ISSN 0032-9460, Problems of Information Transmission, 2016, Vol. 52, No. 2, pp. 178–199. © Pleiades Publishing, Inc., 2016.
Original Russian Text © F. Baccelli, A.N. Rybko, S.B. Shlosman, 2016, published in Problemy Peredachi Informatsii, 2016, Vol. 52, No. 2, pp. 85–110.

THE INTERNATIONAL DOBRUSHIN PRIZE

Queueing Networks with Mobile Servers: The Mean-Field Approach

F. Baccelli[a], **A. N. Rybko**[b,1,2], **and S. B. Shlosman**[c,2,3]

[a]*Department of Mathematics, University of Texas at Austin, Austin, USA*
e-mail: Baccelli@math.utexas.edu
[b]*Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia*
e-mail: rybko@iitp.ru
[c]*Aix Marseille Université, Université de Toulon, CNRS, CPT, Marseille, France*
Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia
e-mail: shlos@iitp.ru

Received August 17, 2015; in final form, January 12, 2016

Abstract—We consider queueing networks which are made from servers exchanging their positions on a graph. When two servers exchange their positions, they take their customers with them. Each customer has a fixed destination. Customers use the network to reach their destinations, which is complicated by movements of the servers. We develop the general theory of such networks and establish the convergence of the symmetrized version of such a network to some nonlinear Markov process.

DOI: 10.1134/S0032946016020071

1. INTRODUCTION

In this paper we start studying a model of a queueing network comprised of moving servers. The servers are moving over the set of nodes of a graph G in such a way that at any time each node harbors a single server. Customers enter the network at each node. When a customer c arrives to some node, it joins the queue of the server currently harbored by this node. In the following we will simply say that a customer joins the queue at this node. Customer c has also a designated exit node $D(c)$, which it needs to reach in order to exit the network. In order to reach its destination, a customer has to visit a series of intermediate servers. The time a customer spends at a server depends on the service discipline of the server. Once a customer c being served by a server at node v leaves, it is sent to the server located at the adjacent node v' which is closest to the destination node $D(c)$. Once it gets to $D(c)$, the customer leaves the system.

The main feature of the network we are considering here is that servers are moving over the graph. This, while customers are waiting to be served, two servers located at adjacent nodes can move by simultaneously swapping their locations. When two servers operate such a swap, each one

[1] The International Dobrushin Prize for 2015 was awarded to Alexander Nikolaevich Rybko. The prize was presented on July 21, 2015, at Sinai's seminar of the Kharkevich Institute for Information Transmission Problems of the Russian Academy of Sciences.

[2] The results of Sections 3–5 were obtained at the Institute for Information Transmission Problems of the Russian Academy of Sciences at the expense of the Russian Science Foundation, project no. 14-50-00150.

[3] Supported in part by the Labex Archimede (ANR-11-LABX-0033) and the A*MIDEX project (ANR-11-IDEX-0001-02), funded by the "Investissements d'Avenir" French Government programme managed by the French National Research Agency (ANR).

brings along all the customers buffered in its queue. Thus, if the server at v currently containing c moves, then the distance between c and its destination $D(c)$ might change.

In such networks with moving servers, new effects take place, which are not encountered in the usual situation with stationary servers. For example, it can happen that "very nice" networks—i.e., networks with fast servers and low load—become unstable. Here instability means that, in a large network, queue sizes becomes bigger and bigger with time. In contrast, for the same parameters, the queues remain finite in the network with stationary servers.

This instability, which appears as a result of the movement of servers, will be a subject of our forthcoming paper [1]. In the present paper we focus on a mean-field approach to such networks with moving servers. The mean-field version of the network consists of N copies of the latter, interconnected in a mean-field manner. We show that in the limit $N \to \infty$ the network state process converges to a nonlinear Markov process (NLMP). The present paper is focused on the existence of the NLMP and on this convergence theorem. We also have results showing that ergodic properties of this NLMP are related to stability/instability properties of our prelimit networks. This relation between the ergodicity and stability will not be discussed here and will be the object of the future paper [1].

The mean-field idea. Here we remind the reader about the mean-field approach, which originates in statistical mechanics. Let us explain the main ideas on the simplest example, namely the *Ising model*. The Ising model features a collection of spin variables $\sigma_i = \pm 1$, which are assigned to sites i of the integer lattice $\mathbb{Z}^d$ of dimension $d \geq 1$, or to sites of a finite set $V \subset \mathbb{Z}^d$. The joint distribution of the spins $\sigma_V = \{\sigma_i,\, i \in V\}$ is governed by the *Hamiltonian* H_V, which for the Ising model is given by

$$H_V(\sigma_V) = -\sum_{i\sim j} \sigma_i \sigma_j, \tag{1}$$

where the summation is over the pairs $i, j \in V$ of nearest neighbor lattice sites. The stochastic *weight* $w(\sigma_V)$ of a configuration σ_V is taken to be

$$w(\sigma_V) = \exp\{-\beta H_V(\sigma_V)\},$$

where the parameter $\beta > 0$ is called the *inverse temperature*. The corresponding probability distribution $\Pr(\sigma_V)$ is obtained via normalization, $\Pr(\sigma_V) = w(\sigma_V)/Z(V,\beta)$, where the *partition function* $Z(V,\beta)$ is just the sum: $Z(V,\beta) = \sum\limits_{\sigma_V} w(\sigma_V)$. With some care this definition can be extended to the case of infinite V, including $V = \mathbb{Z}^d$. The extension $V \to \mathbb{Z}^d$ is called the *thermodynamic limit*.

In spite of its simplicity, the Ising model turns to be quite interesting and nontrivial. It still provides hard problems, and there are still long-standing open questions about it, particularly so for the infinite model. To cope with these difficulties, physicists came up with the mean-field approximation for the model. Instead of the d-dimensional graph $V \subset \mathbb{Z}^d$, one considers the model on the complete graph $\mathfrak{K}_N$ with $N+1$ vertices, and replaces the Hamiltonian (1) by

$$H_N(\sigma_{\mathfrak{K}_N}) = -\frac{1}{N} \sum_{i,j} \sigma_i \sigma_j, \tag{2}$$

where the summation is now taken over all pairs $i \neq j$. The extra factor $\dfrac{1}{N}$ in (2) is needed in order to make the "interaction" between the single spin σ_i and the rest of the world stay finite in the thermodynamic limit $N \to \infty$. The other definitions remain the same.

The mean-field Ising model is much easier to study than its lattice version. The reason is, of course, the larger symmetry of the graph $\mathfrak{K}_N$. One can also say that the mean-field model corresponds to dimension $d = \infty$. What is surprising is that the mean-field models do capture some relevant properties of the lattice models when the dimension d is not too low. For example, model (2) undergoes *phase transition* if the temperature β^{-1} is low enough. The same is true for model (1) provided that $d \geq 2$, while for $d = 1$ there is no transition. The literature on the Ising model and the mean-field models is huge. A general result stating that the mean-field models do capture the key features of the lattice models starting from "reasonable" dimensions is proved in [2].

Mean-field limits play a key role in the study of queueing systems in a variety of contexts. A typical example is that of systems with selection of the shortest queue [3]. The approach was later generalized in [4]. See also the papers [5, 6], where some general facts (like the Poisson hypothesis) about the behavior of these mean-field limiting processes are proved.

Nonlinear Markov processes. Nonlinear Markov processes are central in the theory of mean-field limit approximations. They were introduced in [7]. Here we recall what is meant by nonlinear Markov processes. We do this for the simplest case of discrete time Markov chains taking values in a finite set S, $|S| = k$. For the general case, see [5, Sections 2 and 3]. In the discrete case, the set of states of a Markov chain is the simplex Δ_k of all probability measures on S, $\Delta_k = \{\mu = (p_1, \ldots, p_k) : \ p_i \geq 0,\ p_1 + \ldots + p_k = 1\}$. The matrix $P(i,j)$ of transition probabilities defines the Markov evolution, which is the linear map $P^{\text{lin}}\colon \Delta_k \to \Delta_k$, given by $\mu \rightsquigarrow \mu P$.

By definition, a *nonlinear Markov chain* is a family of transition probability matrices P_μ, $\mu \in \Delta_k$, such that the matrix entry $P_\mu(i,j)$ is the probability of going from i to j in one step, starting in the state μ. The (nonlinear) evolution $P^{\text{n-lin}}\colon \Delta_k \to \Delta_k$ is then defined by $\mu \rightsquigarrow \mu P_\mu$. Both P^{lin} and $P^{\text{n-lin}}$ are deterministic dynamical systems on Δ_k.

Ergodic properties of linear Markov chains are settled by the Perron–Frobenius theorem. In particular, if the linear map P^{lin} is such that the image $P^{\text{lin}}(\Delta_k)$ belongs to the interior $\operatorname{Int}(\Delta_k)$ of Δ_k, then there is precisely one point $\mu \in \operatorname{Int}(\Delta_k)$ such that $P^{\text{lin}}(\mu) = \mu$, and for every $\nu \in \Delta_k$ we have the convergence $P^n(\nu) \to \mu$ as $n \to \infty$. In the nonlinear case of $P^{\text{n-lin}}$ we are dealing with a more or less arbitrary dynamical system on Δ_k, and the question about the stationary states of the chain or about measures on Δ_k which are invariant under $P^{\text{n-lin}}$ cannot be settled in general.

In what follows we will alternatively use the following two equivalent points of view: either one can talk about the deterministic dynamical system $P^{\text{n-lin}}$ on the space of probability measures (given by nonlinear differential equations in the continuous case), or else one can consider the stochastic evolution, given by the family $\{P_\mu, \mu \in \Delta_k\}$ of transition probability matrices. The finite-dimensional probability distribution of a trajectory $\omega = \{i_0, i_1, \ldots, i_k\}$ of the nonlinear Markov chain with initial state ν and the family $\{P_\mu\}$ of transition probability matrices is given by

$$\Pr(\omega = \{i_0, i_1, i_2, \ldots\}) = \nu(i_0) P_\nu(i_0, i_1) P_{(\nu P_\nu)}(i_1, i_2) \ldots .$$

2. MAIN RESULT AND LAYOUT OF THE PAPER

2.1. Description of the Network and Main Result

We consider a queueing network, denoted by $\mathcal{K}_1$, with servers jumping on a connected graph

$$G = [V(G), E(G)].$$

We assume that at every node $v \in V$, at any time, there is one server with a queue q_v of customers in the (infinite capacity) buffer of that server, waiting there for service. Every customer c at v has some destination, $D(c) \in V(G)$, and the goal of the customer is to reach this destination. In order to get there, a customer completing its service at v jumps along the edges of G to one of the nodes v' of G which is closest (in the graph distance) to $D(c)$. There it joins the queue of the server currently harbored by node v'. Once the destination of a customer is reached, it leaves the network.

In the meantime, the servers of our network may jump. More precisely, two servers at v and v', which are neighbors in G, can exchange their positions with rate $\beta_{vv'} = \beta_{v'v}$. The queues q_v and $q_{v'}$

then exchange their positions as well. Of course, such an exchange may bring some customers of q_v and $q_{v'}$ closer to their destinations, and some others further away from their destinations. We assume that the rates $\beta_{v'v}$ of these jumps are uniformly bounded by a constant:

$$|\beta_{v'v}| < \beta. \tag{3}$$

The graphs G that we are interested in can be finite or infinite. The case of finite graphs is easier, while the infinite case requires certain extra technical points. In particular, when we talk about functions of the states of our network in the infinite graph case, we will assume that they are either local or quasi-local. We recall that a function f is called *local* if there exists a finite subset $\Lambda \subset V$ such that f depends only on the states of the servers at the nodes $v \in \Lambda$. Similarly, a function f is called *quasi-local* if for some finite subset $\Lambda \subset V$ the dependence of f on the states of the servers at nodes $v \notin \Lambda$ decays exponentially fast in the $\mathrm{dist}(v, \Lambda)$, in some appropriate norm. This exponential decay property will be conserved by our dynamics.

