Diffusion Parameters of Flows in Stable Queueing Networks

Yoni Nazarathy*, Werner Scheinhardt†

November 12, 2013

Abstract

We consider open multi-class queueing networks with general arrival processes, general processing time sequences and Bernoulli routing. The network is assumed to be operating under an arbitrary work-conserving scheduling policy that makes the system stable. An example is a generalized Jackson network with load less than unity and any work conserving policy. We find a simple diffusion limit for the inter-queue flows with an explicit computable expression for the covariance matrix. Specifically, we present a simple computable expression for the asymptotic variance of arrivals (or departures) of each of the individual queues and each of the flows in the network.

Keywords: Queueing Networks, Diffusion Limits, Asymptotic Variance.

1 Introduction

The study of explicit performance measures of stable queueing networks has been at the heart of applied probability and operations research for the past half century. Initial results such as Burke's Theorem [4], indicating that the output of a stationary M/M/1 queue is a Poisson process have motivated the study of queueing output processes with the aim of using the output characteristics of one queue as the input characteristics of a downstream queue. While landmark results such as the product form solution of Jackson networks (c.f. [17] or [19]) have given much hope and practical utility, in the 1960's and 1970's it was well understood that explicit exact queueing network decomposition is in general not attainable. See for example [9], [10] or [11] for classic surveys of queueing networks and their traffic processes.

The lack of explicit solutions in general cases as well as the inability to exactly decouple most networks has motivated the development and study of heuristic queueing network decomposition schemes such as the Queueing Network Analyzer (QNA) [27] (see also [22]), and many subsequent approximation methods (see for example the recent heuristics in [21]). The typical approximating assumption made in such schemes is that each queue in isolation is a G/G/1 queue which can be analysed independently of the other queues. The input process is then approximated by taking into consideration both exogenous arrivals and departures from other upstream queues.

*School of Mathematics and Physics, The University of Queensland, Brisbane, Australia.
†Department of Applied Mathematics, University of Twente, Enschede, The Netherlands.

Some of the key ingredients needed for a network decomposition (such as QNA) are based on

$$\nu_k := \lim_{t\to\infty} \frac{\mathbb{E}[E_k(t)]}{t}, \quad \text{and} \quad \sigma_k^2 = \lim_{t\to\infty} \frac{\mathrm{Var}\Big(E_k(t)\Big)}{t},$$

where $E_k(t)$ is the arrival counting process into queue k:

$$E_k(t) := A_k(t) + \sum_i D_{i,k}(t),$$

with $A_k(t)$ representing the exogenous arrival counting process to that queue and $D_{i,j}(t)$ the number of items that have departed from queue i and immediately arrived to queue j during the time interval $[0,t]$. We refer to the latter counting process as *flow* $i \to j$. The summation in $E_k(t)$ is over all flows $i \to k$.

Finding ν_k exactly is typically a trivial matter based on the network routing matrix and exogenous arrival rates. As opposed to that, σ_k^2 is more complex. In fact, all of the proposed methods to date, only approximate σ_k^2 heuristically by taking into consideration the variability of service times in immediate upstream queues, but most often do not consider dependencies between flows and even when these are considered, the methods are still heuristic, c.f. [21] and references there-in.

Our key contribution in this paper is a simple exact computable expressions for σ_k^2 as well as, related asymptotic covariance terms:

$$\sigma_{i,j} := \lim_{t\to\infty} \frac{\mathrm{Cov}\Big(E_i(t), E_j(t)\Big)}{t}, \tag{1}$$

and the asymptotic variability parameters of flows:

$$\sigma_{i\to j}^2 := \lim_{t\to\infty} \frac{\mathrm{Var}\Big(D_{i,j}(t)\Big)}{t}, \quad \sigma_{i_1\to j_1, i_2\to j_2} := \lim_{t\to\infty} \frac{\mathrm{Cov}\Big(D_{i_1,j_1}(t), D_{i_2,j_2}(t)\Big)}{t}. \tag{2}$$

Our formulas hold for a wide class of stable networks including multi-class queueing networks (of which generalized Jackson networks are a special case). These results may be invaluable in refining queueing network decomposition schemes since the underlying assumption in most schemes is that of renewal inter-queue flows and in this case the squared coefficient of variation (SCV) of the inter-renewal times, c^2, is the asymptotic variance divided by the mean squared. Our results thus guarantee finding an exact SCV in cases where the renewal assumption is valid.

Besides their possible (futuristic) applicability to network decomposition schemes (perhaps being involved in heuristics that also incorporate other characteristics), our current contribution allows to understand the correlation structure between flows. As we demonstrate in an example, there are situations where the sign of the correlation is influenced by the variability of arrival processes and we are able to determine this explicitly.

