J Stat Phys (2010) 140: 819–845
DOI 10.1007/s10955-010-0018-5

Non-Equilibrium Statistical Physics of Currents in Queuing Networks

Vladimir Y. Chernyak · Michael Chertkov · David A. Goldberg · Konstantin Turitsyn

Received: 29 January 2010 / Accepted: 25 June 2010 / Published online: 16 July 2010

Abstract We consider a stable open queuing network as a steady non-equilibrium system of interacting particles. The network is completely specified by its underlying graphical structure, type of interaction at each node, and the Markovian transition rates between nodes. For such systems, we ask the question "What is the most likely way for large currents to accumulate over time in a network ?", where time is large compared to the system correlation time scale. We identify two interesting regimes. In the first regime, in which the accumulation of currents over time exceeds the expected value by a small to moderate amount (moderate large deviation), we find that the large-deviation distribution of currents is universal (independent of the interaction details), and there is no long-time and averaged over time accumulation of particles (condensation) at any nodes. In the second regime, in which the accumulation of currents over time exceeds the expected value by a large amount (severe large deviation), we find that the large-deviation current distribution is sensitive to interaction details, and there is a long-time accumulation of particles (condensation) at some nodes. The transition between the two regimes can be described as a dynamical second order phase

V.Y. Chernyak · M. Chertkov (✉) · D.A. Goldberg · K. Turitsyn
Center for Nonlinear Studies and Theoretical Division, LANL, Los Alamos, NM 87545, USA
e-mail: chertkov@lanl.gov

V.Y. Chernyak
Department of Chemistry, Wayne State University, 5101 Cass Ave, Detroit, MI 48202, USA
e-mail: chernyak@chem.wayne.edu

M. Chertkov
New Mexico Consortium, Los Alamos, NM 87544, USA

D.A. Goldberg
Operations Research Center, MIT, Cambridge, MA 02139, USA
e-mail: dag3141@mit.edu

K. Turitsyn
Landau Institute for Theoretical Physics, Moscow 119334, Russia
e-mail: turitsyn@lanl.gov

transition. We illustrate these ideas using the simple, yet non-trivial, example of a single node with feedback.

Keywords Statistics of non-equilibrium currents · Open queueing networks · Condensation phenomenon · Birth-death processes

1 Introduction

1.1 Non-Equilibrium Statistical Physics and Queueing Networks

The concept of statistical equilibrium is extremely powerful. Once detailed balance (which is synonymic to the equilibrium) is established, one can shortcut a discussion of dynamics and just consider the Gibbs distribution that governs simultaneous correlations in the steady state. On the other hand, if detailed balance is broken no free lunch is guaranteed, and one generally does need to dive into dynamics, even to describe just the steady state. This is an infamous and principal difficulty at the very core of non-equilibrium statistical physics. It is thus of interest to identify a class of non-equilibrium steady systems where the steady state, distinctly different from the Gibbs distribution, can be derived in a straightforward way.

The problem discussed in this manuscript belongs to this class—it is a non-equilibrium statistical physics problem with the steady state known explicitly. More precisely, we study the model generally known as an open queueing network. A general (Markovian) open queueing network, the so-called Jackson network [1], can be described as a random walk of particles (jobs, vehicles, people, computer packets, etc) on a directed graph. The directed links label the transitions between stations/nodes, each of the Poissonian type and thus characterized by a single number (the rate). Each node is characterized by the number of equivalent servers. Collectively this defines a many-particle problem, whose non-equilibrium nature (no detailed balance) follows immediately from the definitions. Adopting the terminology of the Queueing Theory community (a subset of the Operations Research community), the type of service at a node is provided by the $M/M/m/\infty$ queue, which is translated as Markovian input, Markovian output with m servers, and infinite waiting room. (Throughout the manuscript we use the shorter notation $M/M/m$.) An example of such a network is shown in Fig. 1, and more information will be provided below. Note that a similar, and to a degree more general, model is known in statistical physics as the zero range model [2].

In spite of the generally pessimistic non-equilibrium assessment, the stable (i.e. achieving a statistical steady state) open Jackson network allows an explicit and simple solution for the steady state [3]. Similar statements apply to the zero range model [2]. The solution for the steady state is the so-called product form [2–4], where the joint distribution function of the occupation numbers at all the nodes/stations is factorized into a product of marginal probabilities at the separate nodes. Some new and physics-based exposition of this factorized solution is due to [5], and will also be a part of our construction below. We note that a different factorization has also led to the solution of the related Asymmetric Exclusion Processes (AEP) Models, see [6–9] and the references therein.[1]

[1]The general AEP models are significantly different from the Jackson-network models discussed in this manuscript—they can be viewed as a special case of an $M/M/1/1$ queuing model (waiting room with one slot and no particles lost) in contrast with the Jackon network of $M/M/m/\infty$ (infinite waiting room) queues. In the case of a simple one-dimensional chain the AEP model is reducible (with a proper redefinition of the phase space) to the Jackson network model (and other way around) [6–9], and then both models show product-state form solution [2–4]. However, the product-state solvability, which holds for the Jackson network generally, does not extend to the AEP model on general graphs.

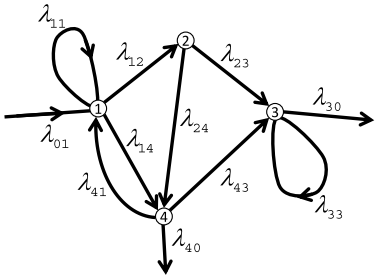

Fig. 1 Example of an open finite queueing network represented by a directed graph. The sample graph consists of four vortexes/stations, labeled 1, 2, 3, 4, with label 0 is reserved for an external (out) node. Transitions between the stations are shown as directed edges. Loops (self-loops), as $1 \to 1$, are allowed. Each graph edge is equipped with a transition rate. Throughout the manuscript all transitions in the network are assumed memoryless, i.e. 0-degree Markovian–Poisson. We also focus on the case of the infinite waiting room, i.e., no particles/jobs are lost, and thus the number of particles accumulated at a node may reach $+\infty$. Another important characteristic of a station/note is the number of tellers/servers. For the sake of simplicity, we assume the jobs/particles to be identical (single-class), and all tellers at any given station processing with the same time-independent Poisson rate

On the other hand, in recent years the quest for universality in non-equilibrium statistical problems has turned to the analysis of currents generated over time, long compared to the system correlation time scale. This route became fruitful and helped to establish some fundamental relations about the symmetry of the currents distribution, known as fluctuation theorems [10–12]. In spite of this partial success, the full description of the currents distribution for a general many-particle and open non-equilibrium problem was (and is) deemed too difficult.[2]

1.2 Large Deviations, Queueing Networks, and Condensation Phenomenon

One of the central questions in the Queueing Theory community is the probability of rare events in queueing networks (large deviations) [15]. A central tool in understanding such events is the notion of the so-called 'fluid-limit' of a queueing network [16], which was originally developed to understand certain questions related to the stability (existence of steady state) for certain complicated networks [17, 18]. We now describe this fluid-limit in greater detail. For a *fixed* queueing network, one scales both *time* and the *initial number of particles in system* by some large integer n, and then normalizes the set of queue lengths by n. Note the important difference with the standard notion of hydrodynamic limit—here, the number of nodes in the network is *fixed*—only time and the initial number of particles are scaled. It is proven in [15, 16] that for a broad class of networks, this scaling has a non-trivial limit (the fluid limit). It is then standard in the Queueing Theory literature to perform the large deviations analysis in the setting of this fluid limit, as opposed to the original (unscaled)

[2]We would like to emphasize here the principal difficulty arising from the fact that the system is both (a) many-particle and (b) open. The analysis of such systems is often limited to the setting in which there are few degrees of freedom. An example is a recent paper of three of us with Puliafito [13] that discusses an explicit expression for the distribution of currents associated with a polymer stretched by external shear flow. In another paper (two of us with Malinin and Teodorescu [14]) considered a setting with multiple degrees of freedom and analyzed the statistics of so-called topological currents. The statistics of currents have also been studied for the aforementioned AEP models over one-dimensional chains, which typically have multiple degrees of freedom [8].

network. There has been great progress in understanding all aspects of these large deviations [19–23]. However, driven by applications to describing buffer overloads in communications systems, most of these results have been geared towards understanding how large queue lengths accumulate over time (see for example [22]). Much less is known about how large currents build up in a queueing network over time, especially for the unscaled (not fluid limit) network.

The statistical physics community has developed several tools (e.g. the fluctuation theorems [10–12]) that are well-suited to studying the large-deviations properties of currents in a network. In light of the aforementioned gap in our understanding of the large deviations of currents in queueing networks and related network models, several researchers in the statistical physics community have recently begun to apply the fluctuation theorems to the study of currents in a variety of network-related models [24–27]. Most of these results have been for models such as the zero-range process, which are tangential (although closely related) to the networks studied by the Queueing Theory community. We bridge this gap by directly applying such an analysis to the canonical model of queueing theory—namely the Jackson network. We note that although the potential application of these tools to Jackson networks is mentioned in [24], the paper closest in spirit to our own is [25]. Indeed, for the 1-D zero-range process the authors identify several regimes in which the model may operate, characterized by the large-deviations properties of currents, and whether or not there is an accumulation of particles at sites over time (so-called condensation phenomenon). Their analysis proceeds by formalizing the system dynamics in a quantum-mechanical/operator-theoretic framework, and then studying the relevant spectral properties and Crámer (large deviation) functions. Our own analysis will very-much parallel that of [25], but for the Jackson network model on general graphs. Other related work can be found in [28], in which the zero-range model on regular lattices is studied in the hydrodynamic limit, as well as [29] where the closely related stochastic lattice gas model is studied.

We also note that the Queueing Theory community has already had some success in applying some of these tools [30–32]. Indeed, these analyses (and our own) rely on the interpretation of queueing systems as interacting particle systems, which (historically) helped lead to many of the great breakthroughs in the understanding of queueing networks (e.g. product-form solution [33]). However, these analyses have been for either an infinite 1-D chain of queues [30, 32] or only in the hydrodynamic limit [31]. In either case, their findings parallel our own, in the existence of certain phase transitions in which condensation may (or may not) occur.

1.3 Currents in a Queueing Networks with Feedback

We now discuss the nature of the currents in a Jackson network in steady state, a topic that has generated much research in the queueing literature [34–38]. Although currents over the entering links are Poissonian by construction, those which leave the system are not obviously Poissonian. However, if the system is stable, these exiting flows are in fact Poissonian [33, 35]. More generally, the flow is Poissonian along all arcs that may not be revisited by a particle, and these flows enjoy several nice properties (such as asymptotic independence) [38]. It was recognized early on in the queueing literature that the statistics of the internal currents with feedback are rather difficult to analyze [35]. The source of these difficulties is the complicated feedback mechanism that arises when particles may revisit an arc. Thus, even for the simple single-node feedback system (which will serve as an enabling example in this manuscript), the statistics of the feedback current are rather complicated [39–41]. Although several properties of the currents in a multi-server feedback queue and its generalizations have been studied in the literature [42–47], most of these results involve showing

that in certain limiting regimes the flows are close to Poissonian under some metric [43, 48–50]—the precise nature of these flows remains poorly understood. Furthermore, it seems that an operator-theory/generating-function approach has not yet been applied to the study of these feedback currents. We will take this approach to derive new understanding of these currents.

1.4 Our Main Results

The main results of our manuscript can be described as follows:

- We present an explicit, detailed operator-theoretic description of a Jackson network in the Doi-Peliti formalism. This expands on the description given in [5], with an eye towards introducing a larger set of the statistical physics community to standard queueing models.
- We identify an "uncongested" regime in Jackson networks, w.r.t. the large-deviation behavior of current. This regime is universal, i.e. interaction-independent, and non-Poissonian. Furthermore, in this regime there is no infinite accumulation of particles at any node (condensation). This universality can be qualitatively explained as follows. In this regime, the given deviation is driven by sample paths in which the number of particles does not diverge. This will generally occur when the given deviations are somewhat mild, and a deviation can be attained without a massive buildup of particles. Then, the impact of one particle 'blocking' another does not contribute asymptotically to the deviation, and thus the system is equivalent asymptotically to one in which all nodes have an infinite number of servers (particles do not interact). We confirm the existence of this regime by demonstrating that the single-node network with feedback falls into this regime for certain parameters, which we compute explicitly.
- We also identify a second, "congested" regime, which is interaction-dependent. In this regime, a given deviation is driven by system primitives that are *dependent* on the number of servers and service times. This will generally occur when the given deviations are more severe, and the only way to attain the given deviation is to have all, or at least some, servers busy for essentially the entire time horizon. In this regime the time-averaged queue length diverges for at least one node, marking the dynamical phase transition between the two regimes as second-order.
- We observe that statistics of time-averaged (over the large observational interval) queue and queue measured at the last moment of time, both conditioned to an atypical (large or small) values of currents, are not identical. The difference is particularly striking in the "congested" regime of the largest currents, where all moments of the former object (queue at the last moment of time) saturate to finite values while all moments of the later object (time-averaged queue) diverge with time. Quite generally this phenomenon can be classified as a breakdown of ergodicity in cases which are atypical with respect to currents. We note that this phenomenon has been previously studied in the queueing theory community, in the context of large deviations and quasi-stationary distributions [51, 52].

