HAL
open science

Poisson Hypothesis for Open Networks at Low Load

Alexandre Rybko, Senya Shlosman, Alexandre Vladimirov

To cite this version:

Alexandre Rybko, Senya Shlosman, Alexandre Vladimirov. Poisson Hypothesis for Open Networks at Low Load. Moscow Mathematical Journal, 2017, 17 (1), pp.145-160. hal-01265620

HAL Id: hal-01265620

https://hal.science/hal-01265620v1

Submitted on 11 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:1408.7034v3 [math.PR] 9 Jun 2015

Poisson Hypothesis for Open Networks at Low Load

A. Rybko[††], Senya Shlosman[†,††] and A. Vladimirov[††]
[†]Aix Marseille Université, Université de Toulon,
CNRS, CPT UMR 7332, 13288, Marseille, France
senya.shlosman@univ-amu.fr
[††]Inst. of the Information Transmission Problems,
RAS, Moscow, Russia
rybko@iitp.ru, shlos@iitp.ru, vladim@iitp.ru

January 9, 2018

Abstract

We study large communication networks of the mean-field type. The input flows to the nodes of the network are supposed to be stationary and with low rate. We show that such a network is ergodic, i.e. it goes to the stationary state, which does not depend on the initial state of the network. This is in contrast with the high load regime, when the large time behavior of the network might depend on its initial state. Our technique is based on the coupling construction, which couples two Non-Linear Markov Processes.

Part of this work has been carried out in the framework of the Labex Archimede (ANR-11-LABX-0033) and of 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). Part of this work has been carried out at IITP RAS. The support of Russian Foundation for Sciences (project No. 14-50-00150) is gratefully acknowledged.

1 Introduction and results

In this paper we study large information networks with low load. Our goal is to show that they obey a certain pattern, which is known under the name of Poisson Hypothesis.

The Poisson Hypothesis is a device to predict the behavior of large queuing networks. It was formulated first by L. Kleinrock in [K], and it was assumed to hold universally. It concerns the following situation. Suppose we have a large network of servers, through which many customers are traveling, being served at different nodes of the network. If the node is busy, the customers wait in the queue. Customers are entering into the network from the outside via some nodes, and these external flows of customers are Poissonian, with constant rates. The service time at each node is random, depending on the node, and the customer. The PH prediction about the (long-time, large-size) behavior of the network is the following:

- consider the total flow $\mathcal{F}$ of customers to a given node $\mathcal{N}$. Then $\mathcal{F}$ is approximately equal to a Poisson flow, $\mathcal{P}$, with a time dependent rate function $\lambda_{\mathcal{N}}(T)$.
- The exit flow from $\mathcal{N}$ – **not** Poissonian in general! – has a rate function $\gamma_{\mathcal{N}}(T)$, which is smoother than $\lambda_{\mathcal{N}}(T)$ (due to averaging, taking place at the node $\mathcal{N}$).
- As a result, the flows $\lambda_{\mathcal{N}}(T)$ at various nodes $\mathcal{N}$ should go to a constant limits $\bar{\lambda}_{\mathcal{N}} \approx \frac{1}{T}\int_0^T \lambda(t)\,dt$, as $T \to \infty$, the flows to different nodes being almost independent.
- The above convergence is uniform in the size of the network.

Note that the distributions of the service times at the nodes of the network can be arbitrary, so PH deals with quite a general situation. The range of validity of PH is supposed to be the class of networks where the internal flow to every node $\mathcal{N}$ is a union of flows from many other nodes, and each of these flows constitute only a small fraction of the total flow to $\mathcal{N}$. If true, PH provides one with means to make easy computations of quantities of importance in network design.

The rationale behind this conjectured behavior is natural: since the inflow is a sum of many small inputs, it is approximately Poissonian. And due

to the randomness of the service time the outflow from each node should be "smoother" than the total inflow to it. (This statement was proven in [RSV1] under quite general conditions.) In particular, the variation of the latter should be smaller than that of the former, and so all the flows should go with time to corresponding constant values.

In the paper [RS1] the Poisson Hypothesis is proven for simple networks in the infinite volume limit, under some natural conditions. The counterexamples to such universal behavior are also known by now. Some were constructed in [RS2], [RSV2].

It was argued in [RSV2], that the violation of the Poisson Hypothesis can happen only at high load, and that the load plays here the same role as the temperature in the statistical mechanics, so that the high (low) load corresponds to low (high) temperature. We have shown in [RSV3] that indeed the Poisson Hypothesis holds for the closed networks with low load. In the present paper we show that the Poisson Hypothesis behavior holds also for the open networks with low load. (The time gap between the paper [RSV3] and the present one is due to the fact that certain general results needed were not available earlier. Now they are established, see [BRS].)

The main technical innovation of our paper is that we prove our result by constructing a suitable coupling of two Non-Linear Markov Processes (in Sections 6.2, 6.3 below). The coupling constructed shows that the two NLMP merge together as time diverges, provided the load of the network is low. It is thus reminiscent of the FKG-uniqueness results of statistical mechanics, when it is obtained via coupling constructions for systems with attraction, [H].

In the next Section we describe the basic finite network, from which large networks of the mean-field type can be built. This mean-field construction is explained in the Section 3, and the corresponding Non-Linear Markov Process (NLMP) is defined in Section 4. The general properties of the networks with low input flows are established in the Section 5. The Section 6 contains the main result of the paper – that of ergodicity of the NLMP, corresponding to the low inflow. It is established by using the coupling of Sections 6.2, 6.3.

2 Basic network

The main object of our study are large queueing networks that are composed of many identical copies of a finite basic network G. In this section we will

describe this basic network.