We will assume that the degrees of the vertices of G are finite and uniformly bounded by some constant $D(G)$. Of course, this automatically holds in the finite graph case.

The network $\mathcal{K}_N$. In order to make our network tractable, we will study its *symmetrized*, or mean-field, modification, $\mathcal{K}_N$. This means that we pass from the graph G to its mean-field version, the graph $G_N = G \times \{1, \ldots, N\}$, and eventually take the limit $N \to \infty$. By definition, the graph G_N has the set of vertices $V(G_N) = V(G) \times \{1, \ldots, N\}$; two vertices $(v, k), (v', k') \in V(G_N)$ define an edge in $E(G_N)$ if and only if $(v, v') \in E(G)$. As we shall see, the restriction of our state process from $G \times \{1, \ldots, N\}$ to the subgraph $G \equiv G \times \{1\}$ goes, as $N \to \infty$, to a nonlinear Markov process on G, which is a central object of our study. We denote by $\mathcal{K}_N$ the network on the graph G_N. The limiting network $\mathcal{K}$ (which can be analyzed by the NLMP mentioned above) is the limit of the networks $\mathcal{K}_N$ on $G \times \{1, \ldots, N\}$. For the limit to exist, the rate of exchange $\beta_{vv'}$ should be renormalized as we pass from G to G_N. For a server located at $(v, k) \in V(G \times \{1, \ldots, N\})$, the swap with the server (v', k'), where node v' is a neighbor of node v, should have the rate

$$\frac{\beta_{vv'}}{N}.$$

This implies that the server at (v, k) will exchange its position with one of the servers positioned at the nodes $v' \times \{1, \ldots, N\}$ with the rate $\beta_{vv'}$, independent of N.

Each customer has a class, $\varkappa \in K$, where K is some finite alphabet of classes. If a customer of class $\varkappa$ completes its service at the server at v and goes to the server at v', it then gets a new (deterministic) class $\varkappa' = \mathcal{T}(\varkappa; v, v')$. Once a server finishes serving a customer, it chooses another one from its queue, according to the class of the customers present in the queue and the service discipline. It can happen that the service of a customer is preempted if a customer with higher priority comes, and then the interrupted service is resumed after an appropriate time.

The random service time η of a client c, which starts its service at the node v, depends on the client class $\varkappa$ and on the node v; i.e., $\eta = \eta(\varkappa, v)$. If the servers at v and v' swap, which can happen if $\beta_{vv'} \neq 0$, while the client c was served for time τ, then the distribution of its remaining service time at v' is just the conditional distribution of the random variable $\eta(\varkappa, v') - \tau$ under the condition that $\eta(\varkappa, v') > \tau$. We do not assume that η is exponential.

Every customer c in $\mathcal{K}_N$ has its destination node, $D(c) = v \in V(G)$ (or, equivalently, a destination set $v \times \{1, \ldots, N\} \subset V(G_N)$). In spite of the fact that our servers do change their positions, this location $D(c)$ does not does not change with time. Customer c tries to get to its destination node; in order to do so, if it is located at (v, k) and finishes its service there, then it goes to the server at (v', n), where $v' \in G$ is the neighbor of v which is closest to $D(c)$. If there are several such v', one is chosen uniformly at random. The coordinate $n \in \{1, \ldots, N\}$ is chosen uniformly at random as well. If at the end of the service it happens that v is at distance 1 from $D(c)$ or that v coincides with $D(c)$, then the customer leaves the network. However, while the customer c is waiting for service completion at the server at (v,k), then nothing special happens with it, even if $v = D(c)$; this server might drift away from node $D(c)$, and the distance between c and its destination $D(c)$ might then increase during its waiting time.

A more formal definition of the Markov process describing the evolution of the network $\mathcal{K}_N$ will be given in Section 5.1.

Our main result is the proof of the convergence of the network $\mathcal{K}_N$ to some nonlinear Markov process, which is the limiting mean-field system (see Section 2.5). The proof is based on the characterization of the infinitesimal operator Ω_N of the continuous time Markov process describing $\mathcal{K}_N$. We want to pass to the limit $N \to \infty$, since in this limit the nature of the process becomes simpler. The key observation is that if this limit exists, then the arrivals to each server at every time is a Poisson point process (with time-dependent rate function). Indeed, the flow to every server is the sum of N flows of rates $\sim \dfrac{1}{N}$. Since the probability that a customer served at a given node revisits this node goes to zero as $N \to \infty$, the arrivals to a given server in disjoint intervals are asymptotically independent in this limit.

In order to check that the limit $N \to \infty$ exists, we will formally write down the limiting infinitesimal generator Ω. We will then show that it defines a (nonlinear) Markov process. Finally, we will check that the convergence $\Omega_N \to \Omega$ is such that the Trotter–Kurtz theorem applies.

The rest of the paper is organized as follows. We start with the limiting process, namely the NLMP. We describe its state space in Section 2.2 and its possible jumps, together with its evolution equation, in Section 2.5. Section 2.5 formulates our main results (in Theorem 1): the first one concerns the existence of the NLM process and the second is about the convergence of the networks $\mathcal{K}_N$ to it as $N \to \infty$. The existence result is proved in Section 3. Section 4 is devoted to various compactification arguments. We leverage these arguments in Section 5 to check the applicability of the Trotter–Kurtz theorem, which is used to prove the convergence result.

2.2. State Space of the Mean-Field Limit

We describe below the configuration space of the mean-field limit, which will be referred to as the *Comb*. The state of the mean-field limit process will then be a probability distribution on the Comb.

At any given time, at each node $v \in G$, we have a server with a finite ordered queue q_v of customers, $q_v = \{c_i\} \equiv \{c_i^v\} \equiv \{c_1^v, \ldots, c_{l(q_v)}^v\}$, where $l(q_v)$ is the length of the queue q_v. The customers are ordered according to their arrival times to this server. We will denote by $C(q_v)$ the customer of queue q_v which is being served, and we denote by $\tau(C(q_v))$ the amount of service that this customer has already received. (We need to keep track of this, since service times are not exponential in general.) It can happen that the queue q_v has customers of lower priority than $C(q_v)$, which have already received some service but whose service was postponed due to the arrival of higher priority customers.

Let $i^*(q_v)$ be the location of the customer $C(q_v)$ in the queue q_v, i.e., $C(q_v) \equiv c_{i^*(q_v)}^v$. The service discipline is some rule R_v to choose the location $i^*(q_v)$ of the customer which has to be served. In what follows we assume that the rule R_v is some function of the sequence $\varkappa_1, \ldots, \varkappa_{l(q_v)}$ of customer classes and of the sequence $D(c_1^v), \ldots, D(c_{l(q_v)}^v)$ of their destinations, so that

$$i^*(q_v) = R_v\big[\{\varkappa_1, \ldots, \varkappa_{l(q_v)}\}, \{D(c_1^v), \ldots, D(c_{l(q_v)}^v)\}\big].$$

We assume for simplicity that the function R_v depends on $\{D(c_1^v), \ldots, D(c_{l(q_v)}^v)\}$ only through the relative distances $\operatorname{dist}(D(c_i^v), v)$. In what follows we consider only conservative disciplines. This means that the server cannot be idle if the queue is not empty.

The state of a server consists of

1. The classes $\varkappa_i \equiv \varkappa_i(c_i) \in K$, $|K| < \infty$, of its customers. We will denote by $\bar{\varkappa}(c)$ the class of the customer c once its service at the current server is over. It is defined by the class function $\mathcal{T}(\varkappa; \cdot, \cdot)$ mentioned above;
2. The destination nodes $v_i = D(c_i) \in V(G)$ which the customers want to reach;
3. The amount of service already acquired by the customers $c_1^v, \ldots, c_{l(q_v)}^v$. This will be denoted by $\tau_1, \ldots, \tau_{l(q_v)}$. This vector will be called for brevity the vector of τ-times. At any given time, the only τ variable which is growing is $\tau_{i^*(q_v)} \equiv \tau[C(q_v)] \equiv \tau[c_{i^*(q_v)}^v]$. Sometimes we will write $c^v \equiv c^v(\varkappa, \tau, v')$ for a customer located at v, of class $\varkappa$, which has already received the amount τ of service, and whose destination is $v' \in V$.

The space of possible queue states at v is denoted by M_v. The "coordinates" in M_v are those listed in the three items above. Thus, M_v is a countable union of finite-dimensional positive orthants. The orthants are indexed by finite strings of pairs $w = \{(\varkappa_i, D_i),\ i = 0, 1, \ldots, l\}$, where $\varkappa_i \in K$ and $D_i \in V$. Such a string corresponds to a queue with l customers with types $\varkappa_i$ and destinations D_i. A point in the orthants represents a vector of received service times $\tau_1, \ldots, \tau_l$ for the l customers. Each M_v will also be called a Comb.

In the following, in order to ease the notation, we will not always list all the indices that our variables depend on. For example, for certain questions, the destinations of the customers present at a server are not important and will be omitted.

Let $M = \prod_{v \in V} M_v$. The state of the NLMP is a probability measure μ on the product space M. We denote by $\mathcal{P}$ the space of all probability measures on M. It turns out that we will encounter only product measures on M. We will discuss this point below; see also [8], where we prove a simple extension of de Finetti's theorem.

2.3. Possible Jumps of the Mean-Field Limit Process

We list here all possible jumps of the NLM process and their rates.

2.3.1. Arrival of external customers. An external customer $c^v(v')$ of class $\varkappa \in K$ arrives to the server at node v, with destination $D(c) = v'$, with rate $\lambda = \lambda(\varkappa, v, v')$. We assume that

$$\sum_{\varkappa, v'} \lambda(\varkappa, v, v') < C, \tag{4}$$

uniformly in v. We will write that the queue state $q = \{q_u,\ u \in V\}$ changes to $q' = \{q_u,\ u \in V\} \oplus c^v(v')$. The τ-times of the customers present before the arrival stay the same, and the newly arrived has its τ-time equal to 0. The associated jump rate $\sigma_{\mathrm{e}}(q, q')$ is

$$\sigma_{\mathrm{e}}(q, q') = \lambda(\varkappa, v, v'). \tag{5}$$

2.3.2. Service completion. It is easy to see that the customer in service at node v, which received the amount τ of service, finishes its service at v with the rate $\dfrac{\mathcal{F}'_{\varkappa(C(q_v)),v}(\tau)}{1 - \mathcal{F}_{\varkappa(C(q_v)),v}(\tau)}$, where $\mathcal{F}_{\varkappa,v}$ denotes the distribution function of the service time. For future use we assume that this rate has a limit as $\tau \to \infty$. We also assume that it is uniformly bounded by a constant $\mathcal{F} < \infty$. The queue state $q = \{q_u,\ u \in V\}$ changes to $q' = \{q_u,\ u \in V\} \ominus C(q_v)$, so we denote this rate by

$$\sigma_{\mathrm{f}}(q, q') = \frac{\mathcal{F}'_{\varkappa(C(q_v)),v}(\tau)}{1 - \mathcal{F}_{\varkappa(C(q_v)),v}(\tau)} \leq \mathcal{F}. \tag{6}$$

The τ-times of the other customers stay the same. (Queuing theorists might be surprised by these "departures without arrivals," whereas customers do not necessarily leave the network. As we shall see, in the mean-field limit, any single departure from v to v' has no effect on the state of the queues of v'. This can be explained by the uniform routing to the N mean-field copies in the pre-limit and by letting N tend to infinity. However, the sum of the departure processes from all copies of the servers at v leads to a positive arrival rate from v to v', which is evaluated in Section 2.3.4 below.)