1.5 Outline

The manuscript is organized as follows. In Sect. 2, we present a technical introduction to the dynamics of the Jackson network in terms of the physics-native Doi-Peliti ("quantum" or "second-quantized") technique [53–55]. There we formally characterize the Master Equation (ME) that governs the system evolution and steady state, as well as the joint distribution of densities (occupation numbers). We customize this description to the $M/M/\infty$, $M/M/1$,

and general $M/M/m$ models in Sects. 2.1, 2.2, and 2.3, respectively. In Sect. 3, which represents the core of the manuscript, we adopt the Doi-Peliti technique to analyze the ME for the joint distribution function of densities (that reside at the nodes) and currents (that reside at the links). We also show how the coherent-state technique provides a complete description of the ground state (eigen-value and eigen-function) in terms of the relevant evolution operator. Section 3 is partitioned into four Subsections. In Sect. 3.2, we discuss the universal (and statistically typical) "uncongested" regime. We also describe the boundary of the "uncongested" region in the space of currents, and comment on the associated dynamical phase transition. Our analysis procedes by invoking an auxiliary construction for the left eigenfunction of the evolution operator, which is discussed in Appendix. The transitions between the "uncongested" and (partially) "congested" regimes are discussed in Sect. 3.3. The general theory is illustrated in Sect. 3.4 for a single-node system with feedback. In Sect. 4, to validate the theory, we describe a full spectral solution for this single-feedback problem. In Sect. 5 we draw several conclusions and discuss future directions for research.

2 The Doi-Peliti-Massey Operator Technique for a Generic Birth-Death Process

This section introduces notation and describes the main operational rule of the Jackson network in terms of the statistical physics native Doi-Peliti technique [53–55]. We note that a very similar (but less explicit) formulation was derived in [5].

As we will see, in this context the Doi-Peliti technique is closely related to the operator-theoretic framework formulated by Massey [56, 57] in the Queueing Theory community. We start by introducing the quantum-mechanics based bra(c)ket notation. We then discuss the product-form solutions for the stationary problems associated with the $M/M/\infty$, $M/M/1$, and generic $M/M/m$ networks in Sects. 2.1, 2.2, and 2.3, respectively.

The network (e.g. the one shown in Fig. 1) is represented by the directed graph, $(\mathcal{G}_0, \mathcal{G}_1)$, where $\mathcal{G}_0, \mathcal{G}_1$ marks the set of vertices and directed edges of the graph (respectively). If at some instance t, node j has a queue of size n, one says that the node is in the state represented by the ket-vector $|n\rangle$, where $n = 0, 1, \ldots$. Then, any "pure" state of the network will be denoted by the ket-vector $|\boldsymbol{n}\rangle$, where the components n_i of a vector $\boldsymbol{n} = (n_i | i \in \mathcal{G}_0)$ are labeled by the network nodes (vertices). If a state $|\boldsymbol{n}\rangle$ is realized with the probability $P(\boldsymbol{n})$, we say that the entire network is in the following "mixed" state

$$|s\rangle = \sum_{\boldsymbol{n}} P(\boldsymbol{n}) |\boldsymbol{n}\rangle, \quad \sum_{\boldsymbol{n}} P(\boldsymbol{n}) = 1, \tag{1}$$

where the last condition reflects the fact that the total probability equals unity. Here and below we formally assume that $P(\boldsymbol{n}) = 0$ whenever any component of the vector $\boldsymbol{n}$ is negative.

It is convenient to introduce a Hilbert space of $\mathcal{G}_0$-dimensional analytic functions of the vector variable $\boldsymbol{z} = (z_i | i \in \mathcal{G}_0)$

$$\mathcal{P}(\boldsymbol{z}) = \sum_{\boldsymbol{n}} P(\boldsymbol{n}) \prod_{i \in \mathcal{G}_0} z_i^n, \tag{2}$$

which is also known as the generating function in the theory of birth-death processes [58, 59].

The "quantum" (pure) states are transformed by the following creation and annihilation operators:

$$\hat{a}_j^+ | \ldots, n_j, \ldots \rangle = | \ldots, n_j + 1, \ldots \rangle, \qquad \hat{a}_j | \ldots, n_j, \ldots \rangle = n_j | \ldots, n_j - 1, \ldots \rangle. \tag{3}$$

The normalization condition in (1), i.e. conservation of probability, reads

$$\langle \mathbf{0}| \exp\left(\sum_{j\in\mathcal{G}_0} \hat{a}_j \right) |s\rangle = 1, \tag{4}$$

where the vacuum state $|\mathbf{0}\rangle \equiv |0,\ldots,0\rangle$ corresponds to the empty queue over the entire network.

In these notations ME becomes

$$\partial_t |s\rangle = \hat{H}|s\rangle, \tag{5}$$

where $\hat{H}$ is the Hamiltonian operator of the Q-network. In an integrated form, (5) is equivalent to

$$|s(t)\rangle = \hat{U}(t)|s(0)\rangle, \quad \hat{U}(t) \equiv T\exp\left(\int_0^t dt' \hat{H} \right), \tag{6}$$

where T exp is defined as a time-ordered exponential, i.e. the product of time-discretized operators. Furthermore, it becomes normal exponential if the parameters of $\hat{H}$ (i.e. transition rates) do not carry explicit time dependence.

Note that the Hamiltonian $\hat{H}$ is always real, as it represents probabilities which are positive and bounded. Thus the normalization conditions (4) and (5), which should be enforced by the theory for any feasible $\boldsymbol{n}$, result in

$$\langle \mathbf{0}| \exp\left(\sum_j \hat{a}_j \right) \hat{H} = 0, \tag{7}$$

where we have used standard bra-vector notations. Recall that an operator acting on the bra-vector from the right generates a bra-vector, and all the features of left operations can be extracted directly from the normal definition of the bra-(c)-ket scalar product, $\forall n,m$: $\langle n|m\rangle = \delta(n,m)$. Stating it differently, $\langle \mathbf{0}| \exp(\sum_j \hat{a}_j)$ is the left eigen-vector of the Hamiltonian with zero eigen-value.

The expectation value of a (dummy) operator $\hat{\bullet}$ over a state $|s\rangle$ is

$$\langle \hat{\bullet} \rangle \equiv \langle \mathbf{0}| \exp\left(\sum_{j\in\mathcal{G}_0} \hat{a}_j \right) \hat{\bullet} |s\rangle, \tag{8}$$

and according to (7), the corresponding Heisenberg (evolution) equation becomes

$$\partial_t \langle \hat{\bullet} \rangle = \langle [\hat{\bullet}, \hat{H}] \rangle. \tag{9}$$

Here we have assumed that $\hat{\bullet}$ does not have an explicit time dependence, and $[\hat{A}, \hat{B}]$ is the standard notation for a commutator.

2.1 $M/M/\infty$ Network

The generic form of the ME for a network of $M/M/\infty$ queues is

$$\frac{\partial}{\partial t} P(\boldsymbol{n};t) = \sum_{(i,j)\in\mathcal{G}_1}^{i,j\neq 0} \lambda_{ij} \big((n_i+1) P(\ldots, n_i+1, \ldots, n_j-1, \ldots; t)$$

$$
\begin{aligned}
&- n_i P(\dots, n_i, \dots, n_j, \dots; t)\big) \\
&+ \sum_{i\in\mathcal{G}_0} \lambda_{0i} \left(P(\dots, n_i - 1, \dots; t) - P(\dots, n_i, \dots; t)\right) \\
&+ \sum_{i\in\mathcal{G}_0} \lambda_{i0} \left((n_i + 1)P(\dots, n_i + 1, \dots; t) - n_i P(\dots, n_i, \dots; t)\right). \qquad (10)
\end{aligned}
$$

Here (i, j) stands for the directed edge of the network corresponding to a job transfer from site i to site j, with Poisson rate λ_{ij}; and λ_{0j},λ_{j0} are the Poisson rates for job injection and removal to/from the network at site j (respectively). Applying summation over properly weighted $\boldsymbol{n}$-states to both sides of (10), and using the relations (5), (6), one arrives at the following Hamiltonian:

$$
\hat{H}_\infty = \sum_{(i,j)\in\mathcal{G}_1}^{i,j\neq 0} \lambda_{ij}(\hat{a}_j^+ - \hat{a}_i^+)\hat{a}_i + \sum_{i\in\mathcal{G}_0} \left(\lambda_{0i}(\hat{a}_i^+ - 1) + \lambda_{i0}(1 - \hat{a}_i^+)\hat{a}_i\right), \qquad (11)
$$

which was first derived for the problem in [54].

We further introduce a path-integral representation. The analytic structure of the theory is as follows:

$$
\mathcal{P}(z; t) = \int \frac{d\boldsymbol{\zeta} d\boldsymbol{\zeta}'}{(2\pi i)^{|\mathcal{G}_0|}} W(z, \boldsymbol{\zeta})\mathcal{P}(\boldsymbol{\zeta}'; 0) \exp(-\boldsymbol{\zeta}\boldsymbol{\zeta}'), \qquad (12)
$$

$$
W(z, \boldsymbol{\zeta}) = \int_{\boldsymbol{\eta}(0)=\boldsymbol{\zeta}}^{\boldsymbol{\eta}'(t)=z} \mathcal{D}\boldsymbol{\eta}\mathcal{D}\boldsymbol{\eta}' \exp\left(z\boldsymbol{\eta}(t) - \int_0^t dt' \left(\boldsymbol{\eta}'(t')\dot{\boldsymbol{\eta}}(t') - \mathcal{H}_\infty(\boldsymbol{\eta}'(t'), \boldsymbol{\eta}(t))\right)\right), \qquad (13)
$$

where $\mathcal{H}_\infty(\boldsymbol{\eta}', \boldsymbol{\eta})$ corresponds to $\hat{H}_\infty$ expressed as a polynomial over the creation/annihilation operators, so that the creation operators are all positioned on the left from the annihilation operators (normal ordering), and $\hat{a}_j^+$,$\hat{a}_j$ are replaced by η_j' and η_j (respectively). Thus, for a general birth-death model (11) over a network $\mathcal{G}$, one derives

$$
\mathcal{H}_\infty(\boldsymbol{\eta}', \boldsymbol{\eta}) = \sum_{(i,j)\in\mathcal{G}_1}^{i,j\neq 0} \lambda_{ij}(\eta_j' - \eta_i')\eta_i + \sum_{i\in\mathcal{G}_0} \left(\lambda_{0i}(\eta_i' - 1) + \lambda_{i0}(1 - \eta_i')\eta_i\right). \qquad (14)
$$

As usual, the path-integrals in (13) should be understood as the continuous limit of the following discretized multiple integral (see [54] for explanations and accurate validation of the proper discrete-time regularization):

$$
\begin{aligned}
W(z, \boldsymbol{\zeta}) = \lim_{N\to\infty} \int \prod_{l=1}^{N-1} \frac{d\boldsymbol{\eta}_l d\boldsymbol{\eta}_l'}{(2\pi i)^{|\mathcal{G}_0|}} \\
\times \exp\left(z\boldsymbol{\eta}_{N-1} + \Delta\mathcal{H}(z, \boldsymbol{\eta}_{N-1}) + \sum_{l=1}^{N-1} \left(-\boldsymbol{\eta}_l'(\boldsymbol{\eta}_l - \boldsymbol{\eta}_{l-1}) + \Delta\mathcal{H}(\boldsymbol{\eta}_l', \boldsymbol{\eta}_{l-1})\right)\right), \qquad (15)
\end{aligned}
$$

where $\Delta = t/N$.