2.1 G : Markov chain of queues

The network G consists of a finite number of servers, s, also called nodes. At any given time t, each node is populated by a finite number of customers (also called particles) of different classes. The set J of all classes is finite, $|J| < \infty$. It is convenient from the very beginning to include into the class j of the customer the address of server $s(j) \in G$ where the customer is currently located. The customers wait in the queue for their turn to be served by the node. The order of service is governed by the service discipline. After being served, the customer either jumps to some node of G (becoming there the last in the queue), or leaves the network.

The traffic of customers in G is governed by the routing matrix P. This is a $|J| \times |J|$ matrix, whose rows and columns correspond to the classes of customers in the network. The element p_{ij} of P is the probability of a customer c of class i to change its class from i to j after the service. As soon as the service of c is over, the customer jumps with probability p_{ij} to the node $s(j)$. With the complementary probability $p_{i0} = 1 - \sum_{j \in J} p_{ij}$ the customer leaves the network.

2.2 Queues

We restrict consideration to Markov networks where the service time of each customer is exponentially distributed. We also consider service disciplines that depend only on the order of arrival of customers to the queue (and not on the interarrival times or any other history). Namely, for each node n and each queue $x = (j_1, ..., j_k)$ of length k at this node, a collection $\gamma_n(x, j_r)$: $r = 1, \ldots, k$ of service rates is defined by the service discipline. The formal definition will be given later.

For the alphabet $J = \{1, \ldots, |J|\}$ we define the configuration space

$$X = J^* = \bigcup_{k=0,1,\ldots} J^k$$

of finite words in this alphabet, with $J^0 = \{\varnothing\}$ and $J^1 = J$. We will use the notation $x = (x_1, \ldots, x_k) \equiv (j_1, \ldots, j_k)$ for the list of all customers in the network. Denote $|x| = k$.

The order of customers in the queue coincides with the order of their arrival, that is, x_1 is the class of the oldest customer in the queue and x_k is that of the last arrived. (The customer x_1 is not necessarily the first to be served, since we consider a rather general class of service disciplines.)

2.3 Network events

The current state $x \in X$ can be changed as a result of arrival events and service events. An *arrival event* of type j takes x to

$$x \oplus j = \{j_1, \ldots, j_{|x|}, j\}.$$

Service event may only happen if $x \neq \emptyset$. If $x = \{j_1, \ldots, j_{k-1}, j_k, j_{k+1}, \ldots, j_{|x|}\}$, i.e. $k \leq |x|$, then the service event operation $\ominus *_k$ is the jump from x to

$$x \ominus *_k = \{j_1, \ldots, j_{k-1}, j_{k+1}, \ldots, j_{|x|}\}.$$

With some abuse of notation we also denote this jump as $x \ominus j_k$. Service and arrival events may happen simultaneously, if a particle whose service is finished, joins the same queue again, see below.

2.4 Markov dynamics

The random evolution of the network G is a continuous-time Markov process $\mathcal{G}$ on the countable configuration space X. Let us write down the rates of all arrival and service events as functions of the current configuration $x \in X$.

The external arrivals of clients of type $j \in J$ is a Poisson process with rate $\lambda_j(t)$. We will consider the time homogeneous case, when $\lambda_j(t) \equiv \lambda_j$, though for technical reasons we will also use sometime the general case of time-dependent rates $\lambda_j(t)$. The external arrival events of customers of different types $j \in J$ are independent of each other and of the state of the system.

A service event of type $i \in x$ happens at appropriate location in x at rate $\gamma(x, i)$. If the customer i stays in the network and turns into customer j, then the configuration of the system becomes $x \ominus i \oplus j$. That can happen with probability p_{ij}. Another option is that the customer leaves the network, and the network state becomes $x \ominus i$. That happens with probability $p_{i0} = 1 - \sum_{j \in J} p_{ij}$.

Apart from the pairs of associated service and arrival events mentioned above, all the events in the system happen independently of each other. The

non-zero transition rates $r(x,y)$ of the process G are thus equal to

$$r(x, x \oplus j) = \lambda_j(t), \qquad j \in J, \tag{1}$$

$$r(x, x \ominus i) = p_{i0}\gamma(x,i), \qquad i \in x, \tag{2}$$

$$r(x, x \ominus i \oplus j) = p_{ij}\gamma(x,i), \qquad i \in x,\ j \in J. \tag{3}$$

The state of the process $\mathcal{G}$ at the time t is given by a probability measure μ_t on X. We denote by

$$\mathcal{P}(X) = \{\mu \in \mathbb{R}^X : \mu(x) \geq 0, \quad x \in X, \quad \sum_{x \in X} \mu(x) = 1\}$$

the space of probability measures on X.

3 Mean-field networks

In order to construct the mean-field networks G_M, we take the colection of M independent copies of G, and interconnect them as follows. The service event i at any given copy of G is followed by the corresponding arrival event j at any other copy of G with probability $\frac{1}{M}p_{ij}$. With the complementary probability, $p_{i0} = 1 - \sum_{j \in J} p_{j_i j}$, no arrival event happens, and the customer leaves the system.