Our main result, Theorem 1, is formulated as a simple functional central limit theorem (FCLT) and assumes that the network processes satisfy a functional strong law of large numbers (FSLL) and the primitive processes satisfy FCLTs. In dealing with a stable queueing network, this could be viewed as a "fundamental" diffusion limit result similar to some of the results summarized in [7]. To the best of our knowledge this result has been

overlooked by previous authors working on basic diffusion limits of queueing networks. This is probably due to the fact that most of the exciting research in the field of diffusion approximations of queueing networks in the past three decades, has focused on critically loaded networks (c.f. [3], [8], [26] [29], as well as many other key references summarized in [7], [12] [23] and [28]). The seminal paper [6], does consider diffusion approximations for queueing networks in all regimes (under-loaded, balanced and over-loaded), yet the inter-queue flows are not considered in that paper.

As described in our main diffusion result, the asymptotic variability of flows is driven by two components: (i) The variability of the arrival flows. (ii) The variability resulting from the Bernoulli routing. In stable networks, the variability of queue sizes (related also to service time distributions) does not play a role.

Since asymptotic variability of flows only depends on the interplay of the arrival process variability and the Bernoulli routing, we are also motivated to present an alternative way for quantifying the asymptotic variability parameters: *networks with zero service times*. In such networks, jobs that arrive to the network traverse instantaneously through the classes/queues until they depart, and hence the total count of jobs passing on flow $i \to j$ is

$$\sum_k \sum_{\ell=1}^{A_k(t)} N_{i,j|k}(\ell),$$

where the outer sum is over all queues and $N_{i,j|k}(\ell)$ are counts of the number of passes on $i \to j$ for the ℓ'th job arriving exogenously to k. Using elementary calculations, we find the asymptotic variances and covariances of such processes and prove they are the same as those originating from the diffusion parameters. Relating the diffusion limit parameters to the actual asymptotic variability values of the flows requires uniform integrability (UI). In addition to the diffusion limits, we establish this UI. It is the zero service time view which allows us to establish UI and related the diffusion parameters to asymptotic variability parameters.

The structure of the sequel is as follows: in Section 2 we summarize our results in a main theorem together with the notation and assumptions of the model. The following three sections constitute the proof. In Section 3 we present the calculation of the diffusion parameters and diffusion limit. In Section 4 we present the alternative view of the network based on zero service times. In Section 5 we relate the diffusion parameters to asymptotic variance and establish the required UI. We then follow with Section 6 where we present a numerical example. Readers are encouraged to read this section in conjunction with Section 2. Closing remarks are in Section 7.

2 Model and Main Result

We consider open networks subject to Bernoulli routing. These can either be the generalized Jackson queueing networks as described in [7] or more generally, open multi-class queueing networks (MCQN) operating under an arbitrary policy. See for example [2] for an extensive overview of the various models. We now outline the notation and assumptions of our network, followed by the main result.

Denote the classes/queues as $k = 1, \ldots, K$ and use the index 0 to denote the outside world. Let $T_k(t)$ denote the work (in units of time) allocated towards serving class k during

the time interval $[0,t]$. Further assume counting processes, $A_k(t)$ and $S_k(t)$ representing the number of exogenous arrivals to class k during $[0,t]$ and the number of jobs served during uninterrupted service in class k during $[0,t]$ respectively. The actual number of jobs served during $[0,t]$ is $S_k\big(T_k(t)\big)$. For $i=1,\ldots,K$ and $j=0,\ldots,K$, let $\Phi_{i,j}(\ell)$ denote the number of items routed from class i to class j out of the first ℓ items served at i, with,

$$\sum_{j=0}^{K}\Phi_{i,j}(\ell)=\ell,\quad i=1,\ldots,K. \tag{3}$$

Our *inter-queue flows* are the following counting processes,

$$D_{i,j}(t)=\Phi_{i,j}\Big(S_i\big(T_i(t)\big)\Big),\quad i=1,\ldots,K,\ j=0,\ldots,K. \tag{4}$$

Let $Q_k(t)$ denote the number of items in the queue of class k at time t. We assume $Q_k(0)=0$ and have,

$$Q_k(t)=E_k(t)-\sum_{j=0}^{K}D_{k,j}(t), \tag{5}$$

where the (total) arrival process to queue k is,

$$E_k(t)=A_k(t)+\sum_{i=1}^{K}D_{i,k}(t). \tag{6}$$

In the treatment below, the vectors Q,T,A,E and S (and their "bar", "hat" and "tilde" versions as defined below) are treated as K-dimensional column vectors. Further, let Φ and D be K^2 dimensional column vectors with the elements ordered in lexicographic order with the elements $D_{k,0}$ omitted. For example,

$$D=\Big(D_{1,1},\ldots,D_{1,K},D_{2,1},\ldots,D_{2,K},\ldots\ldots\ldots\ldots,D_{K,1},\ldots,D_{K,K}\Big)'.$$