2.3.3. Servers jumping. Assume that the server at v jumps and exchanges its position with the one at v'. As a result, the queue q_v is replaced by a (random) queue Q distributed according to the distribution law $\mu_{v'}(dQ)$, where $\mu_{v'}$ is the marginal of μ on $M_{v'}$. Thus, the state of the server at v changes from q_v to q'_v, which is drawn from the distribution $\mu_{v'}$ on $M_{v'}$. The rate is

$$\sigma_{\mathrm{ex}}(q_v, Q)\, dQ = \beta_{vv'}\, \mu_{v'}(dQ). \tag{7}$$

Of course, the new queue Q at v comes equipped with its own τ-times. Note that the server at v' does not change its state after the swap. That is again the manifestation of the mean-field structure of our model. In the pre-limit (namely for $N < \infty$), the jump event that we discuss corresponds to the swap between the servers at $(v, 1)$ and at (v', k), for some $k = 1, 2, \ldots, N$. But the probability that $k = 1$ goes to zero as $N \to \infty$.

2.3.4. Arrival of transit customers. For each $v \in G$ we introduce the set $\mathcal{N}(v)$ of all vertices of G which are neighbors of v. Let the node $v' \in \mathcal{N}(v)$. Assume that a customer c^v of class $\varkappa$ located in the server at node v completes its service there. What are the chances that this customer goes to node v'? For this to happen it is necessary that

$$\operatorname{dist}(v, D(c^v)) = \operatorname{dist}(v', D(c^v)) + 1 \qquad \text{and} \qquad \operatorname{dist}(v', D(c^v)) > 0.$$

If there are several such nodes in $\mathcal{N}(v)$, then all of them have the same chance. If $\operatorname{dist}(v, D(c^v)) \le 1$, then customer c^v leaves the network immediately. Let $E(v, D(c^v))$ be the number of such nodes:

$$E(v, D(c^v)) = \#\{w \in \mathcal{N}(v) :\ \operatorname{dist}(v, D(c^v)) = \operatorname{dist}(w, D(c^v)) + 1\}.$$

Thus, for every pair $v, D \in V$ of sites with $\operatorname{dist}(v, D) \ge 2$, we define the function $e_{v,D}$ on the sites $w \in V$:

$$e_{v,D}(w) = \begin{cases} \dfrac{1}{E(v,D)} & \text{if } w \in \mathcal{N}(v) \text{ and } \operatorname{dist}(v, D) = \operatorname{dist}(w, D) + 1, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Then, in state μ, the rate of transits of customers arriving to node v' with class $\varkappa$ and destination $w \neq v'$ is given by

$$\begin{aligned} &\sigma_{\mathrm{tr}}(q, q \oplus c^{v'}(\varkappa, w)) \equiv \sigma^{\mu}_{\mathrm{tr}}(q, q \oplus c^{v'}(\varkappa, w)) \\ &\quad= \sum_{v \in \mathcal{N}(v')} \int d\mu(q_v)\, e_{v, D(C(q_v))}(v') \frac{\mathcal{F}'_{\varkappa,v}(\tau(C(q_v)))}{1 - \mathcal{F}_{\varkappa,v}(\tau(C(q_v)))} \delta(\bar{\varkappa}(C(q_v)), \varkappa) \delta(D(C(q_v)), w), \end{aligned} \tag{9}$$

with δ being the Kronecker delta function. Here $\bar{\varkappa}$ is the class that the customer $C(q_v)$ gets after its service is completed at v. Again, the τ-times of customers already present stay the same, and the newly arrived has its τ-time set to 0.

Note that the two rates (7) and (9) do depend on the measure μ, which is the source of the nonlinearity of our process.

2.4. Evolution Equations

For a warm-up we begin with the case where the measure $\mu = \prod \mu_v$ on $M = \prod\limits_v M_v$ has a nice density. The general situation will be treated below; see Proposition 2.

For $q \in M_v$ denote by $e(q)$ the last customer in queue q and by $l(q)$ the length of the queue. Note that $e(q)$ can also be denoted by $c_{l(q)}$ and that the quantity $\tau(e(q))$ denotes the amount of service that this customer has already received.

We then have

$$\frac{d}{dt}\mu_v(q_v,t)=\mathcal{A}+\mathcal{B}+\mathcal{C}+\mathcal{D}+\mathcal{E}, \tag{10}$$

where $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$, and $\mathcal{E}$ are operators acting on μ, described below.

The operator $\mathcal{A}$ corresponds to service progress:

$$\mathcal{A}=-\frac{d}{d\tau_{i^*(q_v)}(q_v)}\mu_v(q_v,t). \tag{11}$$

The operators $\mathcal{B}$ and $\mathcal{C}$ correspond to changes in queue q_v due to customers arriving from the outside and from other servers. For $\mathcal{B}$, we have

$$\mathcal{B}=\delta(0,\tau(e(q_v)))\mu_v(q_v\ominus e(q_v),t)[\sigma_{\mathrm{tr}}(q_v\ominus e(q_v),q_v)+\sigma_{\mathrm{e}}(q_v\ominus e(q_v),q_v)], \tag{12}$$

where q_v is obtained from $q_v\ominus e(q_v)$ by the arrival of $e(q_v)$ from v', and $\delta(0,\tau(e(q_v)))$ takes into account the fact that if the last customer $e(q_v)$ has already received some amount of service, then it cannot have just arrived from the outside or from another server (see (9) and (5)). For $\mathcal{C}$, we have

$$\mathcal{C}=-\mu_v(q_v,t)\sum_{q'_v}[\sigma_{\mathrm{tr}}(q_v,q'_v)+\sigma_{\mathrm{e}}(q_v,q'_v)]. \tag{13}$$

The operator $\mathcal{D}$ corresponds to service completions:

$$\mathcal{D}=\int\limits_{q'_v:\,q'_v\ominus C(q'_v)=q_v} d\mu_v(q'_v,t)\,\sigma_{\mathrm{f}}(q'_v,q'_v\ominus C(q'_v))-\mu_v(q_v,t)\sigma_{\mathrm{f}}(q_v,q_v\ominus C(q_v)), \tag{14}$$

where the first term describes the situation where the queue state q_v arises after a customer was served in the queue q'_v (longer by one unit) such that $q'_v\ominus C(q'_v)=q_v$, while the second term describes the completion of service of a customer in q_v.

Finally, the operator $\mathcal{E}$ corresponds to the exchange of the servers:

$$\mathcal{E}=\sum_{v'\in\mathcal{N}(v)}\beta_{vv'}[\mu_{v'}(q_v,t)-\mu_v(q_v,t)]. \tag{15}$$

Remark 1. Equation (10)–(15) has to be understood as follows: instead of talking about the evolution of measure $\mu(t)$, one has to consider the evolution of functionals of the measures. Thus, let b be a bounded function on M with bounded derivatives. Consider the functional $B(\mu)=\int\limits_M b\,d\mu$ on the state space. The equation should be understood as an equation on $B(\mu(t))$ of the form

$$\frac{d}{dt}B(\mu(t))=(\Omega(B))(\mu(t)),$$

for a certain operator Ω. This operator is written explicitly below, in equations (30)–(34). Since its exact expression is not important now, we postpone its presentation.

2.5. Main Theorem

Before stating the main result, we have to make some important assumptions and observations on the states of the network. Since we have to compare the networks $\mathcal{K}_N$ and the limiting network $\mathcal{K}$, we need to describe their states by probability distributions on the same space. This can be achieved thanks to the permutation symmetry of the networks $\mathcal{K}_N$. For this to hold, we assume that the initial state of $\mathcal{K}_N$ is $\prod_{v\in V} \mathcal{S}_{v,N}$-invariant, where each permutation group $\mathcal{S}_{v,N}$ permutes the N servers at the node v. Then the state has the same symmetry property at all later times. After taking the quotient by the action of the permutation group $(\mathcal{S}_N)$, the configuration at any vertex $v \in V$ can be described by the atomic probability measure

$$\Delta_N^v = \sum_{k=1}^{N} \frac{1}{N} \delta_{(q_{v,k},\tau)}$$

on M_v, where τ is the received service vector of customers in queue $q_{v,k}$. We define $\Delta_N = \{\Delta_N^v\}$.

We study the limit of $\mathcal{K}_N$ as $N \to \infty$. For this limit to exist, we need to choose initial states of the networks $\mathcal{K}_N$ appropriately. More precisely, we assume that the initial states ν_N of the networks $\mathcal{K}_N$, which are atomic measures on M with atom weights $1/N$, converge weakly to the state ν of the limiting network $\mathcal{K}$.

Below, a function $\mathcal{G}$ on the space of probability measures $\mathcal{P}$ will be said to be nice at infinity if $\mathcal{G}$ can be extended to a continuous function on the compactification $\bar{\mathcal{P}}$ of $\mathcal{P}$; see Section 5. We are now in a position to state the main result.

Theorem 1. *Let $S_{N,t}$ be the semigroup* $\exp\{t\Omega_N\}$ *associated with the infinitesimal generator Ω_N of network $\mathcal{K}_N$ described in Section* 2.1 *and formally defined in* (22)–(26). *Let S_t be the semigroup* $\exp\{t\Omega\}$ *associated with the infinitesimal generator Ω of network $\mathcal{K}$ defined in* (10)–(15) *for "nice" states and in* (30)–(34) *for the general case.*

1. *The semigroup S_t is well defined; i.e., for every measure ν on M, the trajectory $S_t(\nu)$ exists and is unique. In addition, this semigroup is Feller.*
2. *Let ν_N be the initial states of the networks $\mathcal{K}_N$. Then the measures $S_{N,t}(\nu_N)$ and $S_t(\nu_N)$ are close to each other in the following sense. For every function $\mathcal{G}$ on the space of the probability measures $\mathcal{P}$ which is continuous in the weak topology and is nice at infinity, we have*

$$\lim_{N\to\infty} \sup_{\nu_N} |\mathcal{G}(S_{N,t}(\nu_N)) - \mathcal{G}(S_t(\nu_N))| = 0.$$

Remark 2. The initial state ν_N is a measure on M (an atomic one). It is a usual initial configuration of a usual Markov process. The measure $S_{N,t}(\nu_N)$, $t > 0$, is a random measure on M. However, the measure $S_t(\nu_N)$ is a *nonrandom* measure on M. Our theorem is a statement of the same type as the law of large numbers for the random measure $S_{N,t}(\nu_N)$. One can also say that if the measures $\nu_N \to \nu$ weakly, then for each $t > 0$ we also have $S_{N,t}(\nu_N) \to S_t(\nu)$ weakly.

3. THE NONLINEAR MARKOV PROCESS: EXISTENCE

In this section we prove Statement 1 of Theorem 1.