3.1 $\mathcal{G}_M$: the mean-field type process

The resulting Markov process $\mathcal{G}_M$ on the configuration product space

$$X^M = \{\mathbf{x} = (x_1, \ldots, x_M) : x_m \in X,\ m = 1, \ldots, M\}$$

can be described formally as follows. Let us introduce the notations:

$$\mathbf{x} \oplus_m j = (x_1, \ldots, x_m \oplus j, \ldots, x_M)$$

and for $i \in x_m$

$$\mathbf{x} \ominus_m i = (x_1, \ldots, x_m \ominus i, \ldots, x_M).$$

Then the transition rates of the process $\mathcal{G}_M$ are given by the relations

$$r(\mathbf{x}, \mathbf{x} \oplus_m j) = \lambda_j(t), \tag{4}$$

$$r(\mathbf{x}, \mathbf{x} \ominus_m i) = (1 - \sum_{j \in J} p_{ij})\gamma(x_m, i), \tag{5}$$

$$r(\mathbf{x}, \mathbf{x} \ominus_m i \oplus_{m'} j) = \frac{p_{ij}\gamma(x_m, i)}{M}, \tag{6}$$

for all $\quad m, m' = 1, \ldots, M$, $i \in x_m$, $j \in J$.

3.2 Factor-process

The permutation group S_M acts on X^M, and the process $\mathcal{G}_M$ is invariant under this action. The factor space can be identified with the atomic measures on X. So we will use the notation

$$\mathbf{x} = \frac{1}{M} \sum_{m=1}^{M} \delta_{x_m}, \tag{7}$$

where δ_x is the unit measure at $x \in X$.

Denote by $\mathcal{P}_M(X) \subseteq \mathcal{P}(X)$ the subspace of all measures of the form (7). The factor-process, denoted by the same symbol, $\mathcal{G}_M$, becomes a continuous-time Markov process on the (countable) configuration space $\mathcal{P}_M(X)$. Its state at time t is a probability measure

$$\mu_t^M = \{\mu_t^M(\mathbf{x}) : \mathbf{x} \in \mathcal{P}_M(X)\}$$

on $\mathcal{P}_M(X)$ and thus on $\mathcal{P}(X)$. The expectation $\mathbb{E}\mu_t^M$ is an element of $\mathcal{P}(X)$ but, in general, not an element of $\mathcal{P}_M(X)$.

4 Non-linear Markov process

Let us consider the limit of $\mathcal{G}_M$ as $M \to \infty$. To this end, let us choose one of M copies of G in the system G_M – say, the first one, G_1 – and observe for $t \geq 0$ the arrival and service events only at this copy. In other words, we take a projection $\mathcal{H}_M$ of $\mathcal{G}_M$ to G_1, integrating out the remaining degrees of freedom.

If we want to know the evolution of the process $\mathcal{H}_M$, then, in addition to the initial configuration $x_0 \in X$ we need, of course, to specify the initial state $\mu_{t=0}^M \in \mathcal{P}_M(X)$ of the whole ensemble.

Suppose that the initial measures μ_0^M converge weakly to some $\mu_0 \in \mathcal{P}(X)$ as $M \to \infty$, and consider the limiting process $\mathcal{H} = \lim_{M\to\infty} \mathcal{H}_M$. The existence and uniqueness of the limiting process is proven in [**?**]. This stochastic process will be referred to as the *non-linear Markov process* (NLMP) $\mathcal{H}$. To specify its evolution we need to know the initial configuration $x \in X$ and

the initial measure $\mu_0 \in \mathcal{P}(X)$. The NLMP $\mathcal{H}$ is not a Markov process on X because of its dependence on the whole measure μ_0 and not just on $x(0)$.

The current configuration of this extended process is the pair $(\mu_t, x(t))$, and this configuration defines the distribution of the future configurations $(\mu_s, x(s))$, $s \geq t$, uniquely. Moreover, the measure μ_s is defined by μ_t in a deterministic way, for $s \geq t$. The measure μ_t is a solution of a system of differential equations which we write down below. Once the family $\{\mu_t, t \geq 0\}$ is given, the process $x(t)$ becomes an inhomogeneous Markov process on X, see below for more details.

4.1 The process equations

The evolution equations for the state μ_t of the process $\mathcal{H}$ are conveniently written in terms of *flow rates* which we define now.

The *outflow rate* of class j customers from the system $\mathcal{H}$ at time t is given by

$$u_j(t) = \sum_{x \in X} \mu_t(x) \sum_{r: x_r = j} \gamma(x, x_r), \qquad j \in J \tag{8}$$

The *inflow rate* of class j customers to the system is then equal to

$$v_j(t) = \lambda_j(t) + w_j(t), \qquad j \in J, \tag{9}$$

where

$$w_j(t) = \sum_{i \in J} p_{ij} u_i(t), \qquad j \in J, \tag{10}$$

is the *internal inflow rate*.

Thus the total inflow is a time-inhomogeneous (vector) Poisson process, whose rate vector is $v(t) = \lambda(t) + w(t)$. The queueing process $x(t)$ is then an inhomogeneous Markov process corresponding to the time-dependent arrival process with rates $v_j(t)$, $j \in J$, and the service rates $\gamma(x, i)$, $i \in x$.

An arrival event of type j converts the configuration from x to $x \oplus j$. The rate of this arrival equals to $\mu(x) v_j$. Analogously, a service event of type i converts x to $x \ominus i$. Its rate is $\mu(x) \gamma(x, i)$.

We have described the dynamics of $\mathcal{H}$ in terms of arrival and service events by specifying their rates and the corresponding changes of the measure μ. The time evolution of $\mu_t \in \mathcal{P}(X)$ is given by the following countable system

of ordinary differential equations:

$$\dot{\mu}_t = \sum_{x \in X} \mu_t(x) \left[\sum_{r=1}^{|x|} \gamma(x, x_r)(\delta_{x \ominus x_r} - \delta_x) + \sum_{j \in J} v_j(t)(\delta_{x \oplus j} - \delta_x) \right]. \qquad (11)$$

Here we treat the measure μ_t as a vector with coordinates $\mu_t(x)$, and δ_y-s denotes the δ-measures at various points $y \in X$. The first sum in brackets corresponds to service events that happen at configurations $x \in X$, $|x| > 0$, with (configuration-dependent) rates $\gamma(x, j)$, and the second sum corresponds to arrival events that happen at all configurations $x \in X$ with (configuration-independent) rates $v_j(t)$, $j \in J$.