Finally, one finds that the creation-annihilation (11) and the path-integral (13), (15) formulations of the birth-death process also allow for the following simple "differential" interpretation for the analytic function defined in (2):

$$
\partial_t \mathcal{P}(z; t) = \hat{\mathcal{H}}_\infty(z)\mathcal{P}(z; t), \qquad (16)
$$

$$\hat{\mathcal{H}}_\infty(\boldsymbol{z}) = \mathcal{H}_\infty(\boldsymbol{z}, \partial_{\boldsymbol{z}}) = \sum_{(i,j)\in\mathcal{G}_1}^{i,j\neq 0} \lambda_{ij}(z_j - z_i)\partial_{z_i} + \sum_{i\in\mathcal{G}_0} \left(\lambda_{0i}(z_i - 1) + \lambda_{i0}(1 - z_i)\partial_{z_i}\right). \quad (17)$$

Thus the mapping from the creation-annihilation operators to the poly-differential operators (holomorphic representation) is

$$\left(\hat{a}^+, \hat{a}\right) \to (z, \partial_z). \quad (18)$$

Looking for a steady state (time-independent) solution of (17) in the exponential form

$$\mathcal{P}_\infty(\boldsymbol{z}) = \exp\left[\sum_{i\in\mathcal{G}_0} h_i(z_i - 1)\right], \quad (19)$$

and substituting the ansatz into (16), one arrives at the following set of conditions on $\boldsymbol{h}$:

$$\sum_{i\in\mathcal{G}_0} (\lambda_{0i} - \lambda_{i0}h_i) = 0, \quad (20)$$

$$\forall i \in \mathcal{G}_0: \quad -h_i \sum_{j\neq 0}^{(i,j)\in\mathcal{G}_1} \lambda_{ij} + \sum_{j\neq 0}^{(j,i)\in\mathcal{G}_1} \lambda_{ji}h_j + \lambda_{0i} - \lambda_{i0}h_i = 0. \quad (21)$$

Although it seems that there is one more condition than the number of variables, the conditions are dependent (sum (21) over all vertices of the graph). Therefore, solving the system of inhomogeneous linear equations (21), which we restate for convenience as

$$\hat{\Lambda}\boldsymbol{h} = \boldsymbol{\lambda}_{in},$$
$$\boldsymbol{\lambda}_{in} \equiv (-\lambda_{0i}|i \in \mathcal{G}_0), \ \hat{\Lambda} = (\Lambda_{ij}|i, j \in \mathcal{G}_0), \ \Lambda_{ij} = \begin{cases} -\lambda_{i0} - \sum_k^{(i,k)\in\mathcal{G}_1} \lambda_{ik}, & i = j, \\ \lambda_{ji}, & i \neq j, \end{cases} \quad (22)$$

consists in evaluating

$$\boldsymbol{h} = \hat{\Lambda}^{-1}\boldsymbol{\lambda}_{in}. \quad (23)$$

Here the existence of the steady state solution requires that: (a) $\hat{\Lambda}$ is not singular, and (b) all components of the $\boldsymbol{h}$ vector that solves (23) are positive.

Note that the form of (19) is fully factorized. Thus once the valid solution of (21) is found, the full probability of observing the system in any given state is decomposed into a product of probabilities, each evaluated at the relevant graph vertex. Recall that this occurs in spite of the fact that to find the re-normalized rates h_i one must solve a graph-global linear problem. This strong symmetry of the Poisson-In-Poisson-Out process, observed in spite of the fact that the DB is broken, is referred to (in the Queuing Theory literature) as "quasi"-DB [15, 33].

The special feature (memoryless property) of the exponential distribution is also very transparent in the creation-annihilation language. Indeed, one observes that the exponential (in quantum mechanics also referred to as "coherent") state $\exp(h\hat{a}^+)|0\rangle$ is the eigenfunction of the annihilation operator $\hat{a}$, with the eigen-value h

$$\hat{a}|\mathrm{coh}_\infty(h)\rangle = h|\mathrm{coh}_\infty(h)\rangle, \qquad |\mathrm{coh}_\infty(h)\rangle \equiv \exp(h\hat{a}^+)|0\rangle. \quad (24)$$

Therefore,

$$\hat{H}_\infty|\mathrm{coh}_\infty(\boldsymbol{h})\rangle = \left(\sum_{(i,j)\in\mathcal{G}_1}^{i,j\neq 0}\lambda_{ij}(\hat{a}_j^+ - \hat{a}_i^+)h_i + \sum_{i\in\mathcal{G}_0}\left(\lambda_{0i}(\hat{a}_i^+ - 1) + \lambda_{i0}(1-\hat{a}_i^+)h_i\right)\right) \times |\mathrm{coh}_\infty(\boldsymbol{h})\rangle, \tag{25}$$

and the stationarity condition $\hat{H}_\infty|\mathrm{coh}_\infty(\boldsymbol{h})\rangle = 0$ translates exactly into (21), where the i-th equation correspond to the condition that the c-factor in front of the corresponding $\hat{a}_i^+$ is zero.

2.2 $M/M/1$ Network

Consider a network of $M/M/1$ processes. In this case the ME adopts the following form:

$$\begin{aligned}\frac{\partial}{\partial t}P(\boldsymbol{n};t) = &\sum_{(i,j)\in\mathcal{G}_1}^{i,j\neq 0}\lambda_{ij}\big(P(\ldots,n_i+1,\ldots,n_j-1,\ldots;t) - P(\ldots,n_i,\ldots,n_j,\ldots;t)\big)\\ &+\sum_{i\in\mathcal{G}_0}\lambda_{0i}\left(P(\ldots,n_i-1,\ldots;t) - P(\ldots,n_i,\ldots;t)\right)\\ &+\sum_{i\in\mathcal{G}_0}\lambda_{i0}\left(P(\ldots,n_i+1,\ldots;t) - P(\ldots,n_i,\ldots;t)\right).\end{aligned} \tag{26}$$

Here $\theta(x)$ is the characteristic function of the logical condition x, i.e. it is unity when the condition is satisfied and zero otherwise. The corresponding Hamiltonian operator in (5), (6) is of the form

$$\hat{H}_1 = \sum_{(i,j)\in\mathcal{G}_1}^{i,j\neq 0}\lambda_{ij}(\hat{a}_j^+ - \hat{a}_i^+)\hat{b}_i + \sum_{i\in\mathcal{G}_0}\left(\lambda_{0i}(\hat{a}_i^+ - 1) + \lambda_{i0}(1-\hat{a}_i^+)\hat{b}_i\right). \tag{27}$$

Here $\hat{b}_i$ is a "skewed" annihilation operator (see e.g. [5] for a similar operational rule), such that $\hat{b}_i|n_i\rangle = \theta(n_i>0)|n_i-1\rangle$, and in both (26) and (27) we keep the same notation as in (10), (11) (respectively).

Note that $\hat{b}$ is expressed in terms of $\hat{a}$ and $\hat{a}^+$ in an extremely nonlinear way. However, the representation allows a simple "analytic" interpretation [5]:

$$\hat{b}\sum_n p_n|n\rangle \to \frac{p(z)-p(0)}{z}, \quad \text{where } p(z) = \sum_n p_n z^n. \tag{28}$$

For the introduced generating function representation, the analog of (16), (17) becomes

$$\begin{aligned}\partial_t\mathcal{P}(\boldsymbol{z};t) = &\sum_{(i,j)\in\mathcal{G}_1}^{i,j\neq 0}\lambda_{ij}(z_j - z_i)\frac{\mathcal{P}(\boldsymbol{z};t) - \mathcal{P}(\boldsymbol{z}_{\sim i};t)}{z_i} + \sum_{i\in\mathcal{G}_0}\lambda_{0i}(z_i-1)\mathcal{P}(\boldsymbol{z};t)\\ &+\sum_{i\in\mathcal{G}_0}\lambda_{i0}(1-z_i)\frac{\mathcal{P}(\boldsymbol{z};t) - \mathcal{P}(\boldsymbol{z}_{\sim i};t)}{z_i}.\end{aligned} \tag{29}$$

Here $\boldsymbol{z}_{\sim i} \equiv ((1-\delta_{ij})z_j|\, j\in\mathcal{G}_0)$. In words, this is the vector $\boldsymbol{z}$ with the component z_i replaced by zero.

Let us come back to the skewed-creation-annihilation representation, and note that the coherent states associated with this "skewed" annihilation operator $\hat{b}$ were discussed in [5]. The approach can also be traced back in the Queuing Theory literature to the classic papers of Massey [56, 57] on the operator approach to Jackson networks. The coherent states for $\hat{b}$ are constructed as follows:

$$\hat{b}|\mathrm{coh}_1(h)\rangle = h|\mathrm{coh}_1(h)\rangle, \quad |\mathrm{coh}_1(h)\rangle \equiv \frac{1}{1-h\hat{a}^+}|0\rangle. \tag{30}$$

Then the analog of (25) becomes

$$\begin{aligned}&\hat{H}_1|\mathrm{coh}_1(\boldsymbol{h})\rangle \\ &= \left(\sum_{(i,j)\in\mathcal{G}_1}\lambda_{ij}(\hat{a}_j^+ - \hat{a}_i^+)h_i + \sum_{i\in\mathcal{G}_0}\left(\lambda_{0i}(\hat{a}_i^+ - 1) + \lambda_{i0}(1-\hat{a}_i^+)h_i\right)\right)|\mathrm{coh}_1(\boldsymbol{h})\rangle.\end{aligned} \tag{31}$$

Furthermore, the condition of stationarity, $\hat{H}_1|\mathrm{coh}_1(\boldsymbol{h})\rangle = 0$, translates exactly into (21), where the i-th equation corresponds to the condition that the c-factor in front of the respective $\hat{a}_i^+$ is zero. We conclude that the stationary distribution of the $M/M/1$-network is

$$\mathcal{P}_1(z) = \prod_{i\in\mathcal{G}_0}\frac{1-h_i}{1-h_i z_i}, \tag{32}$$

where $\boldsymbol{h}$ is the solution of (21).

2.3 $M/M/m$ Network

The ME in the general case of an inhomogeneous $M/M/m$-network, with positive integers m_i (number of tellers) assigned to each vertex i of the graph, can be represented by

$$\begin{aligned}\frac{\partial}{\partial t}P(\boldsymbol{n};t) = &\sum_{(i,j)\in\mathcal{G}_1}\lambda_{ij}\big(\theta_{m_i}(n_i+1)\theta(n_j>0)P(\ldots,n_i+1,\ldots,n_j-1,\ldots;t)\\ &-\theta_{m_i}(n_i)P(\ldots,n_i,\ldots,n_j,\ldots;t)\big)\\ &+\sum_{i\in\mathcal{G}_0}\lambda_{0i}\left(\theta(n_i>0)P(\ldots,n_i-1,\ldots;t)-P(\ldots,n_i,\ldots;t)\right)\\ &+\sum_{i\in\mathcal{G}_0}\lambda_{i0}\left(\theta_{m_i}(n_i+1)P(\ldots,n_i+1,\ldots;t)-\theta_{m_i}(n_i)P(\ldots,n_i,\ldots;t)\right),\end{aligned} \tag{33}$$

$$\theta_m(n) = \min(n,m). \tag{34}$$

The evolution operator (Hamiltonian) becomes

$$\hat{H} = \sum_{(i,j)\in\mathcal{G}_1}^{i,j\neq 0}\lambda_{ij}(\hat{a}_j^+ - \hat{a}_i^+)\hat{b}_i^{(m_i)} + \sum_{i\in\mathcal{G}_0}\left(\lambda_{0i}(\hat{a}_i^+ - 1) + \lambda_{i0}(1-\hat{a}_i^+)\hat{b}_i^{(m_i)}\right), \tag{35}$$

$$\hat{b}^{(m)}|n\rangle = \theta_m(n)|n-1\rangle. \tag{36}$$

Thus the problem of finding the stationary solution is reduced (pretty much like before in the $m = \infty$ and $m = 1$ cases) to constructing coherent states for the annihilation operator $\hat{b}^{(m)}$:

$$\hat{b}^{(m)}|\mathrm{coh}_m(h)\rangle = h|\mathrm{coh}_m(h)\rangle, \quad |\mathrm{coh}_m(h)\rangle \equiv g_m(h\hat{a}^+)|0\rangle, \tag{37}$$

$$g_m(x) \equiv \sum_{k=0}^{\infty} \frac{x^k}{\prod_{l=1}^{k} \theta_m(l)} = \frac{m^m}{m!}\frac{1}{1-x/m} + \sum_{k=0}^{m-1} x^k \left(\frac{1}{k!} - \frac{m^{m-k}}{m!}\right). \tag{38}$$

Finally, the full expression for the generating function of the stationary solution over the general network becomes

$$\mathcal{P}(\boldsymbol{z}) = \prod_{i\in\mathcal{G}_0} \frac{g_{m_i}(h_i z_i)}{g_{m_i}(h_i)}, \tag{39}$$

where $\boldsymbol{h}$ is the solution of (21). Therefore, by (2),

$$P(\boldsymbol{n}) = Z^{-1} \prod_{i\in\mathcal{G}_0} \frac{h_i^{n_i}}{\prod_{l_i=1}^{n_i} \theta_{m_i}(l_i)}. \tag{40}$$

Obviously, (39), (40) are consistent with (19) and (32) when $\boldsymbol{m} = (m_i | i \in \mathcal{G}_0)$ is set to $\boldsymbol{m} = \infty$ and $\boldsymbol{m} = \boldsymbol{1}$ (respectively).