Our nonlinear Markovian evolution is a jump process on $M = \prod_v M_v$ with piecewise continuous trajectories $\{q_v : v \in V\}$. Between jumps, the states $q_v \in M_v$ only change in that the $i^*(q_v)$ coordinate of the τ vector of q_v increases at unit speed, in a deterministic way. The queue states $q_v \in M_v$ can also perform various jumps, as is described in Section 2.4. (For the reader concerned with the treatment of the case of V finite, we refer to the standard technique of the interacting particle systems theory; see, e.g., [9].)

Theorem 2. *Under the above symmetry assumptions, for every initial states $\mu(0) = \prod \mu_v(0)$, equation* (10)–(15) *has a solution, which is unique.* (*The equations are understood in the sense of Remark* 1.)

Proof. Assume that the theorem holds true. Then we deduce from the form of the generator that, for all v' and v, the rates $\bar{\lambda}_{v'v}(t)$ of the Poisson point processes of arrivals of customer transiting from the servers at v' to a server at v coincide with the rates $\bar{b}_{v'v}(t)$ of the departure point processes, from tagged servers at nodes v', of customers sent to v (note that these departure processes are not Poisson in general).

Thus, let us look for an operator transforming the input point processes to exit point processes, for which the rates $\bar{\lambda}_{v'v}(t)$ are a fixed point. With that idea in mind, consider an auxiliary system on the same set of servers with the same initial condition $\mu(0)$. Instead of internal Poisson flows of the initial system with rates $\bar{\lambda}_{v'v}(t)$ (which are (hypothetically) determined uniquely by $\mu(0)$), we consider, for each node v and each of its nearest neighbor v', an arrival Poisson flow of customers with an arbitrary rate function $\lambda_{v'v}(t)$. The result of the service at v will then be a collection of (individually non-Poisson) departure flows to certain nodes v''. Since the service time distributions have bounded densities (see Section 2.3.2), the following limit exists:

$$b_{vv''}(t) = \lim_{\Delta t \to 0} \frac{\mathbf{E}(\text{number of customers arrived from } v \text{ to } v'' \text{ in } [t, t+\Delta t])}{\Delta t}.$$

These functions $b_{vv''}(t)$ are rate functions of non-Poisson departure flows. Thus, we have an operator $\Psi_{\mu(0)}$, which transforms the collection $\lambda = \{\lambda_{v'v}(t)\}$ to $b = \{b_{v'v}(t)\}$. Our theorem about the existence and uniqueness will follow from the fact that the map $\Psi_{\mu(0)}$ has a unique fixed point, $\bar{\lambda}$. Note that the rate functions λ and b depend not only on the nodes v and v' but also on the class of customers. Below we often omit some coordinates of these vectors, but we always keep them in mind.

By the same properties of the service times distributions, the functions b are continuous and uniformly bounded. Moreover, without loss of generality we may assume that they are Lipschitz, with a Lipschitz constant ℓ which depends only on $\mathcal{F}$, the supremum of the service rates. Since we look for fixed points, we may assume that the functions λ are bounded as well, and also that they are integrable and Lipschitz, with the same Lipschitz constant. Thus, we restrict the functions λ to be in the class of functions denoted by $L_\ell[0,T]$. In the case of G finite, we put the L_1 metric on our functions, i.e.,

$$\int_0^T |\lambda^1(\tau) - \lambda^2(\tau)|\, d\tau \equiv \sum_{v'v} \int_0^T |\lambda^1_{v'v}(\tau) - \lambda^2_{v'v}(\tau)|\, d\tau.$$

Note that this metric turns $L_\ell[0,T]$ into a complete compact metric space, by the Arzelà–Ascoli theorem. For countably infinite G, we choose an arbitrary vertex $v_0 \in V$ as a "root" and define likewise

$$\int_0^T |\lambda^1(\tau) - \lambda^2(\tau)|\, d\tau \equiv \sum_{v'v} \exp\{-2D(G)[\operatorname{dist}(v_0, v') + \operatorname{dist}(v_0, v)]\} \int_0^T |\lambda^1_{v'v}(\tau) - \lambda^2_{v'v}(\tau)|\, d\tau,$$

where the finiteness of the sum follows from a simple counting argument. We recall that $D(G)$ is the maximal degree in G. The topology on $L_\ell[0,T]$ thus defined is equivalent to the Tikhonov topology; in particular, $L_\ell[0,T]$ is again a complete compact metric space.

The contraction arguments below cover only the case of finite G; the generalization to infinite G is discussed at the end of the proof.

We now show that for every $\mu(0)$, the map $\Psi_{\mu(0)}$ is a contraction on $L_\ell[0,T]$; by the Banach theorem, this will imply the existence and uniqueness of the fixed points for $\Psi_{\mu(0)}$. Without loss of generality we may assume that T is small (since we can iterate our argument).

Let $\{\lambda_v^1(t), \lambda_v^2(t) : t \in [0,T], v \in V\}$ be two versions of the collection λ of Poisson inflows to our servers; assume that for all v

$$\int_0^T |\lambda_v^1(\tau) - \lambda_v^2(\tau)|\, d\tau < \Lambda.$$

We want to estimate the difference $b_0^1(t) - b_0^2(t)$ of the rates of the departure flows at some tagged node $0 \in V$; we show that the difference will be much smaller. Clearly, this will be sufficient. The delicate point here is that the rates $b_0^i(\cdot)$ depend also on the rates $\lambda_v^i(\cdot)$ at $v \neq 0$, due to the possibility of servers jumps; after such jumps, the state at node 0 is replaced by that at the neighboring node.

Let $\tau_1 < \tau_2 < \ldots < \tau_k \in [0,T]$, $k = 0,1,2,\ldots$, be the (random) epochs when the state at 0 is replaced by the state at a neighboring node, due to server jumps. We now derive an estimate on $\int_0^T |b_0^1(t) - b_0^2(t)|\, dt$ under the condition that the number k and the epochs $\tau_1 < \tau_2 < \ldots < \tau_k$ are fixed; since our estimate will be uniform in the conditioning, this will be sufficient. Note that the probability to have k swaps during the time T is bounded from above by $(T\beta)^k$ (see (3)).

Informally, the contraction takes place because the departure rates $b(t)$ for $t \in [0,T]$ with T small depend mainly on the initial state $\mu(0)$: the new customers arriving during time $[0,T]$ have little chance to be served before T if there are customers already waiting. Therefore the "worst" case for us is where in the initial state the server at node 0 is empty, i.e., the measure $\mu_0(0)$ is equal to the measure δ_0, with a unit atom at the empty queue $\varnothing$.

A. Let us start the proof by considering the case where no swaps of the server at $v = 0$ happen, i.e., $k = 0$. Let $\lambda^1(t), \lambda^2(t)$, $t \in [0,T]$, be the rates of two collections of Poisson inflows to the empty server. We want to estimate the distance $\int_0^T |b^1(t) - b^2(t)|\, dt$ between the rates of the departure flows from this server. For this we use a classical coupling between two Poisson inflows.

Consider the integral

$$I_T^\lambda \equiv I_T^{\{\lambda_{v'0}\}} = \sum_{v' \sim 0} \int_0^T |\lambda_{v'0}^1(t) - \lambda_{v'0}^2(t)|\, dt. \tag{16}$$

For every value of the index v', let us consider the region between the graphs of the functions

$$H_{v'}(t) = \max\{\lambda_{v'0}^1(t), \lambda_{v'0}^2(t)\} \qquad \text{and} \qquad h_{v'}(t) = \min\{\lambda_{v'0}^1(t), \lambda_{v'0}^2(t)\}.$$

Let us introduce the auxiliary family of independent rate-one Poisson point processes ω_v, $v \in V$, taking values in $\mathbb{R}_+^2$. Then the integral

$$\int_0^T |\lambda_{v'0}^1(t) - \lambda_{v'0}^2(t)|\, dt$$

is just the expectation of the number of Poisson points $\omega_{v'} \in \mathbb{R}_+^2$ falling between the graphs of $H_{v'}$ and $h_{v'}$ up to time T. Let us introduce the corresponding counting random variable:

$$\alpha_{v'} = \operatorname{card}\{\omega_{v'} \cap \{(s,t) : t \in [0,T],\ h_{v'}(t) \le s \le H_{v'}(t)\}\}.$$

Using the process $\omega_{v'}$, we now define a coupling between the two Poisson inflows. Let us declare the points of $\omega_{v'}$ falling below $h_{v'} = \min\{\lambda_{v'0}^1(t), \lambda_{v'0}^2(t)\}$ as colorless; the points of $\omega_{v'}$ falling between $\lambda_{v'0}^1(t)$ and $h_{0'}(t)$ are declared red, while the points of $\omega_{v'}$ falling between $\lambda_{0'v}^2(t)$ and $h_{0'}(t)$ get the blue color. To every point of $\omega_{v'}$ falling below $H_{v'}$ and having abscissa t, we associate a customer arriving to 0 at time t. Then the colorless plus red customers represent the first Poisson inflow, while the colorless plus blue customers represent the second. This defines defines the coupling between the two input flows. The variable $\alpha_{0'}$ is thus the number of colored customers.

On the event that none of the arrival flows have colored customers, the two departure flows are identical. Therefore, for all v'', the contribution of this event to the integral $\int_0^T |b^1_{0v''}(t) - b^2_{0v''}(t)|\,dt$ is zero. Let $\mathcal{C}$ be the complement of this event; its probability is of order I^λ_T when T is small. Under the condition of the arrival of the colored customer $\bar{c}$, the conditional probability that this customer will affect the departure flow is of the order of $\mathcal{F}T$ (see (6)). Indeed, under this condition, the departure process is affected if either customer $\bar{c}$ is served and departs from the server at 0 in $[0,T]$, or the arrival of $\bar{c}$ prevents the departure of some customer with lower priority whose service was going to be completed in $[0,T]$. Hence, for T small, the (unconditional) probability of $\mathcal{C}$ is $T\mathcal{F}I^\lambda_T$, and therefore its contribution to the integral $\int_0^T |b^1_{0v''}(t) - b^2_{0v''}(t)|\,dt$ is bounded from above by $\mathrm{const}\cdot T\mathcal{F}I^\lambda_T$, which is $\ll \Lambda$.

In the case of a nonempty initial queue the situation is even simpler, since we have more uncolored customers.

B. Consider now the case where the server swaps, i.e., $k > 0$. We argue that the contribution to $\int_0^T |b^1_{0v''}(t) - b^2_{0v''}(t)|\,dt$ of the event that k server swaps happen at node $v = 0$ is of order T^{k+1}.

First consider the case $k = 1$, and let us condition on the event that $\tau_1 = \tau < T$ and the swap of $v = 0$ is with the nearest neighbor node w. Thus, we need to compare the evolution at $v = 0$ defined by the Poisson inputs with rates $\{\lambda^1_{v'0}(t)\}$ and $\{\lambda^1_{w'w}(t)\}$ with that defined by the rates $\{\lambda^2_{v'0}(t)\}$ and $\{\lambda^2_{w'w}(t)\}$. We use the same coupling between pairs of flows as that defined above. Before time τ, the server at $v = 0$ behaves as is described in **A**. At time τ the pair of states $\mu^i_0(\tau)$, $i = 1, 2$, that we compare is replaced by that sampled independently from $\mu^i_w(\tau)$, which then evolve according to the inflows $\{\lambda^i_{v'0}(t)\}$, $i = 1, 2$. All the arguments in **A** apply to the node w as well. Thus, under the foregoing conditions, $\int_0^T |b^1_{0v''}(t) - b^2_{0v''}(t)|\,dt$ is again bounded from above by $\mathrm{const}\cdot T\mathcal{F}\max\bigl(I^{\{\lambda_{v'0}\}}_T, I^{\{\lambda_{w'w}\}}_T\bigr)$, which is $\ll \Lambda$. Here $I^{\{\lambda_{w'w}\}}_T$ is defined as in (16), with 0 replaced by w.