Relations (8)–(11) describe completely the evolution of μ_t. The resulting dynamical system is non-linear since its right-hand side contains terms $\mu_t(x)v_j(t)$, which are quadratic in μ, since $v_j(t)$ depends on μ_t linearly. The existence and uniqueness of the solution μ_t on $[0, +\infty)$ of the system (11) for any initial state μ_0 is proven in [BRS] under minimal assumptions on the network. In particular, the existence and uniqueness hold under the following *rate conditions*: for some $\lambda_+ < \infty$ and $0 < \gamma_- \leq \gamma_+ < \infty$ we have

$$\lambda_j(t) \leq \lambda_+ \quad \text{for all} \quad j \in J,\ t \geq 0, \qquad (12)$$

$$\gamma_- \leq \gamma(x) \leq \gamma_+ \quad \text{for all} \quad x \in X,\ x \neq \emptyset, \qquad (13)$$

where $\gamma(x) = \sum_{r=1}^{|x|} \gamma(x, x_r)$.

Note that under the rate conditions the total inflow rates $v_j(t)$ in the system have a uniform upper bound $V < \infty$. So for each $x \in J$ the function $\mu_t(x)$ is Lipschitz continuous in t, with a Lipschitz constant which does not depend on x, t, or μ_0.

5 Processes with low inflows

5.1 Bounds for a single queue

Now let us consider a server that receives independent Poisson inflows of particles of classes $j \in J$ with rates $\lambda_j(t)$. As in the NLMP model, there is no feedback, that is, served customers never return to our server.

In what follows we assume that $\lambda_j(t) \leq \varepsilon$ for all $t \geq 0$, where $\varepsilon > 0$ is a small parameter. Let us derive some bounds on the length $|x|$ of the queue

at this server. Recall that our service discipline is rather general, but the total service rate $\gamma(x)$ is neither too low nor too high for any $x \neq 0$, see rate conditions (13).

We will derive bounds on the distribution ν_t of queue length, provided the queue is empty at $t = 0$. Denote by Q the random process of queue lengths. The trajectories of Q are step functions, taking values in $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$.

Denote by $\nu_t(k)$ the probability that there are exactly k particles in the queue at time t. We have

$$\sum_{k=0,1,\dots} \nu_t(k) = 1, \quad t \geq 0.$$

Let us compare the process Q with the stationary process Q', defined as follows:

- All the customers in Q' have the same exponentially distributed service time and the FIFO service discipline is used;
- The service rate in the process Q' is equal to the lower bound γ_- on the service rates for the process Q;
- The inflow to Q' is a stationary Poisson flow with rate ε for each type $j \in J$.

For ε small enough such a process Q' exists and is unique. One easily sees that the queues in Q' dominates stochastically those in Q at any time $t \geq 0$. Indeed, by coupling we may organize the pairs of evolutions $q(t)$ and $q'(t)$ in such a way that, apart from synchronous arrivals to both queues, there are additional arrivals to q', and, apart from synchronous service attempts, there are additional service attempts at q. Then $|q'(t)| \geq |q(t)|$ for all $t \geq 0$ since $|q(0)| = 0$.

For the stationary process Q' the explicit form of the queue length distribution ν' is known; it is the geometric distribution:

$$\nu'(k) = (1 - \varkappa)\varkappa^k, \qquad \text{where} \qquad \varkappa = \frac{|J|\varepsilon}{\gamma_-}.$$

We have

$$\nu'(0) = 1 - \varkappa, \quad \sum_{k \geq 1} k\nu'(k) = \frac{\varkappa}{(1 - \varkappa)}, \quad \sum_{k > 1} k\nu'(k) = \frac{\varkappa^2}{(1 - \varkappa)}.$$

Hence, from the domination argument we have

Lemma 1 *Let the initial state $\nu_{t=0}$ has no customers. For $\varkappa = \frac{|J|\varepsilon}{\gamma_-}$,*

$$\nu_t(0) \geq 1 - \varkappa, \quad t \geq 0, \tag{14}$$

so the fraction of non-empty queues is $O(\varepsilon)$;

$$\sum_{k\geq 1} k\nu_t(k) \leq \frac{\varkappa}{(1-\varkappa)},$$

so the expected length of the queue is $O(\varepsilon)$;

$$\sum_{k>1} k\nu_t(k) \leq \frac{\varkappa^2}{(1-\varkappa)}, \tag{15}$$

so the expected length of non-single queues is $O(\varepsilon^2)$.

Let $h(j)$ be the expected remaining number of services for a particle of class j. Clearly, $h(j)$ depends only on the rooting matrix P and is finite since $\rho(P) < 1$ (the network is open). Let us put

$$h = \max_{j\in J} h(j).$$

For a queue $x \in X$ we denote

$$L(x) = \sum_{i=1}^{|x|} h(x_i),$$

where $x = (j_1, \ldots, j_{|x|})$. Finally,

$$L(\mu) = \int_X L(x) d\mu(x).$$

As a consequence of Lemma 1, we get the following upper bound on $L(\nu_t)$:

Lemma 2 *Under the hypothesis of Lemma 1, there exists a constant $C_1 > 0$, which depends only on the routing matrix P, such that*

$$L(\nu_t) \leq h\alpha(\nu_t) + C_1\varepsilon^2, \tag{16}$$

where $\alpha(\nu_t) = 1 - \nu_t(\varnothing)$.