Scaling Limits

For $n=1,2,\ldots$ and a function $U(t)$, denote $\bar{U}^n(t)=U(nt)/n$. We say that a fluid limit of U exists if $\lim_{n\to\infty}\bar{U}^n(t)=\bar{U}(t)$ exists uniformly on compact sets (u.o.c) almost surely. Further, when the limit $\bar{U}(t)$ exists, denote,

$$\hat{U}^n(t)=\frac{U(nt)-\bar{U}(nt)}{\sqrt{n}},\quad n=1,2,\ldots.$$

In cases where the above sequence converges weakly on Skorohod J_1 topology to a limiting process, $\hat{U}(t)$, we denote,

$$\hat{U}^n\Rightarrow\hat{U}.$$

For discrete time processes replace $U(nt)$ by $U(\lfloor nt\rfloor)$. See [7], Ch. 5 for brief background of weak convergence in the context of queueing networks. An extensive treatment is in [28].

Using these scaling definitions and assuming that the fluid limits exist, equations (3), (5) and (6) are easily manipulated to yield for all n,

$$0 = \sum_{j=0}^{K} \hat{\Phi}_{i,j}^n(\ell), \quad \ell = 1, 2, \ldots, \tag{7}$$

$$\hat{Q}_k^n(t) = \hat{A}_k^n(t) + \sum_{j=1}^{K} \hat{D}_{j,k}^n(t) - \sum_{j=0}^{K} \hat{D}_{k,j}^n(t), \quad t \geq 0, \tag{8}$$

$$\hat{E}_k^n(t) = \hat{A}_k^n(t) + \sum_{i=1}^{K} \hat{D}_{i,k}^n(t), \quad t \geq 0. \tag{9}$$

Probabilistic Assumptions

The *primitive processes* of our network model are $A(t)$, $S(t)$ and $\Phi(\ell)$. By this we mean that these processes are used to construct the probability space on which further network processes are defined. For simplicity we assume that $A_k(t)$, $S_k(t)$ and $\Phi_{k,\cdot}(\ell)$ are independent processes. This can be easily relaxed to allow correlations between different arrivals, services and routing but we do not do so here.

We assume that the primitive processes satisfy a functional strong law of large numbers (FSLL) yielding fluid limits,

$$\bar{A}_i(t) = \alpha_i t, \quad \bar{S}_i(t) = \mu_i t, \quad \bar{\Phi}_{i,j}(\ell) = p_{i,j}\ell,$$

with $\alpha_i > 0$, $\mu_i > 0$, $p_{i,j} \geq 0$, and $p_{i,0} = (1 - \sum_{j=1}^{N} p_{i,j}) \geq 0$. We denote by P the $K \times K$ matrix of the $p_{i,j}$, $i, j = 1, \ldots, K$. We assume throughout that P has spectral radius less than 1 so that $I - P'$ is non-singular and the network is open. Denote $\nu = (I - P')^{-1}\alpha$ and $\nu_{i,j} := \nu_i\, p_{i,j}$.

We also assume that the primitive processes satisfy functional central limit theorems (FCLT) laws. Specifically we assume $\hat{A}_i(t)$ are Brownian motions with diffusion coefficients $v_i \geq 0$, and that,

$$\hat{\Phi}_{k,\cdot}(t) = \left(\hat{\Phi}_{k,1}(t), \ldots, \hat{\Phi}_{k,K}(t)\right)', \quad k = 1, \ldots, K,$$

are K-dimensional Brownian motions with covariance matrices Γ_k, having the i, j'th entry $p_{k,i}(\delta_{i,j} - p_{k,j})$, where $\delta_{i,j}$ is the Kronecker delta.

In addition to the FCLT laws we assume that the squares of the diffusion scalings of the arrival processes are uniformly integrable (UI). Namely we assume that for $i = 1, \ldots, K$ and some $t_0 > 0$,

$$\left\{\frac{(A_i(t) - \alpha_i t)^2}{t},\, t \geq t_0\right\} \quad \text{is UI.} \tag{10}$$

Note that in the case of renewal arrival processes this UI property holds (c.f. [15]). The assumptions on $\bar{A}_i(t)$ and $\hat{A}_i(t)$ above are then satisfied when the inter-arrival times have finite second moments $\alpha_i^{-3}v_i^2 + \alpha_i^{-2}$. In the case of Bernoulli routing, the existence of $\bar{\Phi}(\cdot)$ and $\hat{\Phi}(\cdot)$ as specified above follows from the standard FSSLN and FCLT.