The unconditioning over τ brings an extra factor $T\beta$, so the corresponding contribution to $\int_0^T |b^1_{0v''}(t) - b^2_{0v''}(t)|\,dt$ is $\mathrm{const}\cdot T^2\beta\mathcal{F}\max(I^{\lambda_{*0}}_T, I^{\lambda_{*w}}_T) \ll T\Lambda$.

The case of general values of k follows the same lines.

C. The contraction property established above shows that the initial condition $\mu(0) = \{\mu_v(0)\}$ uniquely defines all the inflow rates $\lambda_{v'v}(t)$, for all $t \in [0,T]$. If the graph G is finite, this implies the uniqueness of the measures $\mu(t)$. For infinite graphs, it can in principle happen that different "boundary conditions"—i.e., different evolutions of μ "at infinity"—are a source of nonuniqueness. However, the argument of Part **B** shows that the influence on the origin $0 \in V$ from the nodes at distance R during time T is of the order of T^R, provided that the degree of G is bounded. Therefore, the uniqueness holds for infinite graphs as well. △

Proposition 1. *The semigroup defined by equations* (10)–(15) *is Feller.*

Proof. Consider the inflow rates $\{\bar{\lambda}_{v'v}(t)\}$ built above from the initial state μ_0. Thanks to the properties of service times (see Section 2.3.2), the associated departure flow rates, $\{\bar{b}_{v'v}(t)\}$, are continuous functions of time t. As was already mentioned, the departure flow rates coincide with the inflow rates: $\bar{\lambda}_{v'v}(t) = \bar{b}_{v'v}(t)$ for all v, v', and t. Hence, the functions $\{\bar{\lambda}_{v'v}(t)\}$ are continuous. This in turn implies that the trajectories μ_t are continuous as well, in the weak topology. Obviously, the departure rates $\{\bar{b}_{v'v}(t)\}$ are continuous in the initial state μ_0. Therefore, the dependence of the rates $\bar{\lambda}_{v'v}(t)$ on μ_0 is continuous. Thus, the map $\mu_0 \rightsquigarrow \mu_t$ is continuous. △

4. COMPACTIFICATIONS

In order to study the convergence of our mean-field type networks $\mathcal{K}_N$ to the limiting nonlinear Markov process network $\mathcal{K}$, we want the latter to be defined on a compact state space. This means that we have to add to the graph G the sites G_∞ lying at infinity, obtaining some extended graph $\bar{G} = G \cup G_\infty$, and to allow infinite queues at each node $v \in \bar{G}$. We then have to extend our dynamics to this bigger state space system. The way it is chosen among several natural options is of small importance, since, as we will show, if the initial state of our network assigns zero probability to various infinities, then the same holds for all finite times.

The compactification only plays a technical role here. It allows us to use some standard theorems of convergence of Markov processes. The benefits it brings is that certain observables can be continuously extended to a larger space, see Section 5.

4.1. Compactification $\bar{G}$ of the Graph G

The compactification that we define here is adapted to the network type considered. It uses the fact that if a customer c is located at v and its destination $D(c)$ is w, then the path c taken in order to get from v to w is obtained using the greedy algorithm described above: c chooses, uniformly among all the nearest neighbor (n.n.) sites, the site which brings it one unit closer to its destination.

To define the compactification, we proceed as follows. Let $\gamma = \{\gamma_n \in V\}$ be an n.n. path on G. We want to define a notion of existence of the limit $L(\gamma) = \lim_{n\to\infty} \gamma_n$. If the sequence γ_n stabilizes, i.e., if $\gamma_n \equiv g \in V$ for all n large enough, we define $L(\gamma) = g$. To proceed, for any $v \in V$ we define the Markov chain P^v on G. It is an n.n. random walk such that at each step the walk makes its distance to v to decrease by 1. If there are several such choices, one is chooses uniformly at random. Therefore, the transition probabilities $P^v(u, w)$ are given by the function $e_{u,v}(w)$ defined in (8). If T is some positive integer and u is at distance more than T from v, then we denote by $P_u^{v,T}$ the T-step probability distribution on the trajectory of length T starting at u and heading towards v.

Consider now an infinite n.n. path γ. We say that the limit $L(\gamma) = \lim_{n\to\infty} \gamma_n$ exists if for all $u \in V$ and all T, the limit $\lim_{n\to\infty} P_u^{\gamma_n,T}$ exists. In words, this means that, seen from u, the points γ_n and γ_m are "in the same direction," for n, m large enough. For two paths γ', γ'', we say that $L(\gamma') = L(\gamma'')$ if and only if both limits exist and moreover, for all $u \in V$ and all T, the measures $P_u^{\gamma'_n,T}$ and $P_u^{\gamma''_n,T}$ coincide for all $n \geq n(u, T, \gamma', \gamma'')$ large enough.

Consider the union $\bar{V} = V \cup V^\infty \equiv V \cup \{L(\gamma) : \gamma \text{ is an n.n. path on } G\}$. It is easy to see that the natural topology on $\bar{V}$ makes it into a compact. To give an example, consider the case $G = \mathbb{Z}^2$. Let $f\colon \mathbb{Z}^2 \to \mathbb{R}^2$ be the following embedding: $f(n,m) = \Big(\operatorname{sgn}(n)\Big(1 - \dfrac{1}{|n|}\Big), \operatorname{sgn}(m)\Big(1 - \dfrac{1}{|m|}\Big)\Big)$. Then the closure $\bar{V}$ will be the closure of the image of f in $\mathbb{R}^2$.

We want to build a graph $\bar{G}$ with the set $\bar{V}$ as a vertex set. To do this, we need to specify pairs of vertices which are connected by an edge. If v', v'' both belong to V, then they are connected in $\bar{G}$ if and only if they are connected in G. If v' is in V while $v'' \in V^\infty$, then they are never connected. Finally, if $v', v'' \in V^\infty$, then they are connected if and only if one can find a pair of paths $\gamma', \gamma'' \to \infty$ such that $L(\gamma') = v'$, $L(\gamma'') = v''$, and the sites $\gamma'_n, \gamma''_n \in V$ are n.n. In particular, every vertex $v \in V^\infty$ has a loop attached. Note that the graph $\bar{G}$ is not connected.

4.2. Extension of the Network to $\bar{G}$

This is done in a natural way. Now we have servers and queues also at "infinite" sites. Note that the customers at infinity cannot get to the finite part G of $\bar{G}$. Also, the customers from G cannot get to G^∞ in finite time.

4.3. Compactification of M_v

In order to compactify the manifold M_v, $v \in V(G)$, into the compact set $\bar{M}_v$, we have to add to it various "infinite objects": infinite words, infinite waiting times, and infinite destinations. Since the destinations of the customers in our network are vertices of the underlying graph, the compactification of M_v will depend on it. For the last question, we use the graph $\bar{G}$, with the topology introduced earlier. The manifold of possible queues at v is the disjoint union of the positive orthants $\mathbb{R}_w^+$, where w is a finite word describing the queue $q_v = \{c_i\} \equiv \{c_i^v\}$ at v. Letters of the alphabet, making the word w, are pairs $(\varkappa_i, v_i)$, where $\varkappa_i \in K$ is the customer class and $v_i = D(c_i) \in \bar{G}$ is its destination. Thus, our alphabet is compact. We have to compactify the set W of finite ordered words w by adding infinite words to it.

To do this, we denote by $O(w)$ the reordering of w corresponding to the order of service of the queue w, which is defined by the service discipline R_v. This is just a permutation of w. We say that a sequence w_i of finite words converges as $i \to \infty$ if and only if the sequence of words $O(w_i)$ converges coordinatewise. If this is the case, we denote $\bar{O} = \lim\limits_{i\to\infty} O(w_i)$ and say that $\lim\limits_{i\to\infty} w_i = \bar{O}$. We denote by $\bar{W}$ the set of all finite words, $w \equiv (w, O(w))$, supplemented by all possible limit points $\bar{O}$. We define the topology on $\bar{W}$ by saying that a sequence $w_i \in \bar{W}$ is converging if and only if the sequence $O(w_i)$ converges coordinatewise. In other words, we put on $\bar{W}$ the Tikhonov topology. Since K is finite, $\bar{W}$ is compact in the topology of pointwise convergence.

According to what was said in Section 2.2, our service discipline (and hence the function O) has the following property. Let the sequence $w_i \in W$ of finite words converge in the above sense. Let c be a customer, and consider the new sequence $w_i \cup c \in W$, where customer c is the last arrived. Then we have the implication

$$\lim_{i\to\infty} O(w_i) \text{ exists} \implies \lim_{i\to\infty} O(w_i \cup c) \text{ exists}.$$

The continuity of the transition probabilities in this topology on the set of queues is easily seen; indeed, the closeness of two queues q and q' means that the first k customers served in both of them are the same. But then the transition probabilities $P_T(q, \cdot)$ and $P_T(q', \cdot)$ differ by $o(T^k)$. Thus, the extended process is Feller, as well as the initial one.

The compactifications $\bar{\mathbb{R}}_w^+$ of the orthants $\mathbb{R}_w^+$ are defined in an obvious way: they are products of $|w|$ copies of the compactifications $\bar{\mathbb{R}}^+ = \mathbb{R}^+ \cup \infty$. For the infinite words $\bar{O}$ we consider infinite products, in the Tikhonov topology. The notion of convergence in the union $\bigcup\limits_{w\in\bar{W}} \bar{\mathbb{R}}_w^+$ is that of coordinatewise convergence.

The properties of service times formulated in Section 2.3.2 allow us to extend the relevant rates in a continuous way to a function on $\tau \in \bar{\mathbb{R}}^+$. Moreover, an analog of Proposition 1 holds.

5. PROOF OF CONVERGENCE

This section contains the proof of Statement 2 in Theorem 1. Let $\Omega_N, \Omega\colon X \to X$ be (unbounded) operators on the Banach space X. We are looking for conditions ensuring the convergence of the semigroups $\exp\{t\Omega_N\} \to \exp\{t\Omega\}$ on X as $N \to \infty$.

We will use the following version of the Trotter–Kurtz theorem (which is Theorem 6.1 in [10, ch. 1], with the core characterization taken from Proposition 3.3 of the same ch. 1, where a *core* of Ω is a dense subspace $\bar{X} \subset X$ such that $\forall F \in \bar{X}$

$$\exp\{t\Omega\}(F) \in \bar{X}. \tag{17}$$

Theorem 3 (Trotter–Kurtz). *Let $X^1 \subset X^2 \subset \cdots \subset X$ be a sequence of subspaces of the Banach space X. We assume that there exist projectors $\pi^N\colon X \to X^N$ such that for every $f \in X$ we have $f^N = \pi^N(f) \to f$ as $N \to \infty$. Assume that each of the semigroups $\exp\{t\Omega_N\}$ defined on X^N is a* *strongly continuous contraction. If for every F in the core $\bar{X}$ of Ω we have*

$$\|\Omega_N(F^N) - \Omega(F^N)\| \to 0, \tag{18}$$

then for every $f \in X$

$$\exp\{t\Omega_N\}f^N \to \exp\{t\Omega\}f. \tag{19}$$

We now show how to apply this theorem to our setting. The main ideas of this application were developed in [11].