Note (for the sake of accurateness) that when deriving (40) in the operator formalism we took advantage of the important fact that both left (bra-) and right (ket-) zero eigenvalues of the Hamiltonian (35), described by $\langle \boldsymbol{0}| \exp(\sum_j \hat{a}_j)$ and $\prod_{i\in\mathcal{G}_0} g_{m_i}(h_i\hat{a}^+)|\boldsymbol{0}\rangle$ respectively, are explicitly known.

Note that one can recalculate any moment of n from either (39) or (40). In particular, for the first moment at a station we arrive at

$$\langle n_i \rangle = \frac{\langle \boldsymbol{0}| \exp(\sum_{j\in\mathcal{G}_0} \hat{a}_j)\hat{a}_i^+\hat{a}_i \prod_{k\in\mathcal{G}_0} g_{m_k}(h_k\hat{a}^+)|\boldsymbol{0}\rangle}{\langle \boldsymbol{0}| \exp(\sum_{j\in\mathcal{G}_0} \hat{a}_j) \prod_{k\in\mathcal{G}_0} g_{m_k}(h_k\hat{a}^+)|\boldsymbol{0}\rangle} = \frac{\partial}{\partial z_i}\frac{g_{m_i}(h_i z_i)}{g_{m_i}(h_i)}\bigg|_{z=1}, \tag{41}$$

where $g_m(x)$ is taken from (38) and (as before) $\boldsymbol{h}$ is the solution of (21). Note that the moments are finite only if $h_i < m_i$, which thus defines the condition for Q-network stability (statistical stationarity) [15].

3 Statistics of Network Currents

A (general) queueing network is naturally characterized by the actual current of particles/jobs going through and being processed at each station according to a certain service discipline (e.g. First-In-First-Out (FIFO)). Here the quasi-current characterizes the activities of tellers, not focusing on the dynamics of the individual particles at all. In other words, actual current tracks the dynamics of the jobs/particles, while quasi-current tracks quasi-particles/jobs assuming that all the jobs waiting for service at a station are fully equivalent and not prioritized. Actual current and quasi-current coincide in the case of an $M/M/\infty$ network, when the individual jobs do not interact at all, as well as any network in which all particles (customers) are identical. With this disclaimer, we will be discussing quasi-currents for the remainder of the paper, and refer to them as currents to simplify our exposition.

As shown below, calculating the statistics of the currents in a Jackson network becomes tractable in the two regimes that we will study: the "uncongested" and "congested" regimes.

In the "uncongested" regime, this arises from the product-form symmetry characterizing the regime (an extension of the product-form symmetry discussed earlier), as well as the $\boldsymbol{m}$-independence property (universality). Note that for networks without feedback, the statement of $\boldsymbol{m}$-independence is equivalent to the fact that in steady state, the flow along arcs will be Poisson (with rate independent of the number of servers). As mentioned previously, this phenomenon was discovered earlier in the Queueing Theory literature [38, 60, 61]. However (and to the best of our knowledge), the extension of this statement to the asymptotic (large time) limit for more general networks (unscaled, not in the fluid limit or hydrodynamic limit) has not been formally explored in the literature.

The remainder of this section is partitioned into four Subsections. We start from a general discussion of the current related objects in Sect. 3.1. In Sect. 3.2 we consider the uncongested regime. We also develop several generalizations of the Doi-Peliti technique to account for currents, and develop some machinery necessary for the statement of our results. At the end of the section, our analysis naturally leads to the identification of the uncongested regime's breakdown, namely the identification of a phase transition in the space of currents. We also show that this transition is second-order, thus translating into smoothness of the Crámer function of currents (continuity of the first and the second derivatives) at the transition. In Sect. 3.3, we extend the coherent state formalism to the "congested" regime via a simple reduction of the network graph. We note that a similar decomposition was applied in [25], and is similar in spirit to many such reductions appearing throughout the Queueing Theory literature [62–68].

Finally, in Sect. 3.4 we illustrate the general theory using our enabling example of a single node with feedback.

3.1 Preliminary General Remarks

We will mainly be interested to evaluate the joint distribution function of the currents and queue sizes where the latter are averaged over the entire time horizon. In the following we will use $P(\bar{\boldsymbol{n}}, \boldsymbol{J}|t)$ notation for the main object of interest. However it is technically more convenient to start from another (and to a degree auxiliary) object defined as a joint distribution function of currents and queues where the latter are observed at the final moment of time. We will see below that it is important to differentiate these generally distinct objects.

Let $P(\boldsymbol{n}(t), \boldsymbol{J}|t)$ denote the joint probability distribution function of the queue size at the final moment of time t, $\boldsymbol{n}(t) = (n_i(t)|i \in \mathcal{G}_0)$, and currents accumulated over the $[0; t]$ interval of time, $\boldsymbol{J} = (J_{ij}|(i, j) \in \mathcal{G}_1)$, where the latter are defined on all edges of the graph and the former are defined (as before) on vertices. The ME for this object is the natural generalization of (33), which we now present in operator form (to allow for more compact notations). In particular, the operator form of the ME for $P(\boldsymbol{n}(t), \boldsymbol{J}|t)$ is

$$\partial_t|\boldsymbol{s}(\boldsymbol{n}(t); \boldsymbol{J})\rangle = \left(\hat{H} + \sum_{(i,j)\in\mathcal{G}_1} \hat{J}_{ij}\right)|\mathbf{s}(\boldsymbol{n}(t); \boldsymbol{J})\rangle, \tag{42}$$

$$\forall i, j \neq 0: \quad \hat{J}_{ij} = \lambda_{ij}(\hat{a}_{ij}^{+} - 1)\hat{a}_j^{+}\hat{b}_i^{(m_i)}, \tag{43}$$

$$\hat{J}_{0i} = \lambda_{0i}(1 - \hat{a}_{0i}^{+})\hat{a}_i^{+}, \qquad \hat{J}_{i0} = \lambda_{i0}(\hat{a}_{i0}^{+} - 1)\hat{b}_i^{(m_i)}, \tag{44}$$

where $\hat{H}$ is defined in (35). Here we have assumed that the currents are discrete and positive, and the respective ket-vector is related to the joint PDF of $\mathbf{n}(t)$ and $\mathbf{J}$ as follows:

$$|\boldsymbol{s}(\boldsymbol{n}(t);\boldsymbol{J})\rangle = P(\boldsymbol{n}(t);\boldsymbol{J})|\boldsymbol{n};\boldsymbol{J}\rangle. \tag{45}$$

$\hat{J}_{ij}$ (in (43)) is the operator for the amount of current from site i to site j, and $\hat{a}^+_{ij}$ is the newly introduced creation operator (at edge (i,j)) acting on the space of discrete positive currents. We define operators for incoming and outgoing currents in (44) similarly. Formal solution of (42) is

$$|\mathbf{s}(\boldsymbol{n}(t);\boldsymbol{J})\rangle = \exp\left(t\left(\hat{H} + \sum_{(i,j)\in\mathcal{G}_1}\hat{J}_{ij}\right)\right)|\mathbf{s}(\boldsymbol{n}(0);\boldsymbol{J})\rangle, \tag{46}$$

where the ket-state on the rate correspond to the "initial" steady distribution of queues, described by (40), and zero initial current: $|\mathbf{s}(\boldsymbol{n}(0);\boldsymbol{J})\rangle = |s\rangle \otimes |\boldsymbol{J}=\mathbf{0}\rangle$, where we follow notations introduced in the introduction and $|s\rangle = \sum_{\boldsymbol{n}} P(\boldsymbol{n})|\boldsymbol{n}\rangle$.

It follows that the generating function over the currents $(i,j)\in\mathcal{G}_1$, accounting for the incoming ($(0,i)\in\mathcal{G}_1$) and outgoing ($(i,0)\in\mathcal{G}_1$) arcs, is

$$|\boldsymbol{s}_{\boldsymbol{q}}(\boldsymbol{n}(t))\rangle = \sum_{\boldsymbol{J}}\prod_{(i,j)\in\mathcal{G}_1} q_{ij}^{J_{ij}}|\boldsymbol{s}(\boldsymbol{n}(t);\boldsymbol{J})\rangle. \tag{47}$$

According to our standard birth-death [creation/annihilation] rules, the object described by (47) satisfies

$$\partial_t|\boldsymbol{s}_{\boldsymbol{q}}(\boldsymbol{n}(t))\rangle = \hat{H}_{\boldsymbol{q}}|\boldsymbol{s}_{\boldsymbol{q}}(\boldsymbol{n}(t))\rangle, \quad |\boldsymbol{s}_{\boldsymbol{q}}(\boldsymbol{n}(t))\rangle = \exp(t\hat{H}_q)\sum_{\boldsymbol{n}}P(\boldsymbol{n})|\boldsymbol{n}\rangle = \exp(t\hat{H}_q)|s\rangle, \tag{48}$$

$$\begin{aligned}\hat{H}_{\boldsymbol{q}} = &\sum_{(i,j)\in\mathcal{G}_1}\lambda_{ij}(\hat{a}^+_j - \hat{a}^+_i)\hat{b}^{(m_i)}_i \\ &+ \sum_{(i,j)\in\mathcal{G}_1}^{i\neq 0, j\neq 0}\lambda_{ij}(q_{ij}-1)\hat{a}^+_j\hat{b}^{(m_i)}_i + \sum_{(0,i)\in\mathcal{G}_1}\lambda_{0i}(q_{0i}a^+_i - 1) + \sum_{(i,0)\in\mathcal{G}_1}\lambda_{i0}(q_{i0}-a^+_i)\hat{b}^{(m_i)}_i.\end{aligned} \tag{49}$$

Note that even though our evaluation in (48) applies to the general object, namely to the joint distribution function of currents over the entire network, one may compute the relevant marginals (say the distribution function for the current over a single edge) in a straightforward manner. We also note that the operator $\hat{H}_{\boldsymbol{q}}$ can be obtained by modifying/twisting the evolution operator $\hat{H}$, given by (36), as follows. One weights the off-diagonal terms of $\hat{H}$ in the space of populations $|\boldsymbol{n}\rangle$, namely $\lambda_{ij}\hat{a}^+_j\hat{b}^{(m_i)}_i$, $\lambda_{0i}a^+_i$, and $\lambda_{i0}\hat{b}^{(m_i)}_i$, with the factors q_{ij}, q_{0i}, and q_{i0} respectively. This corresponds to viewing the underlying stochastic process as a Markov chain on an infinite graph, whose nodes are labeled by pure states $|\boldsymbol{n}\rangle$, whereas links represent the set of processes allowed by the evolution operator $\hat{H}$. Within such a picture, the set of $\boldsymbol{q}$ parameters plays the role of discrete gauge fields (vector potentials), and $\hat{H}_{\boldsymbol{q}}$ is interpreted as the evolution operator, "twisted" by the gauge field $\boldsymbol{q}$, as described in [69]. Since here we are dealing with oriented graphs, no constraints are imposed on $\boldsymbol{q}$.