As it turns out, the service processes do not play a role in our limiting results and thus we do not impose FCLT or UI assumptions on S.

Stability Assumptions

The evolution of the system depends on a scheduling policy through $T(t)$. Different variations of this model imply different restrictions on $T(t)$ (single-class, multi-class, pre-emptive, non-preemptive, etc...). We assume the network is operated by a policy such that the resulting processes exhibit the following three assumptions:

(A1) Fluid limits for work allocations exist and satisfy: $\bar{T}_k(t) = \frac{\nu_k}{\mu_k} t$.

(A2) Queues are stable in the sense that $\hat{Q}^n \Rightarrow 0$.

(A3) Queue moment assumption: $\mathbb{E}[\big(Q_k(t)\big)^2] = o(t)$ as $t \to \infty$.

Such policies are known to exist for a variety of models and variations. Extensive treatment is in [2]. Specifically in the single class cases, also known as generalized Jackson networks, (c.f. [7], Ch 7. and references there-in), any work conserving policy achieves (A1) and (A2) if we assume that for all queues, $\nu_k < \mu_k$. In multi-class networks, the necessary stability condition is,

$$\sum_{k \in \mathcal{C}_i} \frac{\nu_k}{\mu_k} < 1, \quad \text{for every server } i.$$

Here $\mathcal{C}_i$ is the set of queues served by server i. Although not every work conserving policy is stable, it is know that stable policies exist. See [2] for an extensive treatment of this subject.

The final assumption (A3) is needed to establish UI. In its own right, it is not easily established for arbitrary models and policies. Nevertheless, it is a very sensible assumption. In fact, for many stable networks it holds that $\mathbb{E}[\big(Q_k(t)\big)^2] < C$ for some $C < \infty$ for all t.

Observe that, assumption (A1) implies,

$$\lim_{n \to \infty} \bar{D}^n_{i,j}(t) := \bar{D}_{i,j}(t) = \bar{\Phi}_{i,j}(\bar{S}_i(\bar{T}_i(t))) = p_{i,j}\nu_i t, \text{ u.o.c..} \tag{11}$$

Main Result

We now set up some matrices and vectors used in our main theorem. Use $\mathbf{1}$ to denote the vector of ones and define the $K \times K^2$ matrix $B := \mathbf{1}' \otimes I$ where $\otimes$ is the Kronecker product. Further denote the $K^2 \times K$ matrix,

$$P_c := \begin{bmatrix} P' e_{1,1} \\ P' e_{2,2} \\ \vdots \\ P' e_{K,K} \end{bmatrix},$$

where $e_{i,j}$ is a $K \times K$ matrix with all entry's 0 except for the i,j'th entry being 1. Now define the $K^2 \times (K + K^2)$ matrix H as,

$$H := \begin{bmatrix} P_c(I - P')^{-1} & I_{K^2} + P_c(I - P')^{-1}B \end{bmatrix},$$

as well as the $(K+K^2)\times(K+K^2)$ covariance matrix for the exogenous arrival processes and the routing processes,

$$\Sigma^{(P)} := \begin{bmatrix} \mathrm{diag}(v_k^2) & & & 0 \\ & \nu_1\Gamma_1 & & \\ & & \ddots & \\ 0 & & & \nu_K\Gamma_K \end{bmatrix},$$

where $\mathrm{diag}(v_k^2)$ is a diagonal matrix with elements v_k^2. Further, for any $i,j\in\{1,\ldots,K\}$ define the K dimensional vector $m(i,j)$ as follows:

$$m(i,j) := (I-P)^{-1}e_{i,i}P_{\cdot,j}, \tag{12}$$

where $P_{\cdot,j}$ is the j'th column of P. As further elaborated on in Section 4, the k'th entry of the column vector $m(i,j)$ is the expected number of transitions from state i to state j in a Markov chain whose transient component is specified by P and initial state is set to k.

We now present our main result.

Theorem 1. *Consider queueing networks described by equations* (3)–(6) *operating under some well defined scheduling policy.*

(i) If assumptions (A1) and (A2) hold then the sequences $\hat{D}^n$ and $\hat{E}^n$ converge weakly to drift-less Brownian motion processes with covariance matrices,

$$\Sigma^{(D)} := H\,\Sigma^{(P)}\,H', \quad \textit{and} \quad \Sigma^{(E)} := \Big(BH+[I_K\ \ 0]\Big)\Sigma^{(P)}\Big(BH+[I_K\ \ 0]\Big)', \tag{13}$$

respectively.