We take for function space X the space $\mathcal{C}(\bar{\mathcal{P}})$ of continuous functions on $\bar{\mathcal{P}}$, with $\bar{\mathcal{P}}$ the compactified version of $\mathcal{P}$ in the weak topology.

The strong continuity of the semigroups $\exp\{t\Omega_N\}$ is straightforward, while that for the semigroup $\exp\{t\Omega\}$ follows from the fact that the trajectories $S_t\mu$ are continuous in t. After compactification, the set of probability measures $\mu \in \mathcal{P}$ on the Comb is replaced by the set of probability measures on its compactification and becomes a compact, $\bar{\mathcal{P}}$, on which the family $S_t\mu$ is equicontinuous. This implies the strong continuity.

It is clear that the operation $\Omega_N(F^N)$ consists of computing finite differences for the function F at some atomic measures, while $\Omega(F^N)$ consists of computing derivatives at the same atomic measures. We will have the convergence (18) when the derivatives exist and can be approximated by finite differences. Therefore, for our space $\bar{X}$ we need differentiable functions; we remind the reader that in the case of an infinite graph G, the quasi-locality property of all functions considered is always assumed. To define $\bar{X}$, we first introduce the norm $\|\cdot\|_1$ on $\bar{\mathcal{P}}$. Let $\mathcal{C}(\bar{M}_v)$ be the space of continuous functions on $\bar{M}_v$ with the sup-norm, $\|\cdot\|$. Let $\mathcal{C}_1(\bar{M}_v) \subset \mathcal{C}(\bar{M}_v)$ be the subspace of differentiable functions having finite norm $\|f\|_1 = \|f\| + \|f'\|$. The space $\bar{\mathcal{P}}$ of product probability measures on $\prod \bar{M}_v$ belongs to the dual space $\prod \mathcal{C}_1^*(\bar{M}_v)$, and we define the norm $\|\cdot\|_1$ on $\bar{\mathcal{P}}$ to be the restriction of the natural norm on $\prod \mathcal{C}_1^*(\bar{M}_v)$.

For the *core* $\bar{X}$ of the operator Ω, we take the set of functions F of the measures $\mu \in \bar{\mathcal{P}}$ satisfying the following properties:

- F is uniformly differentiable (in the sense of Fréchet with respect to the norm $\|\cdot\|_1$);
- F belongs to the domain of the generator $\widehat{\sigma}$ of the time-shift semigroup;
- F is exponentially quasilocal.

The last property means the following. Let ρ be the graph metric on G; if G is not connected (which is the case for our compactification) it can take the value ∞. F is called exponentially quasilocal if there exists a vertex $x \in G$ and a constant $c(F)$ such that for any $\mu \in \bar{\mathcal{P}}$ and any two measures $h_1, h_2 \in \bar{\mathcal{P}}$ supported by

$$\prod_{v\colon \rho(v,x)>R} \bar{M}_v$$

we have $|F(\mu) - F(\mu + h_1 - h_2)| < \exp\{-c(F)R\}$.

We check (18) in Section 5.1. We then need to check that this space $\bar{X}$ of uniformly differentiable functions is preserved by the semigroup $\exp\{t\Omega\}$ (core property). The function $\exp\{t\Omega\}F(\mu)$ is just $F(S_t\mu)$, so if F is differentiable in μ, then the differentiability of $F(S_t\mu)$ follows from that of $S_t\mu$.

To estimate the norm of the Fréchet differential, we have to consider a starting measure $\mu^1 = \prod \mu_v(t=0)$, its perturbation $\mu^2 = \prod(\mu_v + h_v)$ with $h = \{h_v\}$ nontrivial for finitely many v (say), the perturbation having norm $\approx \sum\limits_v \|h_v\|$; then take the difference $\mu^2(T) - \mu^1(T)$ and write it as

$$\mu^2(T) - \mu^1(T) = \Phi(T, \mu^1)h + o(\|h\|)h; \tag{20}$$

and finally show that the norm of the operator $\Phi(T, \mu^1)$ is finite.

Note first that it suffices to prove this for T small, the smallness being uniform in all relevant parameters.

Now, if the increment h is small (even only in the (weaker) $\|\cdot\|_1$ sense, i.e., only from the point of view of the smooth functions—like, for example, a small shift of a δ-measure), then all the flows in our network started from μ^1 and μ^2 differ only a little during a short time, its shortness being a function of the service time distributions only. The amplitude of the flow difference is of the order of $D \times \mathcal{F} \times \|h\|_1$, where D is the maximal degree of the graph G, and $\mathcal{F}$ is defined by (6). In fact, it can be smaller; it is attained for the situations where a server, being empty in the state μ^1, becomes nonempty in the h-perturbed state. The order is computed in the stronger sup-norm $\|\cdot\|$. Therefore, after time T the norm of the difference is such that

$$\|\mu^2(T) - \mu^1(T)\| \leq \|h\| + T \times D \times \mathcal{F} \times \|h\|, \tag{21}$$

which explains our claim about the operator $\Phi(T, \mu^1)$. We will give a formal proof in Section 5.2.

5.1. Generator Comparison

We start with the mean-field type network, made of N copies of the initial network.

We first describe the process as a process of the network containing $N|G|$ servers, and then do the factorization by the product of $|G|$ permutation groups $\mathcal{S}_N$.

The former one will be described only briefly. At each of the $N|G|$ servers, there is a queue of customers. Some of the queues can be empty. As time goes, the queues evolve due to (1) arrivals of external customers; (2) end of service of a customer which then leaves the network; (3) end of service of a customer which then moves to the next server; (4) interchange of two servers. Each of these events leads to a jump of our process. If none of them happens, the process evolves linearly: the variables $\tau(C(q_v))$ grow with rate 1.

Consider the semigroup S_N and its generator Ω_N; their existence is straightforward. Ω_N acts on functions F on measures $\mu_N \in \mathcal{P}_N$ which are atomic with atoms of weight $\dfrac{1}{N}$.

Our goal is now the following. Let $\mu_N \to \mu$ weakly, and let F be a smooth function on measures. Let us look at the limit $\Omega_N(F)(\mu_N)$ and the value $\Omega(F)(\mu)$. For F smooth, we can replace certain differences by derivatives; after this, we will see the convergence $\Omega_N(F)(\mu_N) \to \Omega(F)(\mu)$ in a transparent way.

For this, we apply the multiline formula (22)–(26) for the operator Ω_N to a function F. On each M_v, we have to take a probability measure Δ_N^v of the form $\sum\limits_{k=1}^{N} \dfrac{1}{N}\delta_{(q_{v,k},\tau)}$, where τ is the amount of service already received by customer $C(q_{v,k})$ of $q_{v,k}$ (at the position $i^*(q_{v,k})$ in queue $q_{v,k}$). Let $\Delta_N = \{\Delta_N^v\}$. Then

$$(\Omega_N(F))(\Delta_N) = \sum_v \sum_k \frac{\partial F}{\partial r(q_{v,k}, \tau(C(q_{v,k})))}(\Delta_N) \tag{22}$$

$$+ \sum_v \sum_k \sum_{v'\,\mathrm{n.n.}\,v} \sum_{k'} \frac{1}{N} e_{v,D(C(q_{v,k}))}(v')\sigma_{\mathrm{f}}(q_{v,k}, q_{v,k} \ominus C(q_{v,k}))$$

$$\times [F(J_{v,v';k,k'}(\Delta_N)) - F(\Delta_N)] \tag{23}$$

(where, for a directed edge v, v' and for all pairs of queues $q_{v,k}$, $q_{v',k'}$, we denote by $J_{v,v';k,k'}(\Delta_N)$ the new atomic measure which is the result of the completion of the service of customer $C(q_{v,k})$ in queue $q_{v,k}$ at v and its subsequent jump into queue $q_{v',k'}$, increasing thereby the length of the queue $q_{v',k'}$ at v' by one)

$$+ \sum_v \sum_k \sum_{v'} \sum_{\varkappa} \lambda(\varkappa, v, v')[F(\Delta_N - \frac{1}{N}\delta_{(q_{v,k})} + \frac{1}{N}\delta_{(q_{v,k} \oplus c^v(\varkappa,0,v'))}) - F(\Delta_N)] \tag{24}$$

(here we see the arrival of a new customer of class $\varkappa$ with a destination v')

$$+\sum_v \sum_k \sum_{v'\colon \operatorname{dist}(v,v')\le 1} \delta(D(C(q_{v,k})), v')\sigma_{\mathrm{f}}(q_{v,k}, q_{v,k} \ominus C(q_{v,k})) \times [F(J_{v;k}(\Delta_N)) - F(\Delta_N)] \tag{25}$$

(here we account for the customers which leave the network; the operator $J_{v;k}(\Delta_N)$ denotes the new atomic measure which is the result of the completion of the service of customer $C(q_{v,k})$ in queue $q_{v,k}$ at v and its subsequent exit from the system: $J_{v;k}(\Delta_N) = \Delta_N - \frac{1}{N}\delta_{(q_{v,k})} + \frac{1}{N}\delta_{(q_{v,k} q_{v,k} \ominus C(q_{v,k}))})$

$$+\sum_v \sum_k \sum_{v'\,\mathrm{n.n.}\,v} \sum_{k'} \frac{1}{N}\beta_{vv'}[F(T_{v,k;v',k'}\Delta_N) - F(\Delta_N)] \tag{26}$$

(here the operator $T_{v,k;v',k'}$ acts on Δ_N by exchanging the atoms $\frac{1}{N}\delta_{(q_{v,k})}$ and $\frac{1}{N}\delta_{(q_{v',k'})}$).

Remark 3. If our graph is infinite, the sums in (22)–(26) are infinite. However, they make sense for *local* functions F, as well as for *quasilocal* ones, which depend of far-away nodes (exponentially) weakly.

Now we want to pass in the above formula to a formal limit, obtaining thus a (formal) expression for the limiting operator Ω. It acts on functions F on probability measures on $\prod_v M_v$. To do this, we need to introduce extra notation.

There are two natural maps between the spaces $\mathbb{R}^+_w$ introduced in Section 2.4. One is the embedding

$$\chi\colon \mathbb{R}^+_w \to \mathbb{R}^+_{w\cup c} \tag{27}$$

corresponding to the arrival of the new customer c; it is given by $\chi(x) = (x, 0)$. The other one is the projection

$$\psi\colon \mathbb{R}^+_w \to \mathbb{R}^+_{w\smallsetminus c_{i^*(x)}} \tag{28}$$

corresponding to the completion of the service of the customer $c_{i^*(x)}$ currently served. It is given by $\psi(x) = (x_1, \ldots, x_{i^*(x)-1}, x_{i^*(x)+1}, \ldots, x_{|w|})$. For $|w| = 0$ the space $\mathbb{R}^+_\varnothing$ is a point, and the map $\psi\colon \mathbb{R}^+_\varnothing \to \mathbb{R}^+_\varnothing$ is the identity. The third natural map $\zeta_{vv'}\colon M_v \times M_{v'} \to M_v \times M_{v'}$ is defined for every ordered pair v, v' of neighboring nodes. It corresponds to the jump of a customer which has completed its service at v to v', where it is going to be served next. It is defined as follows: if the destination $D(C(q_v))$ of the attended customer $C(q_v)$ of the queue q_v is different from v', then

$$\zeta_{vv'}(q_v, q_{v'}) = (q_v \ominus C(q_v), q_{v'} \oplus c(C(q_v))), \tag{29}$$

where the customer $c(C(q_v))$ has the following properties:

1. $D(c(C(q_v))) = D(C(q_v))$;
2. $\varkappa(c(C(q_v))) = \mathcal{T}(\varkappa(C(q_v)); v, v')$;
3. $\tau(c(C(q_v))) = 0$.