It is also useful to consider the distribution function of queue at the finite moment of time, conditioned to specific value of the current generating parameter $\boldsymbol{q}$. This object, and

the respective first moment of queue size, become

$$P_{\boldsymbol{q}}(\boldsymbol{n}(t)) \propto \langle \boldsymbol{n}(t)| \exp\left(t\hat{H}_q\right)|s\rangle, \tag{50}$$

$$\langle n_i(t)\rangle_{\boldsymbol{q}} = \frac{\langle \mathbf{0}| \exp(\sum_{j\in\mathcal{G}_0}\hat{a}_j)\hat{a}_i^+\hat{a}_i \exp(t\hat{H}_q)|s\rangle}{\langle \mathbf{0}| \exp(\sum_{j\in\mathcal{G}_0}\hat{a}_j) \exp(t\hat{H}_q)|s\rangle}. \tag{51}$$

Returning back to our main object of interest (the joint distribution function of the current and of the queue averaged over the time horizon) and following the same formalism/notations, we derive the analogs of (50), (51)

$$P_{\boldsymbol{q}}(\bar{\boldsymbol{n}}) \propto t^{-1}\int_0^t dt' \langle \mathbf{0}| \exp\left(\sum_{j\in\mathcal{G}_0}\hat{a}_j\right) \exp\left((t-t')\hat{H}_q\right)|\boldsymbol{n}(t')\rangle\langle \boldsymbol{n}(t')| \exp\left(t'\hat{H}_q\right)|s\rangle, \tag{52}$$

$$\langle \bar{n}_i\rangle_{\boldsymbol{q}} = \frac{\int_0^t dt' \langle \mathbf{0}| \exp(\sum_{j\in\mathcal{G}_0}\hat{a}_j) \exp((t-t')\hat{H}_q)\hat{a}_i^+\hat{a}_i \exp(t'\hat{H}_q)|s\rangle}{\int_0^t dt' \langle \mathbf{0}| \exp(\sum_{j\in\mathcal{G}_0}\hat{a}_j) \exp(t\hat{H}_q)|s\rangle}. \tag{53}$$

3.2 Uncongested Regime

We are now in a position to introduce the "uncongested" regime on a formal level. We will characterize this regime in terms of the existence of a special "universal" product-form solution to (48), as well as the finiteness of a particular expectation value (capturing the fact that no condensation of particles occurs at any nodes), at sufficiently large observational time t. Note that in the spirit of our operator-theoretic framework, we define the regime in terms of *both* the queueing network *and* the vector $\boldsymbol{q}$ on which one evaluates the generating function for occupation numbers in the network. Thus a given queueing network may be in the regime for some evaluations of its generating function (certain values of $\boldsymbol{q}$) but not for others. This will then be related to belonging (or not belonging) to the regime for different types of large deviations (of current) through the standard Legandre transform, which maps the generating function (evaluated at different $\boldsymbol{q}$) to the Crámer function (evaluated at different-sized deviations).

We say that a given queueing network is in the "uncongested" regime for a given vector $\boldsymbol{q}$ if the ket vector $|\boldsymbol{s}_{\boldsymbol{q}}(\boldsymbol{n})\rangle$ is dominated by the ground state of $\hat{H}_q$, i.e. at sufficiently large time, $|\boldsymbol{s}_{\boldsymbol{q}}(\boldsymbol{n})\rangle \sim \exp(-\Delta(\boldsymbol{q})t)|\mathrm{coh}_{\boldsymbol{m}}(\boldsymbol{h}(\boldsymbol{q}))\rangle$, holds. In other words, the spectrum of $\hat{H}_q$ is such that its ground state is separated from the excited states by a finite gap, and thus at the times much large than inverse value of the gap the solution is completely described by the ground state only, thus providing respective universality. We also require that $\langle \bar{n}_i\rangle_{\boldsymbol{q}} < \infty$ for all nodes i of the network in the "uncongested" regime, ensuring that the expected number of particles (size of the queue averaged over time) does not diverge at any nodes as $t \to \infty$.

The statistics of queueing networks in the "uncongested" regime may be analyzed using the techniques from Sect. 2. In particular, we substitute $|\boldsymbol{s}_{\boldsymbol{q}}(\boldsymbol{n})\rangle$ by $\sim \exp(-\Delta(\boldsymbol{q})t)|\mathrm{coh}_{\boldsymbol{m}}(\boldsymbol{h}(\boldsymbol{q}))\rangle$ in (48) to arrive at

$$\sum_{i\in\mathcal{G}_0}(q_{0i}\lambda_{0i} - \lambda_{i0}q_{i0}h_i(\boldsymbol{q})) = \Delta(\boldsymbol{q}), \tag{54}$$

$$\forall i \in \mathcal{G}_0: \quad -h_i(\boldsymbol{q})\sum_{j\in\mathcal{G}_0}^{(i,j)\in\mathcal{G}_1}\lambda_{ij} + \sum_{j\in\mathcal{G}_0}^{(j,i)\in\mathcal{G}_1} q_{ji}\lambda_{ji}h_j(\boldsymbol{q}) + \lambda_{0i} - \lambda_{i0}h_i(\boldsymbol{q}) = 0. \tag{55}$$

This set of relations generalizes the stationary ($\boldsymbol{q}=\mathbf{1}$) relations (20), (21), and are thus consistent with $\Delta(\mathbf{1})=0$. Equations (54), (55) describe the right (ket) eigen-function of the ground-state of the evolution operator/Hamiltonian (49), while the corresponding left (bra)-eigenfunction is described in Appendix.

Note that, replacing all internal q-variables by unity, the basic set of equations for $\boldsymbol{h}$ does not depend on the remaining q_{0i} and q_{i0} components. It follows that the respective $\boldsymbol{h}$ are identical to the one derived before for the stationary $\boldsymbol{h}(\mathbf{0})$ (no currents) setting. Moreover, $\Delta(\boldsymbol{q}_0)=\sum_i\left((q_{0i}-1)\lambda_{0i}-\lambda_{i0}(q_{i0}-1)h_i(\mathbf{0})\right)$. This observation translates (after the obvious Legandre transform) into the statement that all the currents entering and leaving the network are asymptotically Poisson, which (as already mentioned) was known previously in the Queuing literature [33, 61].

For the sake of simplicity, hereafter we exclude the incoming and outgoing currents from consideration, which is achieved by setting $q_{i0}=q_{0i}=1$. The consistency of (54) and (55) (the two are just generalized versions of (21), (20)) translates into the following expression for the lowest eigenvalue:

$$\Delta(\boldsymbol{q})=-\sum_{(i,j)\in\mathcal{G}_1}^{i,j\in\mathcal{G}_0}h_i(\boldsymbol{q})\lambda_{ij}(q_{ij}-1). \tag{56}$$

For sufficiently large t (it is our Large Deviation Parameter and thus should at least be significantly larger than the correlation time of the system) we obtain the following asymptotic expression for $P_{\boldsymbol{q}}(t)=\sum_{\boldsymbol{n}}P_{\boldsymbol{q}}(\bar{\boldsymbol{n}})$,

$$P_{\boldsymbol{q}}(t)=\Psi(\boldsymbol{q})\exp(-t\Delta(\boldsymbol{q}))\sim\exp\left(t\sum_{(i,j)\in\mathcal{G}_1}^{i,j\in\mathcal{G}_0}h_i(\boldsymbol{q})\lambda_{ij}(q_{ij}-1)\right), \tag{57}$$

$$\Psi(\boldsymbol{q})\equiv\langle\mathbf{0}|\exp\left(\sum_{i\in\mathcal{G}_0}\bar{h}_i(\boldsymbol{q})\hat{a}_i\right)|\mathrm{coh}_{\boldsymbol{m}}(\boldsymbol{h}(\boldsymbol{q})\rangle. \tag{58}$$

Here, in evaluating the pre-exponential factor $\Psi(\boldsymbol{q})$, we have used the assumption that the main contribution into (52) originates from $t',t-t'\gg 1/\Delta(\boldsymbol{q})$. Thus, $\langle\mathbf{0}|\exp(\sum_i\bar{h}_i(\boldsymbol{q})\hat{a}_i)$ in (58) is the bra-vector defined as the left eigenvector of $\hat{H}_{\boldsymbol{q}}$ with the same eigenvalue $\Delta(\boldsymbol{q})$ (see Appendix for more details).[3] Here certain time-independent pre-factors (which do not impact the asymptotics up to exponential order) are ignored on the rhs. As we are interested in the statistics of the currents that scale (grow) linearly with t, one can replace the sum in (47) by an integral, invert the relation, and arrive at the following saddle-point (large-deviation) expression:

$$P(\boldsymbol{J}|t)\sim\int\mathcal{P}_{\boldsymbol{q}}(t)\prod_{(i,j)\in\mathcal{G}_1}^{i,j\in\mathcal{G}_0}q_{ij}^{-J_{ij}}\sim\exp\left(-t\mathcal{S}(\boldsymbol{J}/t)\right),$$

$$\mathcal{S}(\boldsymbol{j})=\sum_{(i,j)\in\mathcal{G}_1}^{i,j\in\mathcal{G}_0}\left(j_{ij}\ln(q^*_{ij})+h_i(\boldsymbol{q}^*)\lambda_{ij}(1-q_{ij})\right), \tag{59}$$

[3]Note that expression for the analog of $\Psi(\boldsymbol{q})$ correspondent to $P_{\boldsymbol{q}}(\boldsymbol{n}(t))$ is significantly different: $\langle\mathbf{0}|\exp(\sum_{i\in\mathcal{G}_0}\hat{a}_i)|\mathrm{coh}_{\boldsymbol{m}}(\boldsymbol{h}(\boldsymbol{q})\rangle$.

$$\forall (k,l) \in \mathcal{G}_1 \;\&k,l \in \mathcal{G}_0\text{:} \quad j_{kl}/q^*_{kl} = \sum_{(i,j)\in\mathcal{G}_1}^{i,j\in\mathcal{G}_0} \frac{\partial h_i(\boldsymbol{q}^*)}{\partial q_{kl}}(q^*_{ij}-1)\lambda_{ij} + h_k(\boldsymbol{q}^*)\lambda_{kl}, \quad (60)$$

where $\mathcal{S}(\boldsymbol{j})$ is a convex function of its argument, also called the Crámer (or large-deviation) function. The dependence of the Crámer function on the current production $\boldsymbol{j}$ is defined implicitly via (60) and (55).

Since Δ is fully defined by the solution of (55), which is independent of $\boldsymbol{m}$, the resulting expression for the Crámer function is also $\boldsymbol{m}$-independent in the "uncongested" regime. This cancelation is quite remarkable. We note that the degeneracy is especially interesting, since it does not seem to extend to the time-independent pre-factor in $P(\boldsymbol{J}|t)$, $\Psi(\boldsymbol{q})$ (defined in (58)). A Queueing Theory interpretation is that in this regime the large deviations are *not* caused by the interactions of different particles in the network, and thus *the same* kind of deviations would have occurred even if all nodes in the network were of $M/M/\infty$ type (as opposed to $M/M/m$ with $m<\infty$), in which different particles cannot interact and delay one-another in the network.[4]

As explained above, the qualitative assumption that allowed us to extend the product-form ansatz to statistics of currents in (54)–(60) was that the number of particles in the system (size of the queue averaged over time) did not diverge with time. For this assumption to hold, it must be the case that

$$\forall i \in \mathcal{G}_0\text{:} \quad \langle \bar{n}_i \rangle_{\boldsymbol{q}} = \frac{\langle \mathbf{0}| \exp(\sum_{j\in\mathcal{G}_0} \bar{h}_j(\boldsymbol{q})\hat{a}_j)\hat{a}^+_i\hat{a}_i \prod_{k\in\mathcal{G}_0} g_{m_k}(h_k(\boldsymbol{q})\hat{a}^+)|\mathbf{0}\rangle}{\langle \mathbf{0}| \exp(\sum_{j\in\mathcal{G}_0} \bar{h}_j(\boldsymbol{q})\hat{a}_j) \prod_{k\in\mathcal{G}_0} g_{m_k}(h_k(\boldsymbol{q})\hat{a}^+)|\mathbf{0}\rangle} < \infty, \quad (61)$$

where $\boldsymbol{h}(\boldsymbol{q})$ and $\bar{\boldsymbol{h}}(\boldsymbol{q})$ are solutions of (54), (55) and (76), (77) (which describe consistently the right/ket and left/bra ground state eigen-function of the evolution operator/Hamiltonian (49)).