(ii) If in addition to (A1) and (A2), assumption (A3) holds, then the asymptotic variability parameters, as defined in (1) *and* (2)*, can be read off from the diffusion parameters. Namely,*

$$\sigma_{i_1\to j_1,i_2\to j_2} = \Sigma^{(D)}_{(i_1-1)K+j_1,\ (i_2-1)K+j_2}, \qquad \sigma_{i,j} = \Sigma^{(E)}_{i,j}.$$

(iii) An alternative calculation for the asymptotic variability parameters that is valid if assumptions (A1)–(A3) hold is the following:

$$\begin{aligned} \sigma_{i_1\to j_1,i_2\to j_2} &= m_{j_1}(i_2,j_2)\alpha' m(i_1,j_1) + m_{j_2}(i_1,j_1)\alpha' m(i_2,j_2) \\ &\quad +(v^2-\alpha)'\big(m(i_1,j_1)\bullet m(i_2,j_2)\big), \end{aligned} \tag{14}$$

$$\sigma_{i\to j,i\to j} = (1+2m_j(i,j))\alpha' m(i,j) + (v^2-\alpha)'\big(m(i,j)\bullet m(i,j)\big), \tag{15}$$

$$\sigma_{j_1,j_2} = v_{j_1}^2\sum_{i_2=1}^{K} m_{j_1}(i_2,j_2) + v_{j_2}^2\sum_{i_1=1}^{K} m_{j_2}(i_1,j_1) + \sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\sigma_{i_1\to j_1,i_2\to j_2} \tag{16}$$

where $m_k(i,j)$ is the k'th entry of the vector $m(i,j)$ and $(x\bullet y)$ signifies the vector resulting from element-wise product of the vectors x and y.

Proof. The remainder of the paper establishes the proof. (i) is established in Section 3. (iii) is established in Section 4. (ii) relies on the development of (iii) and is established in Section 5. □

We demonstrate the applicability of our result on a specific network example in Section 6.

3 The Diffusion Parameters

Lemmas 1–4 below summarize straight forward algebraic manipulations of the network equations. These then lead to a simple diffusion limit that follows from Donsker's theorem (see [7], Ch. 5-7 or [12] for background). Techniques similar to those employed here are also in [25], applied to queueing networks that generate their own input. The basic idea is to represent the diffusion scaled processes, $\hat{D}^n$ and $\hat{T}^n$ in-terms of the following "tilde" processes,

$$\tilde{\Phi}^n_{i,j}(t) := \hat{\Phi}^n_{i,j}\Big(\bar{S}^n_i\big(\bar{T}^n_i(t)\big)\Big), \quad \text{and} \quad \tilde{S}^n_k(t) := \hat{S}^n_k(\bar{T}^n_k(t)),$$

which in-turn have diffusion limits based on the primitive processes.

Lemma 1. *For $i = 1, \ldots, K$ and $j = 0, \ldots, K$,*

$$\hat{D}^n_{i,j}(t) \;=\; \tilde{\Phi}^n_{i,j}(t) + p_{i,j}\tilde{S}^n_i(t) + p_{i,j}\mu_i\hat{T}^n_i(t). \tag{17}$$

Proof. Use, $D_{i,j}(nt) = \Phi_{i,j}(S_i(T_i(nt))) = \Phi_{i,j}(n\bar{S}^n_i(\bar{T}^n_i(t)))$ and (11) to get,

$$\begin{aligned}
\hat{D}^n_{i,j}(t) &= \frac{\Phi_{i,j}(n\bar{S}^n_i(\bar{T}_i{}^n(t))) - p_{i,j}\nu_i nt}{\sqrt{n}} \\
&= \frac{\Phi_{i,j}(n\bar{S}^n_i(\bar{T}_i{}^n(t))) - p_{i,j}n\bar{S}^n_i(\bar{T}^n_i(t))}{\sqrt{n}} + \frac{p_{i,j}n\bar{S}^n_i(\bar{T}^n_i(t)) - p_{i,j}\mu_i n\bar{T}^n_i(t)}{\sqrt{n}} \\
&\quad + \frac{p_{i,j}\mu_i n\bar{T}^n_i(t) - p_{i,j}\nu_i nt}{\sqrt{n}}.
\end{aligned}$$

Now, (17) follows. □

Denote by M the diagonal matrix with diagonal elements μ_k^{-1}. We now have,

Lemma 2. *The diffusion scaled time allocation can be written as:*

$$\hat{T}^n(t) = M(I - P')^{-1}\Big(\hat{A}^n(t) + B\tilde{\Phi}^n(t) - (I - P')\tilde{S}^n(t)\Big) - M(I - P')^{-1}\hat{Q}^n(t). \tag{18}$$