Proof. Note that $L(x) \leq h|x|$. We estimate $L(\nu_t)$ separately over queues of length one and over longer queues. The first contribution is $\leq h\alpha(\nu_t)$, since $\nu_t(1) \leq \alpha(\nu_t)$, and for the second one we use (15) . ■

5.2 Bounds for the NLMP

Now we return to the NLMP $\mathcal{H}$. Our goal is to prove that if $\varepsilon > 0$ is small, then any solution μ_t converges weakly to the unique stationary solution μ^*. We will do it, first, for the class of *finite weight* initial measures μ_0, that is, for μ_0 with a finite expected length of the queue:

$$\sum_{x \in X} |x| \mu_0(x) < \infty.$$

Note that the state μ_t is finite weight for all $t > 0$ iff μ_0 is finite weight.

First, we will show that there exists a time moment $T(\mu_0)$, such that the expected value of non-empty queues

$$\alpha(\mu_t) = 1 - \mu_t(\varnothing)$$

in our system drops down below the value $C\varepsilon$ for some constant $C > 0$. Second, we prove that for some $\bar{C} > 0$ it remains below the value $\bar{C}\varepsilon$ at all later times. That implies that all the flow rates in the system will stay below $D\varepsilon$ for some $D > 0$.

To make the first step we will construct a Lyapunov function. To implement the second step we 'split' our process $\mathcal{H}$ after the time moment $T(\mu_0)$ in two parts, such that the 'smaller' one has all the non-empty servers at the time $T(\mu_0)$, and the 'larger' part starts from the empty state at $T(\mu_0)$.

We begin by explaining the splitting construction. The splitting of the process $\mathcal{H} = \{\mu_t, t \geq 0\}$ is a pair of processes μ_t^1 and μ_t^2 on X such that

$$\rho\mu_t^1 + (1-\rho)\mu_t^2 = \mu_t, \quad t \geq 0. \tag{17}$$

In words, we mark some part of queues at $t = 0$ and then consider the joint evolution of marked and unmarked queues within the NLMP $\mathcal{H}$. The processes μ_t^1 and μ_t^2 are also NLMP-s, but with non-homogeneous external flows.

Formally, let $\mu_0 = \rho\mu_0^1 + (1-\rho)\mu_0^2$, where μ_0^i, $i = 1, 2$, are probability measures on X and $0 \leq \rho \leq 1$. Let us describe the dynamics of μ_t^i, $i = 1, 2$, $t \geq 0$ by means of internal flows. We define, in the same manner as above,

$$u_j^k(t) = \sum_{x \in X} \mu_t^k(x) \sum_{r: x_r = j} \gamma(x, x_r), \qquad j \in J, \qquad k = 1, 2, \tag{18}$$

and

$$w_j^k(t) = \sum_{i \in J} p_{ij} u_i^k(t), \qquad j \in J, \qquad k = 1, 2. \tag{19}$$

The total inflow will be defined in a different way:

$$v_j'(t) = \lambda_j(t) + \rho w_j^1(t) + (1-\rho) w_j^2(t), \qquad j \in J. \tag{20}$$

Finally, the joint dynamics of μ_t^k, $k = 1, 2$, satisfies the following differential equations:

$$\dot{\mu}_t^k = \sum_{x \in X} \mu_t^k(x) \left[\sum_{r=1}^{|x|} \gamma(x, x_r)(\delta_{x \ominus x_r} - \delta_x) + \sum_{j \in J} v_j'(t)(\delta_{x \oplus j} - \delta_x) \right], \tag{21}$$

for $k = 1, 2$.

Summing up equations (21) over $k = 1, 2$ with factors ρ and $(1-\rho)$, respectively, we conclude that (17) holds. It is easy to see that the evolution of each measure μ_t^k satisfies the equations of an NLMP with a modified routing matrices: $p_{ij}^1 = \rho p_{ij}$, $p_{ij}^2 = (1-\rho)\, p_{ij}$ (fractions of outflows go to the other half of the system) and an additional time-dependent Poisson inflow (the customers that come from the other half of the system).

Now we are prepared to prove the main statement of this section.

Lemma 3 *There exists a constant $\bar{C} > 0$ and, for each initial finite-weight measure μ_0 on X, a time $T = T(\mu_0)$ such that*

$$\alpha(\mu_t) < \bar{C}\varepsilon \quad \textit{for all} \quad t > T.$$

Proof. We use the mean number of remaining service events $L(\mu)$ as the Lyapunov function. The function $L(\mu)$ is finite for any finite-weight measure μ on X, since $L(x) \le h|x|$ for any $x \in X$. Let us check that

$$\frac{d}{dt} L(\mu_t) = \sum_{j \in J} \lambda_j(t) h(j) - S(\mu_t), \tag{22}$$

where

$$S(\mu) = \sum_{x \in X : x \neq \varnothing} \gamma(x) \mu\,(x)$$

is the mean service rate for a queue with the distribution μ (for empty queues this rate is zero). Indeed, the arrival event of type j adds the value $h(j)$

to $L(\mu)$. On the other hand, the expected number of services of a given customer decreases by 1 after its service is over, which explains the second term. Clearly, $S(\mu_t) \geq \gamma_- \alpha(\mu_t)$, where γ_- is the minimal service rate (for nonempty queues). So from (22) we see that if $\gamma_- \alpha(\mu_t) > |J|\varepsilon h$, then the function $L(\mu_t)$ decays. Hence at a certain moment $t = T$ we will have $\alpha(\mu_T) \leq \frac{|J|\varepsilon h}{\gamma_-}$.

