. *Int. J. Appl. Math. Comput. Sci.* **2017**, *27*, 367–384. [CrossRef]
17. Chakravarthy, S. The batch Markovian arrival process: a review and future work. In *Advances in Probability Theory and Stochastic Processes*; Krishnamoorthy, A., Raju, N., Ramaswami, V., Eds.; Notable Publications Inc.: Branchburg, NJ, USA, 2001; pp. 21–29.
18. Buchholz, P.; Kemper, P.; Kriege, J. Multi-class Markovian arrival processes and their parameter fitting. *Perform. Eval.* **2010**, *67*, 1092–1106. [CrossRef]
19. Horváth, G.; Telek, M. Markovian Performance Evaluation with BuTools. In *Systems Modeling: Methodologies and Tools*; Springer: Berlin, Germany, 2019; pp. 253–268.
20. Graham, A. *Kronecker Products and Matrix Calculus with Applications*; Courier Dover Publications: Mineola, NY, USA, 2018.
21. Baumann, H.; Sandmann, W. Numerical solution of level dependent quasi-birth-and-death processes. *Procedia Comput. Sci.* **2010**, *1*, 1561–1569. [CrossRef]
22. Baumann, H.; Sandmann, W. Multi-server tandem queue with Markovian arrival process, phase-type service times, and finite buffers. *Eur. J. Oper. Res.* **2017**, *256*, 187–195. [CrossRef]