Proof. Substituting (17) into (8) we have:

$$\begin{aligned}
\hat{Q}^n_k(t) &= \hat{A}^n_k(t) + \sum_{j=1}^{K}\Big(\tilde{\Phi}^n_{j,k}(t) + p_{j,k}\tilde{S}^n_j(t) + p_{j,k}\mu_j\hat{T}^n_j(t)\Big) \\
&\quad - \sum_{j=0}^{K}\Big(\tilde{\Phi}^n_{k,j}(t) + p_{k,j}\tilde{S}^n_k(t) + p_{k,j}\mu_k\hat{T}^n_k(t)\Big) \\
&= \hat{A}^n_k(t) + \sum_{j=1}^{K}\Big(\tilde{\Phi}^n_{j,k}(t) + p_{j,k}\tilde{S}^n_j(t) + p_{j,k}\mu_j\hat{T}^n_j(t)\Big) - \tilde{S}^n_k(t) - \mu_k\hat{T}^n_k(t) \\
&= \hat{A}^n_k(t) + \sum_{j=1}^{K}\tilde{\Phi}^n_{j,k}(t) - \big(\tilde{S}^n_k(t) - \sum_{j=1}^{K}p_{j,k}\tilde{S}^n_j(t)\big) - \big(\mu_k\hat{T}^n_k(t) - \sum_{j=1}^{K}p_{j,k}\mu_k\hat{T}^n_j(t)\big),
\end{aligned}$$

where in the second step we used (7) and $\sum_{j=0}^{K} p_{i,j} = 1$. In vector/matrix form this reads:

$$\hat{Q}^n(t) \;=\; \hat{A}^n(t) + B\tilde{\Phi}^n(t) - (I - P')\tilde{S}^n(t) - (I - P')M^{-1}\hat{T}^n(t).$$

Now (18) follows by multiplying both sides by $M(I - P')^{-1}$. □

We now have,

Lemma 3.

$$\hat{D}^n(t) = \left[\begin{array}{cc} H & 0_{K\times K} \end{array}\right] \left[\begin{array}{c} \hat{A}^n(t) \\ \tilde{\Phi}^n(t) \\ \hat{S}^n(t) \end{array}\right] - P_c(I-P')^{-1}\hat{Q}^n(t).$$

Proof. Equations (17) are,

$$\hat{D}^n(t) = \tilde{\Phi}^n(t) + P_c\Big(\tilde{S}^n(t) + M^{-1}\hat{T}^n(t)\Big).$$

Substituting (18) in the above, $\tilde{S}^n(t)$ drops out of the equation, and we obtain,

$$\hat{D}^n(t) = \Big(I_{K^2} + P_c(I-P')^{-1}B\Big)\tilde{\Phi}^n(t) + P_c(I-P')^{-1}\hat{A}^n(t) - P_c(I-P')^{-1}\hat{Q}^n(t).$$

□

Observe from Lemma 3 that $\hat{D}^n$ depends on $\hat{S}^n$ only through the queue. We may now represent the analogous result for $\hat{E}^n$, this time omitting the primitive sequence $\hat{S}^n$ from the representation.

Lemma 4.

$$\hat{E}^n(t) = \Big(BH + [I_K \;\; 0]\Big) \left[\begin{array}{c} \hat{A}^n(t) \\ \tilde{\Phi}^n(t) \end{array}\right] - BP_c(I-P')^{-1}\hat{Q}^n(t),$$

Proof. We use (9) and the previous lemma:

$$\begin{aligned} \hat{E}^n(t) &= B\hat{D}^n(t) + \hat{A}^n(t) \\ &= BH \left[\begin{array}{c} \hat{A}^n(t) \\ \tilde{\Phi}^n(t) \end{array}\right] - BP_c(I-P')^{-1}\hat{Q}^n(t) + [I_K \;\; 0] \left[\begin{array}{c} \hat{A}^n(t) \\ \tilde{\Phi}^n(t) \end{array}\right]. \end{aligned}$$

□

We can now establish the diffusion limit in our main theorem:

Proof of Theorem 1, (i): The FCLT assumptions together with applications of the continuous mapping theorem and assumption (A1) imply that $\tilde{\Phi}^n_{k\cdot}(t)$ converge weakly to K-dimensional Brownian motions with covariance matrices $\mu_k \frac{\nu_k}{\mu_k}\Gamma_k$. Since by assumption $\hat{A}(t)$ is Brownian motion, independent of the routing, the covariance matrix of $\left[\begin{array}{c} \hat{A}^n(t) \\ \tilde{\Phi}^n(t) \end{array}\right]$ is $\Sigma^{(P)}$. The result then follows from the representation in Lemmas 3 and 4 and assumption (A2). □

We note that Lemma 2 above can also yield diffusion limits for rate allocations. This appears (7.89), pp.189 in [7]. In fact, there the authors handle a much wider case in which some queues may be critical and/or overloaded. This is originally from [6] (6.14), pg 1498. As stated at the introduction the diffusion limit for D and E did not appear in [6] and subsequent literature. It is insightful to know that we may also obtain joint diffusion limits for T and D or E, yet we do not peruse this here. Further, handling the case of overloaded queues does also not pose any additional technical difficulty. The case of critical queues is in general an open question. It was handled in [1] for the single station queue.