If $D(C(q_v)) = v'$ or if $q_v = \varnothing$, then we put $\zeta_{vv'}(q_v, q_{v'}) = (q_v, q_{v'})$.

Let $\Delta = \lim_{N\to\infty} \Delta_N$. We will use the differentiability of F at Δ, i.e., the existence of the Fréchet differential F' at Δ. This differential will be denoted by $F'_\Delta(\cdot)$; it is a linear functional on the space of tangent vectors to $\mathcal{P}$ at $\Delta \in \mathcal{P}$. Due to the existence of the differential, we have another multiline expression (30)–(34):

$$(\Omega(F))(\Delta) = (\widehat{\sigma}(F))(\Delta) + \sum_v \sum_{v'\,\mathrm{n.n.}\,v} e_{v,D(C(\cdot))}(v')\sigma_{\mathrm{f}} \times F'_\Delta(\zeta_{vv'}(\Delta) - \Delta) \tag{30}$$

(where $\widehat{\sigma}$ is the generator of the time-shift semigroup, $\zeta_{vv'}(\Delta)-\Delta$ is a (signed) measure (see (29)), and $e_{v,D(C(\cdot))}(v')\sigma_{\mathrm{f}}\times(\zeta_{vv'}(\Delta)-\Delta)$ denotes the measure having density $e_{v,D(C(q))}(v')\sigma_{\mathrm{f}}(q_v, q_v\ominus C(q_v))$ with respect to $(\zeta_{vv'}(\Delta)-\Delta)(dq)$)

$$+\sum_v\sum_{v'}\sum_{\varkappa}\lambda(\varkappa,v,v')F'_\Delta(\chi_{v,v';\varkappa}(\Delta)-\Delta) \tag{31}$$

(here $\chi_{v,v';\varkappa}\colon M_v\to M_v$ is the embedding corresponding to the arrival to v of an external customer of class $\varkappa$ and destination v'; see (27))

$$+\sum_v F'_\Delta(\sigma_{\mathrm{f}}\times(\psi^v_{nn}(\Delta)-\Delta)) \tag{32}$$

(here

$$\psi^v_{nn}(q)=\begin{cases}\psi^v(q) & \text{for } q \text{ with } \operatorname{dist}(v,D(C(q)))\le 1,\\ q & \text{for } q \text{ with } \operatorname{dist}(v,D(C(q)))>1,\end{cases} \tag{33}$$

where $\psi^v\colon M_v\to M_v$ is the projection, see (28), and the term $\sigma_{\mathrm{f}}\times(\psi^v_{nn}(\Delta)-\Delta)$ is the (signed) measure having density $\sigma_{\mathrm{f}}(q)$ with respect to the measure $(\psi^v_{nn}(\Delta)-\Delta)(dq)$)

$$+\sum_v\sum_{v'\ \mathrm{n.n.}\ v}\beta_{v'v}F'_\Delta((T_{v'v}\Delta)-\Delta) \tag{34}$$

(here the operator $T_{v'v}$ acts on the measure Δ in the following way: it replaces the component Δ_v of the measure Δ by the measure $\Delta_{v'}$ (via identification between M_v and $M_{v'}$)).

We now check that the limiting operator Ω is the same one that we were dealing with in our study of the nonlinear Markov process, (10)–(15).

Proposition 2. *The operator* (30)–(34) *can be written in the form*

$$(\Omega(F))(\mu)=(\widehat{\sigma}(F))(\mu)+(F'(\mu))(g(\mu)), \tag{35}$$

where $\widehat{\sigma}$ is the generator of the time-shift semigroup acting on our manifold, F' is the Fréchet differential of F, and the (*signed*) *measure $g(\mu)$ is given by the right-hand side of* (10)–(15).

Proof. For the convenience of the reader we repeat here equation (10)–(15):

$$\begin{aligned}\frac{d}{dt}\mu_v(q_v,t)&=-\frac{d}{dr_{i^*(q_v)}(q_v)}\mu_v(q_v,t)\\&\quad+\delta(0,\tau(e(q_v)))\mu_v(q_v\ominus e(q_v))[\sigma_{\mathrm{tr}}(q_v\ominus e(q_v),q_v)+\sigma_{\mathrm{e}}(q_v\ominus e(q_v),q_v)]\\&\quad-\mu_v(q_v,t)\sum_{q'_v}[\sigma_{\mathrm{tr}}(q_v,q'_v)+\sigma_{\mathrm{e}}(q_v,q'_v)]\\&\quad+\Bigg[\int\limits_{q'_v\colon q'_v\ominus C(q'_v)=q_v}d\mu_v(q'_v)\sigma_{\mathrm{f}}(q'_v,q'_v\ominus C(q'_v))]-\mu_v(q_v)\sigma_{\mathrm{f}}(q_v,q_v\ominus C(q_v))\\&\quad+\sum_{v'\ \mathrm{n.n.}\ v}\beta_{vv'}[\mu_{v'}(q_v)-\mu_v(q_v)]\Bigg].\end{aligned}$$

The term $(\widehat{\sigma}(F))(\Delta)$ obviously corresponds to $-\dfrac{d}{dr_{i^*(q_v)}(q_v)}\mu_v(q_v,t)$, and the term

$$\sum_v\sum_{v'\ \mathrm{n.n.}\ v}\beta_{v'v}F'_\Delta((T_{v'v}\Delta)-\Delta),$$

to $\sum\limits_{v'\,\mathrm{n.n.}\,v} \beta_{vv'}[\mu_{v'}(q_v) - \mu_v(q_v)]$. The "external customer arrival" term

$$\sum_v \sum_{v'} \sum_{\varkappa} \lambda(\varkappa, v, v') F'_\Delta(\chi_{v,v';\varkappa}(\Delta) - \Delta)$$

matches the terms

$$\delta(0, \tau(e(q_v)))\mu_v(q_v \ominus e(q_v))\sigma_{\mathrm{e}}(q_v \ominus e(q_v), q_v) - \mu_v(q_v, t) \sum_{q'_v} \sigma_{\mathrm{e}}(q_v, q'_v).$$

The "intermediate service completion" term

$$\sum_v \sum_{v'\,\mathrm{n.n.}\,v} F'_\Delta(e_{v,D(C(\cdot))}(v')\sigma_{\mathrm{f}} \times (\zeta_{vv'}(\Delta) - \Delta))$$

matches the terms

$$\delta(0, \tau(e(q_v)))\mu_v(q_v \ominus e(q_v))\sigma_{\mathrm{tr}}(q_v \ominus e(q_v), q_v) - \mu_v(q_v, t) \sum_{q'_v} \sigma_{\mathrm{tr}}(q_v, q'_v).$$

Finally, the "final service completion" term $\sum\limits_v F'_\Delta(\sigma_{\mathrm{f}} \times (\psi^v_{nn}(\Delta) - \Delta))$ matches

$$\left[\int\limits_{q'_v:\, q'_v \ominus C(q'_v) = q_v} d\mu_v(q'_v)\sigma_{\mathrm{f}}(q'_v, q'_v \ominus C(q'_v))\right] - \mu_v(q_v)\sigma_{\mathrm{f}}(q_v, q_v \ominus C(q_v)). \quad \triangle$$

Let us check that we indeed have the norm-convergence of the operators Ω_N to Ω, the one needed in the convergence statement (18). The norm that we use here is again $\|\cdot\|_1$.

The precise statement that we need is the following.

Proposition 3. *Let F be a function on $\mathcal{P}$, with $\|F\|_1$ finite. We can restrict F on each subspace $\mathcal{P}_N$ and then apply the operator Ω_N, thus getting a function $\Omega_N F$ on $\mathcal{P}_N$. We can also restrict the function ΩF from $\mathcal{P}$ to $\mathcal{P}_N$. Then*

$$\|\Omega_N F - \Omega F\|_1^{\mathcal{P}_N} \le C_N \|F\|_1,$$

with $C_N \to 0$, where $\|\cdot\|_1^{\mathcal{P}_N}$ is the restriction of the norm $\|\cdot\|_1$ to the subspace of functions of the measures $\mathcal{P}_N$.

Proof. We have to compare the operators given by (22)–(26) and (30)–(34) term by term. For example, compare the term (24)

$$\sum_v \sum_k \sum_{v'} \sum_{\varkappa} \lambda(\varkappa, v, v')\left[F\Big(\Delta_N - \frac{1}{N}\delta_{(q_{v,k})} + \frac{1}{N}\delta_{(q_{v,k} \oplus c^v(\varkappa,0,v'))}\Big) - F(\Delta_N)\right],$$

which corresponds to the arrival of a new customer of class $\varkappa$ with a destination v', and the term (31)

$$\sum_v \sum_{v'} \sum_{\varkappa} \lambda(\varkappa, v, v') F'_{\Delta_N}(\chi_{v,v';\varkappa}(\Delta_N) - \Delta_N),$$

where $\chi_{v,v';\varkappa}\colon M_v \to M_v$ is the embedding corresponding to the arrival to v of an external customer of class $\varkappa$ and destination v'. Due to the locality properties of F, it is sufficient to establish the convergence

$$\sum_{k=1}^N \left[F\Big(\Delta_N - \frac{1}{N}\delta_{(q_{v,k})} + \frac{1}{N}\delta_{(q_{v,k} \oplus c^v(\varkappa,0,v'))}\Big) - F(\Delta_N)\right] \to F'_{\Delta_N}(\chi_{v,v';\varkappa}(\Delta_N) - \Delta_N) \tag{36}$$

as $N \to \infty$.

The measure Δ_N is a collection of N atoms, corresponding to queues $q_{v,k}$, $k = 1, \ldots, N$. In the first expression, we change just one of the N atoms, adding a new customer $c^v(\varkappa, 0, v')$ to each of the queues $q_{v,k}$, and then take the sum of the corresponding increments over k. In the second expression, we change all atoms simultaneously, obtaining the measure measure $\chi_{v,v';\varkappa}(\Delta_N)$, and instead of taking the increment $F(\chi_{v,v';\varkappa}(\Delta_N))-F(\Delta_N)$ we take the differential F'_{Δ_N} of the measure $\chi_{v,v';\varkappa}(\Delta_N)-\Delta_N$. To see the norm convergence in (36), let us rewrite the increments

$$F\Big(\Delta_N-\frac{1}{N}\delta_{(q_{v,k})}+\frac{1}{N}\delta_{(q_{v,k}\oplus c^v(\varkappa,0,v'))}\Big)-F(\Delta_N)=F'_{Q_{k,N}}\Big(-\frac{1}{N}\delta_{(q_{v,k})}+\frac{1}{N}\delta_{(q_{v,k}\oplus c^v(\varkappa,0,v'))}\Big),$$

by the intermediate value theorem. Here the points $Q_{k,N}$ are some points on the segments

$$\Big[\Delta_N,\Delta_N-\frac{1}{N}\delta_{(q_{v,k})}+\frac{1}{N}\delta_{(q_{v,k}\oplus c^v(\varkappa,0,v'))}\Big].$$

Note that the norms $\|Q_{k,N}-\Delta_N\|$ obviously go to zero as $N\to\infty$, and so $\|F'_{Q_{k,N}}-F'_{\Delta_N}\|\to 0$ as well. Thus,

$$\begin{aligned}\sum_{k=1}^{N}F'_{Q_{k,N}}&\Big(-\frac{1}{N}\delta_{(q_{v,k})}+\frac{1}{N}\delta_{(q_{v,k}\oplus c^v(\varkappa,0,v'))}\Big)\\ &=F'_{\Delta_N}\Bigg[\sum_{k=1}^{N}\Big(-\frac{1}{N}\delta_{(q_{v,k})}+\frac{1}{N}\delta_{(q_{v,k}\oplus c^v(\varkappa,0,v'))}\Big)\Bigg]\\ &\qquad+\sum_{k=1}^{N}(F'_{Q_{k,N}}-F'_{\Delta_N})\Big(-\frac{1}{N}\delta_{(q_{v,k})}+\frac{1}{N}\delta_{(q_{v,k}\oplus c^v(\varkappa,0,v'))}\Big),\end{aligned}$$

with the second term is uniformly small as $N\to\infty$. By definition,

$$\chi_{v,v';\varkappa}(\Delta_N)-\Delta_N=\sum_{k=1}^{N}\Big(-\frac{1}{N}\delta_{(q_{v,k})}+\frac{1}{N}\delta_{(q_{v,k}\oplus c^v(\varkappa,0,v'))}\Big),$$

and this proves the convergence needed. The other terms are compared in the same manner. △

This completes checking relation (18) of the convergence theorem (Theorem 3).