We now comment on the phase transition which takes us from the uncongested to the congested regime. Picking a direction in the vector space of currents (dimensionality of the space is equal to the number of internal directed edges of the graph) and increasing the length of the vector in this direction, gradually (starting from the domain of values belonging to the "uncongested" regime) we will eventually reach the regime where congestion occurs, i.e. the condition (61) breaks down and particles accumulate over time at some nodes (condensation). To see this heuristically, we note that for large values of q, the main contribution to the generating function will be highly skewed towards very large values of current (due to the fact that q is raised to a power equal to the current), and the most likely way to attain such large values of currents will be for the queueing network to remain totally occupied over the entire time horizon, causing the number of particles in the network to diverge over time. Put in the context of large deviations, the most likely way to attain very large values of currents is to have an accumulation over time of the number of particles in the network (condensation), leading to the system being overloaded over the entire time horizon, enabling all links in the network to generate current continuously over the entire time horizon. We focus in particular on the setting where the "breakdown" of the regime occurs initially at some particular unique node of the network, the expected number of particles in the system diverges at that node (over time), and a certain spectral condition holds at this breakdown

[4] Similar "mysterious" cancelation of the vorticity dependence was reported in [13] in the Crámer function of the entropy production for a polymer stretched by shear-vorticity flow.

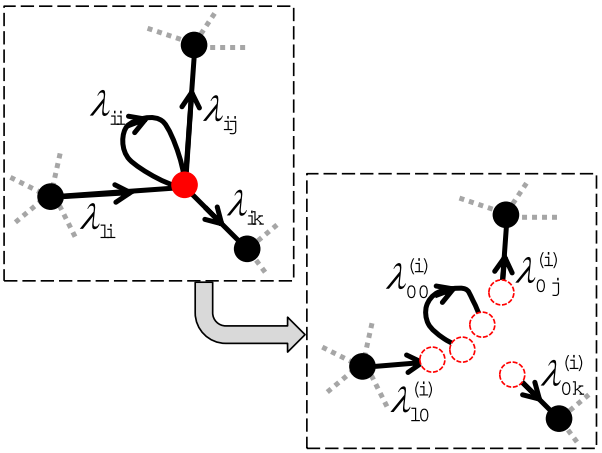

Fig. 2 (Color online) Transformation of the "overloaded" node, discussed in the text. The component of the graph, associated with the node, before and after the transformation are shown in the *left-upper* and *low-right corners* respectively. *Dashed node* on the modified graph correspond to new injections and departures with the Poisson rates exactly equivalent to the respective transition rates on the original graph

point (described in some detail below). We will use these characteristics to formally define our "congested" regime.

We now elaborate on the aforementioned spectral condition. In particular, the transition point from the "uncongested" to the "congested" regime has an interpretation in terms of the closing of the gap between the ground state and the low boundary of the continuous spectrum of (49). This "closing the gap" picture also assumes that other excited discrete states (which are present already in the case of a single station with feedback and $m > 1$, see the discussion in Sect. 4) do not cross the factorized ground state found above. This spectral interpretation also leads to an immediate conclusion/consequence in terms of the Crámer function shape at the values of the current that correspond to the considered uncongested-to-congested transition. Indeed, this "closing the gap" scenario translates into continuity of the Crámer function and its first derivatives at the transition. Stated differently (in the jargon of phase transition theory), the dynamical uncongested-to-congested transition with respect to the currents is second-order, and $1/\langle \bar{n}_i \rangle_{\boldsymbol{q}}$ plays the role of the order parameter at the congested node.

Our final remark is about comparison of the distribution of queue averaged over time with respective distribution measured at the final moment t. The asymptotic analysis, described above for uncongested regime and extended in the following Subsection to congested regime, suggests that $\langle n_i(t) \rangle_{\boldsymbol{q}}$ is finite at any value of q and $t \to \infty$, in particular for $q = q_c$ where $\langle \bar{n}_i \rangle_{\boldsymbol{q}} = \infty$. Moreover, one conjectures that the two distribution functions, $P_{\boldsymbol{q}}(\bar{n})$ and $P_{\boldsymbol{q}}(n(t))$, are different at any values of q except of the special one correspondent to the minimum of the Crámer function achieved at $\boldsymbol{J} = \langle \bar{J} \rangle$. This statement may be interpreted as a breakdown of ergodicity for any current but the special one correspondent to the steady distribution.

3.3 Congested Regime

We now consider a heuristic decomposition approach to extend the product-form description, utilized above in the uncongested regime, beyond the domain bounded by (61). We note that similar decomposition techniques have been considered throughout the Queueing theory literature to understand how networks become congested [62–68, 70, 71].

Consider, for example, crossing the boundaries of (61) along a direction in the $\boldsymbol{q}$-space, and thus violating the condition at some $\boldsymbol{q}_*$ at a single node, say i. Then, exactly at the point

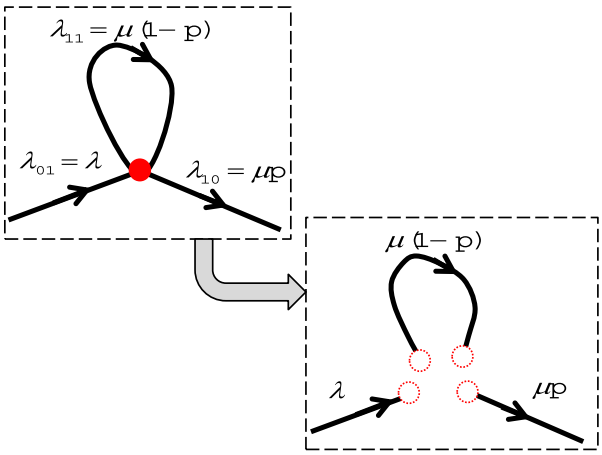

Fig. 3 (Color online) Original graph and transformed graph correspondent to uncongested and congested regimes for example of the single station with the feedback

of crossing we can simply assume that the node is always congested (has an infinite queue in the waiting room), and thus the tellers at the node stay busy over the entire time horizon. This translates into the following obvious modification of the network graph: remove the node, and associate all the in/out edges for the node of the old graph to the new open-system in/out nodes with exactly the same rates. The transformation is shown in Fig. 2. In the new domain (54)–(55) should be considered on the modified graph, and one needs to add the appropriate constant to $\Delta(\boldsymbol{q})$ from (56) to guarantee its continuity at $\boldsymbol{q}_*$. Equivalently, one must compensate all probabilities to reflect the rare event that the given node remained busy over the entire duration.

The general construction is illustrated in Fig. 3, and will also be illustrated in the next Subsection on our enabling example of a single node system with feedback. We note, of course, that in general the most likely way for rare events to occur in queueing networks may be quite complicated (see for example [22]), and what was explained above should be considered as an approximation which is not proven rigorously.

We conclude this general construction by explaining the formal status of our derived results. All the asymptotic derivations so far, discussed in both the congested and uncongested cases, have relied on several important assumptions. The most important of these is the equivalence of the coherent state solution to the lowest eigen-value (ground state) solution. We have assumed that this (physically very plausible) assumption holds, and focused primarily on constructing the coherent state solution explicitly. We note, however, that our derivations have generally been non-rigorous in nature, and this assumption was not justified in any particular case. A considerable difficulty associated with formalizing our results (and justifying this assumption) lie in the fact that the coherent state approach does not allow us to analyze the entire spectrum of the operator. In view of the above, we find it important to validate this assumption, at least for a special case. We now proceed along these lines by studying the aforementioned case of a single station with feedback.

3.4 Single Station with Feedback

Perhaps the simplest example of a (non-trivial) queueing network (as discussed in Sect. 1.3) consists of a single station, one incoming channel/edge, one outgoing channel/edge, and one self-loop, all characterized by Poisson processes with rates λ, μp and $\mu(1-p)$ respectively. The network is shown in the upper left corner of Fig. 3.

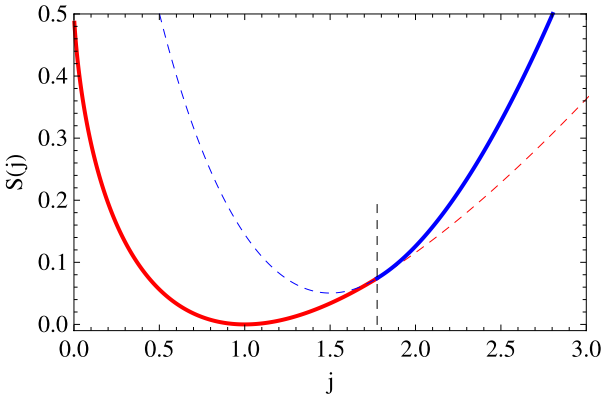

Fig. 4 (Color online) Crámer function, $\mathcal{S}(j)$, of the feedback current shown for $\lambda = m = 1$, $p = 1/2$ and $\mu = 3$. *Thick red* and *thick blue curves* describe the uncongested and congested domains respectively. *Dashed black line* marks the value of j_c

In this example, the only interesting current is associated with the self-loop (as the others are Poissonian in the steady-state, see Sect. 1.3). We thus focus exclusively on the analysis of the current along the feedback arc, and the associated generating function (evaluated for the scalar q and corresponding current j).

We start by considering the "uncongested" regime. Then (54), (56) become

$$h = \frac{\lambda}{\mu(1-(1-p)q)}, \qquad \Delta = -\frac{\lambda(q-1)(1-p)}{1-(1-p)q}. \tag{62}$$

According to (59), (60), this results in

$$j < j_c: \quad \mathcal{S}(j) = \lambda(1-p/2) - \frac{\sqrt{p\lambda(4j+p\lambda)}}{2} + j\ln\left(\frac{2j+p\lambda-\sqrt{p\lambda(4j+p\lambda)}}{2j(1-p)}\right). \tag{63}$$

To identify the "uncongested" regime breakdown we calculate $\bar{h}$, according to (77):

$$\bar{h} = \frac{p}{1-(1-p)q}. \tag{64}$$

Substituting into (58) results in

$$\Psi(q) = g_m(\bar{h}(q)h(q)), \qquad \langle \bar{n} \rangle_q = \frac{\partial}{\partial z}\left.\frac{g_m(\bar{h}(q)h(q)z)}{g_m(\bar{h}(q)h(q))}\right|_{z=1}. \tag{65}$$

From the above and (38), we find that the expected queue length remains bounded over time iff $h\bar{h} < m$, and thus the critical value of congestion at which point one shifts regimes is

$$q_c = \frac{1-\sqrt{\frac{\lambda p}{m\mu}}}{1-p}. \tag{66}$$

The congested regime that occurs at $q > q_c$ and $j > j_c$ corresponds to the situation when the single server is always occupied, and thus (conditioning on this event) jobs continually feedback over time as a $m\mu(1-p)$ standard Poisson process. Therefore, in the congested regime

$$q > q_c: \quad \Delta = -\mu(1-p)m(q-1) + \left(\sqrt{\lambda} - \sqrt{p\mu m}\right)^2. \tag{67}$$

Here the $(\sqrt{\lambda}-\sqrt{p\mu m}\,)^2$ term (a constant with respect to q) is found in accordance with (62) and the condition of Δ-continuity at $q=q_c$. Alternatively, in the congested regime Δ can be identified with the lower edge s_- of the continuous spectrum of $\hat{H}_q$, given by (74). The spectrum of $\hat{H}_q$ for a simple model under consideration is analyzed in some detail in Sect. 4. Performing the Legendre transform on (67), we arrive at the following expression for the Crámer function of the feedback current in the congested regime

$$j>j_c:\quad \mathcal{S}(j)=\left(\sqrt{\lambda}-\sqrt{p\mu m}\right)^2+m\mu(1-p)+j\ln\left(\frac{j}{em\mu(1-p)}\right). \tag{68}$$

Comparing (63) with (68) we also observe that the Crámer function is smooth (first derivative is continuous) at $j=j_c$, thus confirming that the dynamical phase transition described here is of the second order (continuous). The change of the Crámer function shape across the transition is shown in Fig. 4.

The obtained expression for the Crámer function in the congested regime (68) has a very simple and transparent interpretation. Since the server is always occupied, the feedback current generation is a standard Poisson process, so that $\mathcal{S}(j)=\mu m(1-p)+j\ln(j/(em\mu(1-p)))$, where $e^{-t\mathcal{S}_0}$ is the probability to keep the server busy, presented with exponential accuracy. The latter is given by the probability $e^{-t\mathcal{S}_0(\omega)}$ of creating the incoming and outgoing currents of the same value ω to provide the marginal stability of the server, maximized with respect to ω. Since both currents are generated by independent Poisson processes with the rates λ and $p\mu m$, respectively, we have $\mathcal{S}_0(\omega)=\lambda+pm\mu+\omega\ln(\omega^2/(\lambda p\mu m)e^2)$, and minimization with respect to ω results in $\mathcal{S}_0=(\sqrt{\lambda}-\sqrt{p\mu m}\,)^2$, which reproduces (68).