4 The Zero Service Time View

In this section we refer to the queues as *nodes* to make it clear that there is actually no queueing taking place. For the ℓ'th customer arriving exogenously first to node k, denote $N_{j|k}(\ell)$ as the number of times that the customer visits node j, and denote $N_{i,j|k}(\ell)$ as the number of times that the customer traverses on the flow $i \to j$. Thus, $N_{j|k}(\ell) = \sum_{i=1}^{K} N_{i,j|k}(\ell)$. Define now,

$$\breve{D}_{i,j}(t) := \sum_{k=1}^{K} \sum_{\ell=1}^{A_k(t)} N_{i,j|k}(\ell), \quad \text{and} \quad \breve{E}_k(t) := A_k(t) + \sum_{i=1}^{K} \breve{D}_{i,k}(t) = A_k(t) + \sum_{k'=1}^{K} \sum_{\ell=1}^{A_k(t)} N_{k|k'}(\ell).$$

The process $\breve{D}_{i,j}(t)$ is a count of the number of items passing from node i to node j up to time t as if service times are 0. In particular, the ℓ'th customer who arrives at node k by time t ($\ell = 1, \ldots, A_k(t)$) makes an "instantaneous tour" through the nodes, passing $N_{i,j|k}(\ell)$ times on the flow $i \to j$. Similarly, $\breve{E}_k(t)$ is the count of the number of jobs arriving to queue k either exogenously or passing through the network assuming that service times are 0.

By considering both $D(\cdot)$ and $\breve{D}(\cdot)$ on the same probability space, we have that *a.s.*,

$$D_{i,j}(t) \leq \breve{D}_{i,j}(t).$$

Denote now,

$$\breve{N}_{i,j}(t) := \breve{D}_{i,j}(t) - D_{i,j}(t).$$

This is the number of future passes on $i \to j$ by customers that are currently in the system (where service times are generally non-zero) at time t. It is obvious from the Markovian nature of the routing that,

$$\breve{N}_{i,j}(t) \quad =^d \quad \sum_{k=1}^{K} \sum_{\ell=1}^{Q_k(t)} N_{i,j|k}(\ell), \tag{19}$$

where the equality is in distribution and for given k,

$$\big\{ \big(N_{i,j|k}(\ell),\ i,j \in \{1,\ldots,K\}, i \neq j \big),\ \ell = 1,2,\ldots \big\},$$

is an i.i.d. sequence (of K^2 dimensional random vectors) whose distribution is induced by a discrete time Markov chain on state space $\{0,1,\ldots,K\}$ with transition matrix,

$$\tilde{P} = \begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{1} - P\mathbf{1} & P \end{bmatrix}.$$

To construct $N_{i,j|k}(\ell)$, denote by $\{X_n^k\}$ a sequence of states generated by the above Markov chain with $P(X_0 = k) = 1$ for $k \in \{1,\ldots,K\}$. Then for $i \neq j$, $N_{i,j|k}(\ell)$ is distributed as,

$$N_{i,j|k} \ := \ \sum_{n=1}^{\infty} \mathbf{1}\{X_{n-1}^k = i, X_n^k = j\},$$

and thus $N_{j|k}(\ell)$ is distributed as,

$$N_{j|k} \ := \sum_{n=0}^{\infty} \mathbf{1}\{X_n^k = j\}.$$

Since the queueing network is open (P is sub-stochastic), the only recurrent class in this Markov chain is $\{0\}$ and thus the random variables $N_{i,j|k}$ are proper. It is also a standard exercise to show that they have finite mean and variance.

Denote now,

$$\breve{\sigma}_{i,j} := \lim_{t\to\infty} \frac{\mathrm{Cov}\Big(\breve{E}_i(t), \breve{E}_j(t)\Big)}{t}, \quad \text{and} \quad \breve{\sigma}_{i_1\to j_1, i_2\to j_2} := \lim_{t\to\infty} \frac{\mathrm{Cov}\Big(\breve{D}_{i_1,j_1}(t), \breve{D}_{i_2,j_2}(t)\Big)}{t}.$$

As we show now, under assumption (A3), these variability parameters (of the zero-service time flows), are the same as the variability parameters of the system with queueing:

Proposition 1. *Assume assumption (A3) holds. Then,*

$$\breve{\sigma}_{i,j}^2 = \sigma_{i,j}^2, \quad \textit{and} \quad \breve{\sigma}_{i_1\to j_1, i_2\to j_2} = \sigma_{i_1\to j_1, i_2\to j_2}. \tag{20}$$