It remains to prove that for all $t \geq T$ the value of $\alpha(\mu_t)$ will never exceed $\bar{C}\varepsilon$ for some constant $\bar{C}$ which does not depend on μ_0 and T. To do this we use the splitting construction expalined above, with $\rho = \alpha(\mu_T)$, the initial states $\mu_T^1(\cdot) = \mu_T(\cdot | x \neq \varnothing)$, $\mu_T^2(\varnothing) = 1$. By definition, the first portion – $\rho\mu_t^1(\cdot)$ – of our process μ_t remains small for all times $t > T$, since ρ is small. The total mass of the remaining measure $(1-\rho)\mu_t^2$ is close to one, but the state μ_T^2 is empty, and the rate of external flow is of the order of ε. Therefore one can expect that the state μ_t^2 is sparcely populated at all times $t > T$. This is indeed the case:

Claim. *Let $\lambda(t) = \sum_{j \in J} \lambda_j(t)$ and for all $t \geq 0$ we have $\lambda(t) \leq \varepsilon$. Let the initial state $\mu_0 (\equiv \mu_T^2)$ has no customers, i.e. $\mu_0(\varnothing) = 1$. Then there exists a constant $C > 0$ such that for all $t \geq 0$*

$$\alpha(\mu_t) \leq C\varepsilon.$$

Proof of the Claim. Note first that for some $C_3 > 0$ we have

$$\dot{L}(t) \leq C_3\varepsilon - \gamma_- \alpha(\mu_t), \quad t \geq 0, \tag{23}$$

as it follows from (22). Note also, that if $\dot{L}(\mu_t) \geq 0$, then $\alpha(\mu_t) \leq \frac{C_3\varepsilon}{\gamma_-} \equiv \alpha_+$. We will show that for all $t \geq 0$ we have $\alpha(\mu_t) < \bar{\alpha} \equiv (1+h)\alpha_+$, provided ε is small.

Suppose the opposite, and let $\bar{T} > 0$ be the first occasion when $\alpha(\mu_t)$ reaches the level $\bar{\alpha}$. Denote by T_+ the last occasion before $\bar{T}$ when $\alpha(\mu_{T_+}) = \alpha_+$. Note that the total inflow rate $|u(t)|$, $t \leq \bar{T}$ satisfies the inequality

$$|u(t)| \leq \varepsilon + \gamma_+ \bar{\alpha} = C_4\varepsilon.$$

Therefore, from (16) we have the inequality

$$L(\mu_{T_+}) \leq h\alpha_+ + C_5\varepsilon^2.$$

Since $\dot{L}(\mu_t) \leq 0$ for all $t \in \left[T_+, \bar{T}\right]$, we have

$$L(\mu_{\bar{T}}) \leq L(\mu_{T_+}) \leq h\alpha_+ + C_5\varepsilon^2. \tag{24}$$

On the other hand, we always have that $L(\mu) \geq \alpha(\mu)$, so in particular

$$L(\mu_{\bar{T}}) \geq \alpha(\mu_{\bar{T}}) = h\alpha_+ + \alpha_+$$

which contradicts (24) when $\varepsilon > 0$ is small enough, and the claim follows.

The proof of Lemma 3 is thus also finished. ■

5.3 Stationary states of the NLMP

Let us assume that the external inflow is stationary, that is, $\lambda(t) \equiv \lambda = (\lambda_j)$, $t \geq 0$. If the corresponding stationary state μ^λ of $\mathcal{H}$ exists, then the constant internal inflow rate w^* is found from the equation

$$P\left(\lambda + w^*\right) = w^*, \tag{25}$$

that is,

$$w^* = \left(I - P\right)^{-1} P\lambda. \tag{26}$$

Let us assume that $\lambda_j \leq \varepsilon$, $j \in J$, where $\varepsilon > 0$ is a small parameter.

Lemma 4 *If ε is small enough, then $\mathcal{H}$ has a unique stationary state $\mu(t) \equiv \mu^*$.*

The uniqueness follows from (25) and (26). For the existence, it suffices to prove that the network without feedback is ergodic under the stationary Poisson inflow of rate $\lambda + w^*$. The components of this rate vector are of order ε, hence, the assertion follows from the Foster criterion, see, for instance, [B], or directly from Lemma 1.

6 Ergodicity of the NLMP with low inflow

6.1 The main Theorem

Now we turn to the main result of the paper.

Theorem 5 *For a given G, there exists an $\varepsilon > 0$ such that for all $\lambda = (\lambda_j)$ with $\lambda_j < \varepsilon$, $j \in J$ the NLMP dynamics μ_t with finite-weight initial measure μ_0 converges to the unique equilibrium μ^λ: for every $x \in X$*

$$\lim_{t\to\infty} |\mu_t(x) - \mu^\lambda(x)| = 0.$$

We begin the proof with a technical lemma.

Lemma 6 *There exist $C, \varepsilon_0 > 0$ such that, for any $\varepsilon < \varepsilon_0$ and any initial finite-weight measure μ_0, there exists a time $T = T(C, \varepsilon, \mu_0)$ such that*

$$\mathbb{E}_{\mu_t} |x| < C\varepsilon \qquad \textit{for all} \quad t > T.$$

Proof. This result follows from Lemma 2 and Lemma 3. ■

6.2 The coupling of Non-Linear Markov Processes

We are going to construct a coupling $\mathcal{H}' \otimes \mathcal{H}''$ of two NLMP $\mathcal{H}'$ and $\mathcal{H}''$, which live on the same graph, have the same inflows, and differ only by their initial states, μ_0' and μ_0''. We remind the reader that the coupling of the processes μ_t' and μ_t'' is a process M_t on $X \times X$ with marginals μ_t' and μ_t''. In our case the process M_t will be again a NLMP. Its initial state can be any coupling of μ_0' and μ_0''; for example the product $\mu_0' \times \mu_0''$ will go.