4 Direct Analysis of the Singe-Station Feedback System

As shown in the previous sections, an understanding of the evolution operator ground state structure is essential for predicting long-time statistics of currents. We have also argued that the ground state can be described analytically for networks with general graphical structure. However, these arguments were indirectly dependent on certain information about the rest of the spectrum—specifically on the fact that the ground state is separated from the continuous spectrum by a gap which collapses at the dynamical phase transition. In general, information on the entire spectrum is difficult to obtain. The main point of this Section is to gain a broader understanding of the simple single-node network with feedback along these lines. Therefore in this Subsection we analyze the full spectrum of the problem with the single feedback and we confirm the general picture of the 'gap emergence', and collapse at the phase transition point suggested above on the basis of only partial (ground state) analysis.

Our starting point is the dynamical equation $P_q(n;t)=\sum_{j=0}^{\infty}q^jP(n,j;t)$ for the generating function of current j via the feedback arc. Here $P(n,j;t)$ is the joint probability distribution function of the number of particles, n standing at the queue at the time t, and the number of particles, j, that have passed through the feedback loop by time t. Following directly the proper generalization, according to (43)–(49) of the ME (33) we derive:

$$\begin{aligned}\frac{\partial}{\partial t}P_q(n;t)=&\lambda\left(P_q(n-1;t)-P_q(n;t)\right)+\mu p\left(\theta_m(n+1)P_q(n+1;t)-\theta_m(n)P_q(n;t)\right)\\&+\mu(1-p)(q-1)\theta_m(n)P_q(n;t),\end{aligned} \tag{69}$$

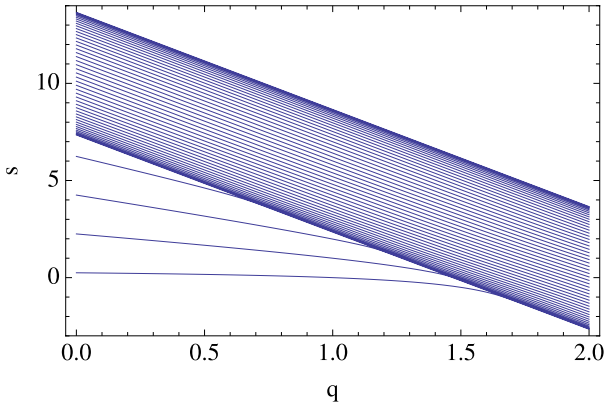

Fig. 5 (Color online) Spectrum of the evolution operator as a function of q for $m = 5$, $\lambda = 0.5$, $\mu = 2.0$, $p = 0.5$ and $N = 50$. The *lower bottom line* corresponds to the ground state solution $\Delta(q)$

where the last term on the r.h.s. accounts for the current. Performing the Laplace transform of $P_q(n;t)$ over time, $P_{s;q}(n) \equiv \int_0^\infty \exp(st) P_q(n;t) dt$ (with s considered as a spectral parameter), we arrive at the following spectral equation:

$$\begin{aligned}(s - \lambda - \mu p \theta_m(n) + \mu(1-p)(q-1)\theta_m(n)) P_{s;q}(n) \\ = -\lambda P_{s;q}(n-1) - \mu p \theta_m(n+1) P_{s;q}(n+1).\end{aligned} \tag{70}$$

Here the inhomogeneous part, dependent on the initial condition at $t = 0$, is ignored.

The "boundary" conditions over n for (70) read $P_{s;q}(-1) = 0$ and $P_{s;q}(n \to \infty) = 0$. We will relax the last condition, assuming the following natural finite waiting room regularization: $P_{s;q}(N + m) = 0$ for some sufficiently large value of N. We will study the spectrum at finite N, and show towards the end of the calculations that the $N \to \infty$ limit is well defined. We note that the particular choice of regularization turns out to be unessential for the limits we consider. In order to find the eigenvalues (spectrum) we solve the recurrence relation (70) for general values of s, and use the aforementioned boundary condition, and the normalization condition $P_{s;q}(0) = 1$. To summarize, the set of conditions that complement (70) to provide the full description of the spectrum are

$$P_{s;q}(-1) = 0, \qquad P_{s;q}(0) = 1, \qquad P_{s;q}(N + m) = 0, \tag{71}$$

and we are mainly interested in studying the $N \to \infty$ limit.

A numerical solution of the eigenvalue problem is illustrated in Fig. 5 for $m = 5$, $\lambda = 0.5$, $\mu = 2.0$, $p = 0.5$ and $N = 50$. The simulations show emergence at $q = 1$ of three discrete eigenstates well separated from the (still discrete) band of states. We tested numerically as one increases N, the $q = 0$ value of the ground state s (top line in Fig. 5) approaches 0. We observed that the asymptotic dependence of the ground state eigen-value on s is fully consistent with the prediction of the ground-state coherent-state theory explained above in Sect. 3.4, specifically with (62). Moreover, our prediction of q_c (where the gap between the ground state and the continuous band collapses), as given by (66), is fully consistent with the respective dependencies of the gap collapse in our computations. Let us also mention that the two non-ground state eigen-values observed at $q = 1$ never cross with each other or with the ground state at $q > 1$, and merge into the continuous band (consequently one after another) at values of q smaller than q_c. Experimenting with larger m, we observe that the discrete spectrum becomes equidistant in the $m \to \infty$ limit, which is fully consistent with the interaction-free nature of the limit.

We now analyze the spectrum analytically. Let us begin by making some preliminary observations. First, the linear dependence on s implies that the function $P_{s;q}(n)$ is an n-th order polynomial in s, suggesting that there are always exactly $N+m$ solutions of the last condition in (71), $P_{s;q}(N+m)=0$. Second, at $n>m$ the factor $\theta_m(n)$ becomes constant and the recursive relations (70) can be solved analytically. Thus, looking for solution of (70) at $n>m$ in the $P_{s;q}(n)=c_+\rho_+^n+c_-\rho_-^n$ form, one obtains the following expression for $\rho_\pm$ (which depends on q, s and the rates):

$$\rho_\pm=\frac{1}{2\mu pm}\Big(\lambda+\mu pm-\mu(1-p)(q-1)m-s \\ \pm\sqrt{(\lambda+\mu pm-\mu(1-p)(q-1)m-s)^2-4\lambda\mu pm}\Big). \tag{72}$$

The boundary condition $P_{s;q}(N+m)=0$ translates into the following condition (dependent on $P_{s;q}(m)$ and $P_{s;q}(m-1)$):

$$P_{s;q}(N+m)=\frac{(P_{s;q}(m)-\rho_-P_{s;q}(m-1))\rho_+^N+(P_{s;q}(m)-\rho_+P_{s;q}(m-1))\rho_-^N}{\rho_+-\rho_-}=0. \tag{73}$$

The above has different types of solutions dependent on the values of s with respect to the following two threshold values:

$$s_\pm=\lambda+\mu pm+\mu(1-p)(q-1)m \\ \pm\sqrt{4\lambda\mu pm}=\left(\sqrt{\lambda}\pm\sqrt{\mu pm}\right)^2-\mu(1-p)(q-1)m. \tag{74}$$

At $s<s_-$, both eigenvalues $\rho_\pm$ are real and $\rho_+>\rho_-$. In this case, as we are interested in the $N\to\infty$ limit, one can simply ignore the ρ_-^N contributions in (73), thus leading to the following relation: $P_{s;q}(m)-\rho_-P_{s;q}(m-1)=0$. This then replaces the last condition in (71), w.r.t. describing $P_{s;q}(n)$ at $0\le n\le m$. We were not able to solve this reduced system of equations analytically at any values of m, and thus to find the spectrum in its full glory. However, the information just provided is already sufficient for a heuristic derivation of the lowest eigenvalue, and the value of the gap between it and the continuous band.

We focus on a particular single form solution of the reduced system of equations for $P_{s;q}(n)$ with $0\le n\le m$, corresponding to $s=\Delta$ and $P_{s;q}(n)=h^n/n!$ for $n\le m$, with h and Δ from (62). In this case, one has $\rho_+=\lambda/(\mu ph)$ and $\rho_-=h/m$, so this is indeed the ground state solution. However, let us note for the sake of accurateness that we have not formally proven that this special solution always corresponds to the lowest eigenvalue (the ground state). However, this is exactly what we observed in our numerical experiments with different values of m.

Equation (73) also allows the continuous spectrum of the operator to be identified. At $s_-<s<s_+$ we have $|\rho_+|=|\rho_-|$, and one has to keep both terms in (73). As long as $|P_{s;q}(m+1)-\rho_-P_{s;q}(m)|=|P_{s;q}(m)-\rho_+P_{s;q}(m-1)|$, (73) has at least N solutions. At $N\gg 1$ this region turns into the continuous band of the spectrum. We conjecture that the system of equations does not have any solutions for $s>s_+$. (Once again, this is confirmed in simulations but we do not have an explicit way of proving it.)

Finally, we conclude that the transition between the uncongested regime and congested regime takes place at $\Delta=s_-$ which corresponds exactly to $q=q_c$ (from (66)). As one can easily see, the value of h at this point is equal to $h=\sqrt{\lambda/\mu p}<1$, so the queue length in the final moment does not diverge.

5 Conclusions and Path Forward

Let us briefly recall the highlights of this manuscript. Our analysis was focused on the generating function of currents over a Jackson (queueing) network, with an eye towards analyzing how large currents accumulate over time. We began by giving some relevant background in Queueing Theory, and discussing some tie-ins with recent work in statistical physics. We then adopted the Doi-Peliti technique for describing the dynamics of the Jackson network. We used this formalism to show that the ground state of the respective evolution operator has a well-defined and analytically tractable product/coherent state form in a particular regime. These results were translated into an implicit analytical expression for the Crámer function of currents in this "uncongested" regime, where the ground state was well separated by a gap from the continuous spectrum. We also observed that crossing the surface in the phase space of observed currents, where the spectral gap collapses, corresponds to the congestion of a node in the network. We suggested a heuristic graph-reduction scheme which allows the statistics of currents to be described in this partially congested setting. Finally, we validated the general results using the example of a single node feedback system, where many of our assumptions could be verified directly by performing an explicit analysis of the evolution operator spectrum.

We consider this study more like an opening for further exciting research along the following lines:

- Following the discussion in Sect. 3.3, we naturally conjecture that in a large network, gradually increasing the observed current(s) will lead to a transition from the uncongested to the fully congested regime via a number of steps, each characterized by an increase in the number of congested nodes. It would be interesting and important to explore the specific sequence of phase transitions separating the space of observed currents into cells. A particularly interesting question concerns the algorithmic complexity of identifying these cells and exploring their geometric structure (e.g., possible convexity).
- We have considered a queuing model with the transition rates independent of the node occupation numbers. This condition can be relaxed and such an extension of our theory should allow for non-uniform dependence of the rates on the occupation numbers.
- Extensions of the current-statistics theory to the so-called multi-class networks, where different classes of particles are treated differently at the stations (for example having different priorities) are less trivial, yet we anticipate such a generalization is still possible.
- In this manuscript we have considered solely open queuing networks. It would be interesting and instructive to generalize the theory of current statistics to the case of closed (particles are not leaving or entering the network) and semi-open (particles are injected into the system and leaving it, but in such a way that the total number of particles is conserved) networks.
- We may also consider networks of fixed structure, yet with rates changing in time, for example in a periodic fashion. To describe statistics of currents in this case, and especially in the regime where the typical correlation time of the rate changes are comparable to the inverse rates, constitutes another interesting future challenge. (Note that an approach blending the techniques described in this manuscript with the ones discussed recently in the context of the so-called Jarzynski equality [72] and work relations [73, 74], both closely related to the subject of fluctuation theorems [10–12], may prove fruitful for this task.)
- The assumption of infinite waiting room was crucially important for advancing the product/coherent state decomposition to the statistics of currents. Analyzing the current statistics in the regimes when all or some waiting rooms are of a finite capacity is yet another challenging task, as this finiteness brings in a new type of inter-particle interaction.

- It would be an interesting challenge to put the ideas presented in this paper on a more rigorous mathematical foundation. This would aid greatly in understanding the formal relationship between the regimes identified in this paper and the sample path large deviations properties of queueing networks. We note that the authors are currently undertaking preliminary work along these lines, with the work directed more towards the Queueing Theory community.
- Our analysis suggests that in large queuing networks one may expect emergence of multiple transitions as one increases the current. More generally, and in the spirit of [75, 76], it would be interesting to explore the field of the so-called qualitative queuing network theory via the methods/techniques discussed in this manuscript.
- Finally, all of the above should be used not only to study existing (man- or nature- made) networks, but also to guide construction of future technological networks with desired properties. In other words, we suggest to use this analysis for control and optimization of networks in new areas such as power, and even more generally energy, distribution.