5.2. Fréchet Differential Properties

Here we check the core property (17).

Proposition 4. *The semigroup is uniformly differentiable in t. In the notation of Section 5 (see (21)) this means that for the Fréchet differential $h(t)=[\mathcal{D}\mu^1(t)](h)$ of the trajectory $\mu^1(t)$ at the point $\mu^1(t)$ in the direction h we have*

$$\big\|\mu^2(t)-\mu^1(t)-[\mathcal{D}\mu^1(t)](h)\big\|_1\le O(\|h\|_1^2)$$

uniformly in $t\le T$ and μ^1, provided that T is small enough.

Proof. To write the equation for the Fréchet differential $h(t)=[\mathcal{D}\mu^1(t)](h)$ of trajectory $\mu^1(t)$ at the point $\mu=\mu^1(t)$ in the direction h, we have to compare the evolving measures $\mu^1(t)$ and $\mu^2(t)$, which are solutions of equation (10)–(15) with initial conditions μ and $\mu+h$, and keep the terms linear in h. In what follows we use the notation $\sigma_{\mathrm{tr}}^{\mu}$, where the superscript refers to the state in which the rate σ_{tr} is computed; see (9).

For $h=\{h_v,v\in V\}$ we have

$$\frac{d}{dt}h_v(q_v,t)=-\frac{d}{dr_{i^*(q_v)}(q_v)}h_v(q_v,t)\tag{37}$$

(derivative along the direction $r(q_v)$)

$$+ \delta(0, \tau(e(q_v)))h_v(q_v \ominus e(q_v))[\sigma^{\mu}_{\mathrm{tr}}(q_v \ominus e(q_v), q_v) + \sigma_{\mathrm{e}}(q_v \ominus e(q_v), q_v)]$$
$$+ \delta(0, \tau(e(q_v)))\mu_v(q_v \ominus e(q_v))[\sigma^{h}_{\mathrm{tr}}(q_v \ominus e(q_v), q_v)], \tag{38}$$

(q_v is created from $q_v \ominus e(q_v)$ by the arrival of $e(q_v)$ from v', and $\delta(0, \tau(e(q_v)))$ accounts for the fact that if the last customer $e(q_v)$ was already served for some time, than it cannot arrive from the outside; see (9) and (5))

$$- h_v(q_v, t) \sum_{q'_v} [\sigma^{\mu}_{\mathrm{tr}}(q_v, q'_v) + \sigma_{\mathrm{e}}(q_v, q'_v)] - \mu_v(q_v, t) \sum_{q'_v} [\sigma^{h}_{\mathrm{tr}}(q_v, q'_v)] \tag{39}$$

(the queue q_v is changing due to customers arriving from the outside and from other servers)

$$+ \left[\int\limits_{q'_v \colon q'_v \ominus C(q'_v) = q_v} dh_v(q'_v)\sigma_{\mathrm{f}}(q'_v, q'_v \ominus C(q'_v)) \right] - h_v(q_v)\sigma_{\mathrm{f}}(q_v, q_v \ominus C(q_v)) \tag{40}$$

(here the first term describes the creation of the queue q_v after a customer was served in a queue q'_v (longer by one customer) such that $q'_v \ominus C(q'_v) = q_v$, while the second term describes the completion of service of a customer in q_v)

$$+ \sum_{v' \text{ n.n. } v} \beta_{vv'}[h_v(q_v) - h_v(q_v)] \tag{41}$$

(the β's are the rates of exchange of the servers).

The existence of the solution to the (linear) equation (37)–(41) follows by the Peano theorem, while the uniqueness of the solution is implied by the estimate (see (4))

$$\|h(t)\| \le \|h(0)\| e^{Ct},$$

which follows from the Grönwall's estimate.

Finally, we want to estimate the remainder

$$\rho(t) = [\mu + h](t) - \mu(t) - [\mathcal{D}\mu(t)](h).$$

Here $\mu + h \equiv \mu(0) + h(0) \equiv [\mu + h](0)$ is a small perturbation of μ, $[\mu + h](t)$ is its evolution, and $[\mathcal{D}\mu(t)](h)$ is the application of the Fréchet differential of the map $\nu(0) \rightsquigarrow \nu(t)$ computed at the point μ and applied to the increment h. Note that $\rho(0) = 0$ and it satisfies the equation

$$\begin{aligned}
\frac{d}{dt}\rho_v(q_v, t) = &- \frac{d}{dr_{i^*(q_v)}(q_v)}\rho_v(q_v, t) \\
&+ \delta(0, \tau(e(q_v)))\rho_v(q_v \ominus e(q_v))\sigma_{\mathrm{e}}(q_v \ominus e(q_v), q_v) \\
&+ \delta(0, \tau(e(q_v)))\rho_v(q_v \ominus e(q_v))\sigma^{[\mu+h]}_{\mathrm{tr}}(q_v \ominus e(q_v), q_v) \\
&+ \delta(0, \tau(e(q_v)))[\mu + h]_v(q_v \ominus e(q_v))\sigma^{\rho}_{\mathrm{tr}}(q_v \ominus e(q_v), q_v) \\
&+ \delta(0, \tau(e(q_v)))h_v(q_v \ominus e(q_v))\sigma^{h}_{\mathrm{tr}}(q_v \ominus e(q_v), q_v) - \\
&- \rho_v(q_v, t) \sum_{q'_v} [\sigma^{[\mu+h]}_{\mathrm{tr}}(q_v, q'_v) + \sigma_{\mathrm{e}}(q_v, q'_v)] \\
&- h_v(q_v, t) \sum_{q'_v} \sigma^{h}_{\mathrm{tr}}(q_v, q'_v) - [\mu + h]_v(q_v, t) \sum_{q'_v} \sigma^{\rho}_{\mathrm{tr}}(q_v, q'_v) \\
&+ \left[\int\limits_{q'_v \colon q'_v \ominus C(q'_v) = q_v} d\rho_v(q'_v)\sigma_{\mathrm{f}}(q'_v, q'_v \ominus C(q'_v)) \right] \\
&- \rho_v(q_v)\sigma_{\mathrm{f}}(q_v, q_v \ominus C(q_v)) + \sum_{v' \text{ n.n. } v} \beta_{vv'}[\rho_{v'}(q_v) - \rho_v(q_v)].
\end{aligned}$$

Note that the initial condition for the last equation is $[\rho]_v(q,t) = 0$. The terms on the right-hand side which do not contain ρ are

$$\delta(0, \tau(e(q_v)))h_v(q_v \ominus e(q_v))\sigma^h_{\mathrm{tr}}(q_v \ominus e(q_v), q_v) - h_v(q_v, t)\sum_{q'_v} \sigma^h_{\mathrm{tr}}(q_v, q'_v),$$

which is of the order of $\|h\|^2$. Therefore, by Grönwall's inequality the same bound holds uniformly for the function $[\rho]_v(q,t)$, provided that $t \le T$ with T small enough. △

Proposition 5. *The set of uniformly differentiable quasilocal functions is a core of the generator of our semigroup.*

Proof. Follows from Proposition 4, via the chain rule, and the Stone–Weierstrass theorem. △

This implies that condition (17) of Theorem 3 holds as well, so in our case it is indeed applicable.

6. CONCLUSION

In this paper we have established the convergence of the mean-field version of a spatially extended network with jumping servers to a nonlinear Markov process. The configuration of the N-component mean-field network is described by the (atomic) measure $\mu_N(t)$, which randomly evolves in time. We have shown that in the limit $N \to \infty$ we have convergence of the measures $\mu_N(t) \to \mu(t)$, where the evolution $\mu(t)$ is nonrandom. In a sense, this result can be viewed as a functional law of large numbers.

Our results can easily be generalized to the situation where instead of the underlying (infinite) graph G we take a sequence of finite graphs H_n such that $H_n \to G$, consider the N-fold mean-field type networks $H_{n,N}$, and take the limit as $n, N \to \infty$.

REFERENCES

1. Baccelli, F., Rybko, A.N., Shlosman, S.B., and Vladimirov, A., Stability, Metastability and Instability of Moving Networks. I, II, 2015 (in preparation).
2. Biskup, M. and Chayes, L., Rigorous Analysis of Discontinuous Phase Transitions via Mean-Field Bounds, *Commun. Math. Phys.*, 2003, vol. 238, no. 1–2, pp. 53–93.
3. Vvedenskaya, N.D., Dobrushin, R.L., and Karpelevich, F.I., Queueing System with Selection of the Shortest of Two Queues: An Asymptotic Approach, *Probl. Peredachi Inf.*, 1996, vol. 32, no. 1, pp. 20–34 [*Probl. Inf. Trans.* (Engl. Transl.), 1996, vol. 32, no. 1, pp. 15–27].
4. Bramson, M., Stability of Join the Shortest Queue Networks, *Ann., Appl., Probab.*, 2011, vol. 21, no. 4, pp. 1568–1625.
5. Rybko, A.N. and Shlosman, S.B., Poisson Hypothesis for Information Networks. I, II, *Mosc. Math. J.*, 2005, vol. 5, no. 3, pp. 679–704; no. 4, pp. 927–959.
6. Rybko, A., Shlosman, S., and Vladimirov, A., Spontaneous Resonances and the Coherent States of the Queuing Networks, *J. Stat. Phys.*, 2008, vol. 134, no. 1, pp. 67–104.
7. McKean, H.P., Jr., A Class of Markov Processes Associated with Nonlinear Parabolic Equations, *Proc. Nat. Acad. Sci. U.S.A.*, 1966, vol. 56, no. 6, pp. 1907–1911.
8. Pirogov, S., Rybko, A.N., and Shlosman, S.B., Propagation of Chaos for Some Queuing Networks, 2014 (in preparation).
9. Liggett, T.M., *Interacting Particle Systems*, New York: Springer, 1985.
10. Ethier, S.N. and Kurtz, T.G., *Markov Processes: Characterization and Convergence*, New York: Wiley, 1986.
11. Karpelevich, F.I. and Rybko, A.N., Asymptotic Behavior of the Thermodynamical Limit for Symmetric Closed Queueing Networks, *Probl. Peredachi Inf.*, 2000, vol. 36, no. 2, pp. 69–95 [*Probl. Inf. Trans.* (Engl. Transl.), 2000, vol. 36, no. 2, pp. 154–179].