The coupling we construct has a form of a sum: $M_t = R_t + W_t$; the two (positive) processes R_t and W_t are interacting NLMP-s, except that the masses $r_t \equiv R_t (X \times X), w_t \equiv W_t (X \times X)$ of each of them can be less than one, while of course $r_t + w_t = 1$. The processes R_t and W_t are processes on pairs of queues, and we will call them the *red* and *white* pairs; the customers in these queues will be likewise called *red* and *white* customers. The choice of the initial states M_0^r and M_0^w of the processes R_t and W_t are not very important; one option can be $R_0 = M_0, W_0 = 0$. The key property for us will be that the process W_t has its support on the diagonal $X \subset X \times X$, so we will be done if we will show that $w_t \to 1$ as $t \to \infty$.

We start with the verbal **Definition** of the evolution of the (interacting) processes R_t and W_t. We then write down the differential equation, equivalent to that definition.

As usual, we couple the (equal) inflows to $\mathcal{H}'$ and $\mathcal{H}''$ in such a way that the customers arrive in pairs (of identical customers) simulataneously to $\mathcal{H}'$ and $\mathcal{H}''$. The arriving external pairs of customers have white colour. The fraction w_t of arriving identical pairs goes into the process W_t. The remaining amount, r_t, goes into red process, R_t, and the corresponding white arriving customers become red instantly.

The service of white customers also happens in pairs, so the two paired (identical) customers start and end their service sinchronously. After the

service is over, the newly created pair either leaves the network, or else go to the *same* node, according to the routing matrix, i.e. the routing of the two customers is identical. The rule of staying white pair of customers or becoming red customers are then the same as for the arriving external white pairs.

Each red customer c is served individually. If c was participating in the process $\mathcal{H}'$ (resp., $\mathcal{H}''$), then after the service he either leaves the network or jumps to the another node, still being in the process $\mathcal{H}'$ (resp., $\mathcal{H}''$). If c was belonging to the red pair of queues $(q' \oplus c, q'')$, and if $q' \neq \varnothing$ or $q'' \neq \varnothing$, then the resulting pair (q', q'') stays red, while in the case when both $q' = \varnothing$ and $q'' = \varnothing$, the red pair $(c, \varnothing)$ turns into the pair $(\varnothing, \varnothing)$, which is declared white. If c stays in the network and enters into the pair (q', q'') of queues, then the resulting pair $(q' \oplus c, q'')$ is declared red, independently of the colour of the pair (q', q'').

Summarizing, we see that a white particle turns red in two cases: if it joins a red queue, or if a red particle arrives to a white pair of queues (then all the white customers in both queues turn red). The queues in the white pair are identical, while the queues in the red pair may be identical or different. The pair $(\varnothing, \varnothing)$ is always white.

Our goal will be to prove that the fraction of red pairs r_t vanishes as $t \to \infty$. It will imply that the limits of μ'_t and μ''_t, as $t \to \infty$, exist and coinside with μ^λ. In the end we will drop the assumption of μ'_0 to be finite-weight.

6.3 Equations for the coupled NLMP

As was explained above, the state of our coupled process is given by the two measures – white and red:

$$\{W_t(x), R_t(y,z): \quad x, y, z \in X\},$$

where we think of x as a pair $(x, x) \in X \times X$ on the diagonal. Sometime we will omit the subscript t.

Given such a state, let us define the corresponding – white and red – flows. The outflows are

$$u_j(W) = \sum_{x \in X} W(x) \sum_{i: x_i = j} \gamma(x, i),$$

$$u'_j(R) = \sum_{y,z} R(y,z) \sum_{i:y_i=j} \gamma(y,i),$$
$$u''_j(R) = \sum_{y,z} R(y,z) \sum_{i:z_i=j} \gamma(z,i).$$

The corresponding internal inflows are

$$v_j(W) = \sum_k p_{kj} u_k(W),$$
$$v'_j(R) = \sum_k p_{k,j} u'_j(R),$$
$$v''_j(R) = \sum_k p_{k,j} u''_j(R).$$

Let us define $v_j\,(W,R) = v_j(W) + v'_j(R) + v''_j(R)$.

Now we write the differential equations for the white and red measures W and R, which describe formally the evolution explained in the previous section.

$$\frac{d}{dt}W_t(x) = -W_t(x)\left[\sum_j (\lambda_j + v_j(W_t,R_t)) + \sum_{i=1}^{|x|}\gamma(x,i)\right]$$
$$+\ W_t(\bar{x} : \bar{x}\oplus j = x)(\lambda_j + v_j(W_t)) + \sum_{i=1}^{|x|+1}\sum_{\tilde{x}:\tilde{x}\ominus\tilde{x}_i=x} W_t(\tilde{x})\gamma(\tilde{x},\tilde{x}_i)$$
$$+\ \mathbb{I}_{\{x=\varnothing\}}\left(\sum_j \left[R_t(j,\varnothing) + R_t(\varnothing,j)\right]\gamma(j)\right);$$