Acknowledgements We are thankful to David Gamarnik for consulting us on many issues related to Queuing Theory, and Sergey Foss, Bill Massey and Alexander Rybko for enlightening conversations. This material is based upon work supported by the National Science Foundation under CHE-0808910 (VC) and CCF-0829945 (MC via NMC). The work at LANL was carried out under the auspices of the National Nuclear Security Administration of the U.S. DoE at LANL under Contract No. DE-AC52-06NA25396. KT acknowledges support of an Oppenheimer Fellowship at LANL, and DAG work on the project was a part of his summer internship (GRA program) at LANL.

Appendix: Left Ground State of the Hamiltonian

In this Appendix we construct the left ground state of the Hamiltonian (49). We start by recalling that according to (7), the bra-vector $\langle \mathbf{0}| \exp(\sum_i \hat{a}_i)$ is the left zero eigen-function of the Hamiltonian (49) at $\boldsymbol{q} = \mathbf{1}$. However, it ceases to be an eigen-vector at $\boldsymbol{q} \neq \mathbf{1}$. On the other hand, it is easy to check that the "exponential" bra-vector $\langle 0| \exp(\bar{h}\hat{a})$ is in fact a left eigen-vector of the creation operator $\hat{a}^+$, with the eigen-value $\overline{h}$, $\langle 0| \exp(\overline{h}\hat{a})\hat{a}^+ = \overline{h}\langle 0| \exp(\bar{h}\hat{a})$. This suggests searching for a left eigen-vector of the Hamiltonian (49), in the exponential form:

$$\langle \boldsymbol{s}_{\boldsymbol{q}}| = \langle \mathbf{0}| \exp\left(\sum_i \bar{h}_i \hat{a}_i\right). \tag{75}$$

Then, from $\langle \boldsymbol{s}_{\boldsymbol{q}}|\hat{H}_{\boldsymbol{q}} = -\bar{\Delta}(\boldsymbol{q})\langle \boldsymbol{s}_{\boldsymbol{q}}|$ and (49) (combined with utilizing the aforementioned feature of the creation operators) one derives the following set of conditions on the $\bar{\boldsymbol{h}}$ vector (of c-numbers):

$$\sum_i (q_{0i} - \bar{h}_i(\boldsymbol{q}))\lambda_{0i} = \bar{\Delta}(\boldsymbol{q}), \tag{76}$$

$$\forall i\colon \quad \lambda_{i0}\left(q_{i0} - \bar{h}_i(\boldsymbol{q})\right) + \sum_j^{(i,j)\in\mathcal{G}_1} \lambda_{ij}(q_{ij}\bar{h}_j(\boldsymbol{q}) - \bar{h}_i(\boldsymbol{q})) = 0. \tag{77}$$

It is straightforward to verify that the resulting $\bar{\Delta}(\boldsymbol{q})$ and $\Delta(\boldsymbol{q})$ from (54) are identical, $\bar{\Delta}(\boldsymbol{q}) = \Delta(\boldsymbol{q})$.

References

1. Jackson, J.R.: Manag. Sci. **10**, 131 (1963). http://www.jstor.org/stable/2627213
2. Spitzer, F.: Adv. Math. **5**, 246 (1970)
3. Kelly, F.P.: Adv. Appl. Probab. **8**, 416 (1976)
4. Nelson, R.: ACM Comput. Surv. **25**, 339 (1993)
5. Zeitak, R.: Dynamics of Jackson networks: perturbation theory (2007). http://arxiv.org/abs/0708.1718
6. Derrida, B., Domany, E., Mukamel, D.: J. Stat. Phys. **69**, 667 (1992)
7. Derrida, B., Lebowitz, J.L.: Phys. Rev. Lett. **80**, 209 (1998)
8. Derrida, B.: J. Stat. Mech., Theory Exp. **2007**, P07023 (2007). http://stacks.iop.org/1742-5468/2007/P07023
9. Blythe, R.A., Evans, M.R.: J. Phys. A, Math. Gen. **40**, 333 (2007). arXiv:0706.1678
10. Gallavotti, G., Cohen, E.: J. Stat. Phys. **80**, 931 (1995). http://dx.doi.org/10.1007/BF02179860
11. Kurchan, J.: J. Phys. A., Math. Gen. **31**, 3719 (1998). http://stacks.iop.org/0305-4470/31/3719
12. Lebowitz, J.L., Spohn, H.: J. Stat. Phys. **95**, 333 (1999). http://www.springerlink.com/content/u34v2j413047x642
13. Turitsyn, K., Chertkov, M., Chernyak, V.Y., Puliafito, A.: Phys. Rev. Lett. **98**, 180603 (2007), 4 pp. http://link.aps.org/abstract/PRL/v98/e180603
14. Chernyak, V.Y., Chertkov, M., Malinin, S.V., Teodorescu, R.: J. Stat. Phys. **137**, 109 (2009). http://www.springerlink.com/content/17740348qh5j03m7
15. Chen, H., Yao, D.: Fundamentals of Queuing Networks. Springer, Berlin (2001)
16. Chen, H., Mandelbaum, A.: Math. Oper. Res. **16**, 408 (1991)
17. Dai, J.: Ann. Appl. Probab. **5**, 49 (1995)
18. Rybko, A., Stolyar, A.: Probl. Peredachi Inf. **28**, 3 (1992)
19. Atar, R., Dupuis, P.: Stoch. Process. Appl. **84**, 255 (1999)
20. Anantharam, V.: IBM Research Report (1990)
21. Ignatiouk-Robert, I.: Ann. Appl. Probab. **10**, 962 (2000). http://www.jstor.org/stable/2667326
22. Majewski, K., Ramanan, K.: Preprint (2008). http://www.math.cmu.edu/users/kramanan/research/Jackson.pdf
23. Puhalskii, A.: Markov Process. Relat. Fields **13**, 99 (2007). http://www-math.cudenver.edu/~puhalski/publications/jackson_appeared.pdf
24. Merhav, N., Kafri, Y.: J. Stat. Mech., Theory Exp. P02011 (2010)
25. Rakos, A., Harris, R.: J. Stat. Mech., Theory Exp. P05005 (2008)
26. Harris, R., Rakos, A., Schutz, G.: Europhys. Lett. **75**, 227 (2006)
27. Harris, R., Rakos, A., Schutz, G.: J. Stat. Mech., Theory Exp. P08003 (2005)
28. J. Stat. Phys. **123**, 237 (2006). http://www.springerlink.com/content/g422mw36t15782k3
29. Bertini, L., Sole, A.D., Gabrielli, D., Jona-Lasinio, G., Landim, C.: J. Stat. Phys. **107**, 635 (2002). http://www.springerlink.com/content/lcqe21fx62dd71jm/
30. Srinivasan, R.: Math. Oper. Res. 39–50 (1993)
31. Malyshev, V., Yakolev, A.: Ann. Appl. Probab. **6**, 92 (1996)
32. Stolyar, A.: Tech. Rep., Bell Labs Laboratory (2009)
33. Kelly, F.: Reversibility and Stochastic Networks. Wiley, New York (1979)
34. Burke, P.: Oper. Res. **4**, 699 (1956)
35. Beutler, F., Melamud, B.: Oper. Res. **26**, 1059 (1956)
36. Pujolle, G., Soula, C.: In: Proc. 4th International Symposium on Modelling and Performance Evaluation of Computer Systems (1979)
37. Labetoulle, J., Pujolle, G., Soula, C.: Math. Oper. Res. **6**, 173 (1981). http://www.jstor.org/stable/3689132
38. Walrand, J., Varaiya, P.: Math. Oper. Res. **6**, 387 (1981)
39. Burke, P.: IEEE Trans. Commun. **24**, 575 (1976)
40. Bremaud, P.: Z. Wahrscheinlichkeitsth. **45**, 21 (1978)
41. Beutler, F.J., Melamed, B.: Oper. Res. **26**, 1059 (1978). http://www.jstor.org/stable/170265
42. Takacs, L.: Bell Syst. Tech. J. **42**, 505 (1963)
43. Pekoz, E., Joglekar, N.: J. Appl. Probab. **39**, 630 (2002)
44. Disney, R.L., Kiesler, P., Wortman, M.: Queueing Syst. **9**, 353–363 (1991)
45. Disney, R., McNickle, D., Simon, B.: Nav. Res. Logist. Q. **27**, 635 (1980)
46. D'Avignon, G.R., Disney, R.L.: Manag. Sci. **24**, 168180 (1977)
47. D'Avignon, G.R., Disney, R.L.: Tech. Rep. 75-9 (1975)
48. Brown, T., Fackrell, M., Xia, A.: Cosmos **1**, 47 (2005)
49. Brown, T., Weinberg, G., Xia, A.: Stoch. Process. Appl. **87**, 149 (2000)
50. Barbour, A.D., Brown, T.C.: J. Appl. Probab. **33**, 472 (1996). http://www.jstor.org/stable/3215072

51. Kyprianov, E.K.: On the quasi-stationary distributions of the GI/M/1 queue. J. Appl. Probab. **9**(1), 117–128 (1972)
52. Kao, P.: Limiting diffusion for random walks with drift conditioned to stay positive. J. Appl. Probab. **15**(2), 280–291 (1978)
53. Doi, M.: J. Phys. A, Math. Gen. **9**, 1465 (1976). http://stacks.iop.org/0305-4470/9/1465
54. Peliti, L.: J. Phys. Fr. **46**, 1469 (1985). http://dx.doi.org/10.1051/jphys:019850046090146900
55. Peliti, L.: J. Phys. A, Math. Gen. **19**, L365 (1986). http://stacks.iop.org/0305-4470/19/L365
56. Massey, W.A.: Adv. Appl. Probab. **16**, 176 (1984). http://www.jstor.org/stable/1427230
57. Massey, W.A.: J. Appl. Probab. **21**, 379 (1984). http://www.jstor.org/stable/3213647
58. Rozanov, Y.: Processes Aleatories. Mir, Moscow (1975)
59. Gardiner, C.W.: Handbook of Stochastic Methods. Springer, Heidelberg (1983)
60. Melamed, B., Whitt, W.: J. Appl. Probab. **27**, 376 (1990). http://www.jstor.org/stable/3214656
61. Beutler, F.J., Melamed, B.: Adv. Appl. Probab. **9**, 215 (1977). http://www.jstor.org/stable/1426358
62. Mcdonald, D.: Ann. Appl. Probab. **9**, 110 (1999)
63. Adan, I., Foley, R., Mcdonald, D.: Queueing Syst. **62**, 311 (2009)
64. Dai, J.G., Nguyen, V., Reiman, M.I.: Oper. Res. **42**, 119 (1994). http://www.jstor.org/stable/171530
65. Eun, D., Shroff, N.: IEEE/ACM Trans. Netw. **13**, 526 (2005)
66. Eun, D., Shroff, N.: Adv. Appl. Probab. **36**, 893 (2004)
67. Veciana, G.D., Courcoubetis, C., Walrand, J.: In: Proc. IEEE INFOCOM 2003, pp. 466–473 (1994)
68. Bertsimas, D., Paschalidis, I.C., Tsitsiklis, J.N.: Ann. Appl. Probab. **8**, 1027 (1998). http://www.jstor.org/stable/2667173
69. Chernyak, V.Y., Chertkov, M., Malinin, S.V., Teodorescu, R.: Non-equilibrium thermodynamics for functionals of current and density (2007). http://arxiv.org/abs/0712.3542
70. Wischik, D.: Queueing Syst. **32**, 383 (1999)
71. Wischik, D.: Ann. Appl. Probab. **11**, 379 (2000)
72. Jarzynski, C.: Phys. Rev. E **56**, 5018 (1997)
73. Crooks, G.E.: Phys. Rev. E **60**, 2721 (1999)
74. Chernyak, V., Chertkov, M., Jarzynski, C.: Phys. Rev. E **71**, 025102 (2005)
75. Kel'bert, M.Y., Kontsevich, M.A., Rybko, A.N.: Theory Probab. Appl. 379–382 (1989)
76. Dobrushin, R.L., Kelbert, M.Y., Rybko, A.N., Sukhov, Y.M.: In: Dobrushin, R.L., Kryukov, V.I., Toom, A.L. (eds.) Stochastic Cellular Systems: Ergodicity, Memory, Morphogenesis, pp. 183–224. Manchester Univ. Press, Manchester (1990)