$$\frac{d}{dt}R_t(y,z) = -R_t(y,z)\left[\sum_j (\lambda_j + v_j(W_t,R_t)) + \sum_{i=1}^{|y|}\gamma(y,i) + \sum_{i=1}^{|z|}\gamma(z,i)\right]$$
$$+R_t(\bar{y} : \bar{y}\oplus j = y, \bar{z} : \bar{z}\oplus j = z)(\lambda_j + v_j(W_t))$$
$$+R_t(\bar{y} : \bar{y}\oplus j = y, z)v'_j(R)\ +\ R_t(y,\bar{z} : \bar{z}\oplus j = z)v''_j(R)$$
$$+\ \mathbb{I}_{\{(y,z)=(j,\varnothing)\}}W_t(\varnothing)v'_j(R)\ +\mathbb{I}_{\{(y,z)=(\varnothing,j)\}}W_t(\varnothing)v''_j(R)$$
$$+\ \mathbb{I}_{\{(y,z)\neq(\varnothing,\varnothing)\}}\left[\sum_{i=1}^{|y|+1}\sum_{\bar{y}:\bar{y}\ominus\bar{y}_i=y} R(\bar{y},z)\gamma(\bar{y},\bar{y}_i) + \sum_{i=1}^{|z|+1}\sum_{\bar{z}:\bar{z}\ominus\bar{z}_i=z} R(y,\bar{z})\gamma(\bar{z},\bar{z}_i)\right].$$

6.4 Proof of stability

Now the colored dynamics is explicitely defined; it is a non-linear Markov process on the space of colored pairs of queues. Any pair is either white (and then the queues are identical and synchronized) or red (and then the queues evolve independently). The theorem 5 follows from the following

Theorem 7 *If ε is small enough, then $r_t \to 0$ as $t \to \infty$.*

Proof. First, we start the two NLMP $\mathcal{H}'$ and $\mathcal{H}''$ and run them for the time T, given by Lemma 3. After that time we couple them as was explained above. The rates of all the flows in the coupled system do not exceed $C_1\varepsilon$ at any time $t \geq T$. Denote by $h(t)$ the expected number of remaining services *of red particles only* in both $\mathcal{H}'$ and $\mathcal{H}''$. By assumption, $h(0)$ is finite and so $h(t)$ remains finite for all $t \geq 0$. The change of r_t and $h(t)$ in time is due to two factors. The first one is the termination of the service of red particles; it contributes to decline of $h(t)$, which happens with the rate Cr_t. The second one is the creation of new red particles, and it makes $h(t)$ to increase, either when

1. white particles arrive to red queues, or when
2. red particles arrive to white queues.

In the case 1 the rate of emergence of the *new* red particles (and hence the rate of growth of $h(t)$) is at most of the order of εr_t, since the flow of red particles has order r_t, the fraction of *non-empty* white queues has order at most ε and the mean length of these queues is of order 1. In the case 2, the rate of growth of $h(t)$ is also of order εr_t since the flow of white particles has order ε and the fraction of red queues has order r_t. Summarizing, we arrive at the following differential inequality:

$$\frac{d}{dt}h(t) \leq -(C - 2\varepsilon)\, r_t.$$

Since $h(t)$ is positive, the integral $\int_0^\infty r_t dt$ is finite. Together with the fact that $r_t \geq 0$ and the derivative of r_t is bounded, it implies that

$$\lim_{t\to\infty} r_t = 0.$$

■

7 Infinite expected initial length

Here we say few words on extension of Theorem 5 to the case of infinite expected initial queue length. We give a sketch of the proof.

In order to reduce the situation to that of a finite initial length, we separate a small fraction $\beta > 0$ of queues so that the remaining ones are of finite expected length at $t = 0$. The value of β can be chosen arbitrarily small.

The flows to and from the separated part are of order β. We need to prove that, for finite initial weights and for different inflows that differ by order β, the solutions approach each other up to a distance of order β as well.

First, we prove that the fraction of red queues must eventually drop down to the order β. This follows from the differential equations for the dynamics of $L(t)$ and r_t. So we can suppose that $r_T \sim \beta$ at some time T.

At this time T we separate, additionally, all the red queues (their fraction is r_T) and consider the remaining process that starts with no red queues. Then we can use the argument similar to that in the proof of Lemma 6 and show that r_t remains of order β for all $t \geq T$. The convergence to the equilibrium follows since β can be chosen arbitrarily small.

References

[BRS] F. Baccelli, A. Rybko, and S. Shlosman. Queuing networks with varying topology – a mean-field approach, 2013. Eprint arXiv:1311.3898, to appear in Queuing Systems.

[B] Bramson, M. (2008). Stability of Queueing Networks, Springer.

[H] Holley, R. (1974), "Remarks on the FKG inequalities", Communications in Mathematical Physics 36 (3): 227–231.

[K] Kleinrock, L.: Communication Nets, Stochastic Message Flow and Delay, McGraw-Hill Book Company, New York, 1964.

[RS1] A.N. Rybko, S.B. Shlosman: *Poisson Hypothesis for Information Networks.* arXiv:math/0406110 [math.PR], Sinai's Festschrift, Moscow Math. J., v.5, 679-704, 2005, Tsfasman's Festschrift, Moscow Math. J., v.5, 927-959, 2005.

[RS2] Rybko, A. N.; Shlosman, S.B.: *Phase transitions in the queuing networks and the violation of the Poisson hypothesis.* Mosc. Math. J. 8 (2008), no. 1, 159–180.

[RSV1] Rybko, A. N.; Shlosman, S.B. and Vladimirov A. *Self-averaging property of queuing systems,* arXiv:math/0510046 [math.PR], Problems of Information Transmission, 42 , Issue 4 (December 2006) Pages: 344 - 355.

[RSV2] Rybko, A. N.; Shlosman, S.B. and Vladimirov A.: *Spontaneous Resonances and the Coherent States of the Queuing Networks,* J. Stat Phys (2008) 134: 67–104.

[RSV3] A. Rybko, S. Shlosman and A. Vladimirov: *Absence of Breakdown of the Poisson Hypothesis. I. Closed Networks at Low Load,* arXiv:0811.3577 [math-ph], Markov Processes and Related Field, 16, pp.267-285, 2010.