Proof. We present the proof for the asymptotic variability of D, the case of E is similar and is omitted. We have,

$$\begin{aligned}
&\Big|\mathrm{Cov}\Big(\breve{D}_{i_1,j_1}(t), \breve{D}_{i_2,j_2}(t)\Big) - \mathrm{Cov}\Big(D_{i_1,j_1}(t), D_{i_2,j_2}(t)\Big)\Big| \qquad (21)\\
\le\ &\Big|\mathrm{Cov}\Big(D_{i_1,j_1}(t), \breve{N}_{i_2,j_2}(t)\Big)\Big| + \Big|\mathrm{Cov}\Big(D_{i_2,j_2}(t), \breve{N}_{i_1,j_1}(t)\Big)\Big| + \Big|\mathrm{Cov}\Big(\breve{N}_{i_1,j_1}(t), \breve{N}_{i_2,j_2}(t)\Big)\Big|\\
\le\ &\sqrt{\mathrm{Var}\Big(D_{i_1,j_1}(t)\Big)\mathrm{Var}\Big(\breve{N}_{i_2,j_2}(t)\Big)} + \sqrt{\mathrm{Var}\Big(D_{i_2,j_2}(t)\Big)\mathrm{Var}\Big(\breve{N}_{i_1,j_1}(t)\Big)}\\
&+\sqrt{\mathrm{Var}\Big(\breve{N}_{i_1,j_1}(t)\Big)\mathrm{Var}\Big(\breve{N}_{i_2,j_2}(t)\Big)}.
\end{aligned}$$

For any (i,j) we have that both $\mathrm{Var}(D_{i,j}(t))/t$ and $\mathrm{Var}(\breve{N}_{i,j}(t))$ are bounded from above uniformly in t; for the latter this is a consequence of Assumption (A3). Dividing (21) by t and taking $t\to\infty$ we get the result. □

Note: a version of the above result also exists for the mean rates, ν. In this case all that is required is finiteness of the first moments of the queues.

We now express the components of $\breve{\sigma}$ in terms of $\mathbb{E}[N_{i,j|k}]$ and $\mathrm{Cov}(N_{i_1,j_1|k}, N_{i_2,j_2|k})$.

Proposition 2.

$$\begin{aligned}
\breve{\sigma}_{i_1\to j_1, i_2\to j_2} &= \sum_{k=1}^{K} \alpha_k Cov(N_{i_1,j_1|k}, N_{i_2,j_2|k}) + \sum_{k=1}^{K} v_k^2 \mathbb{E}[N_{i_1,j_1|k}]\mathbb{E}[N_{i_2,j_2|k}],\\
\breve{\sigma}_{j_1,j_2} &= v_{j_1}^2 \mathbb{E}[N_{j_2|j_1}] + v_{j_2}^2 \mathbb{E}[N_{j_1|j_2}]\\
&\quad + \sum_{k=1}^{K} \alpha_k Cov\big(N_{j_1|k}, N_{j_2|k}\big) \ + \ \sum_{k=1}^{K} v_k^2\, \mathbb{E}[N_{j_1|k}]\, \mathbb{E}[N_{j_2|k}]\\
&= v_{j_1}^2 \mathbb{E}[N_{j_2|j_1}] + v_{j_2}^2 \mathbb{E}[N_{j_1|j_2}] \ + \ \sum_{i_1=1}^{K}\sum_{i_2=1}^{K} \breve{\sigma}_{i_1\to j_1, i_2\to j_2}.
\end{aligned}$$

Proof. We begin with the asymptotic variability of $\breve{D}$, namely $\breve{\sigma}_{i_1\to j_1, i_2\to j_2}$. For illustration we begin with the variance (even though it is a special case of the covariance calculation that follows). Using the conditional variance rule we get,

$$\mathrm{Var}\Big(\breve{D}_{i,j}(t)\Big) = \sum_{k=1}^{K} \mathrm{Var}\Big(\sum_{\ell=1}^{A_k(t)} N_{i,j|k}(\ell)\Big) = \sum_{k=1}^{K} \Big(\mathbb{E}[A_k(t)]\,\mathrm{Var}(N_{i,j|k}) + \mathrm{Var}(A_k(t))\,\mathbb{E}[N_{i,j|k}]^2\Big).$$

Moving onto the covariance, observe that $N_{i_1,j_1|k}(\ell)$ and $N_{i_2,j_2|k'}(\ell)$ are independent whenever $k \neq k'$, hence,

$$\mathrm{Cov}\Big(\breve{D}_{i_1,j_1}(t), \breve{D}_{i_2,j_2}(t)\Big) = \sum_{k=1}^{K} \mathrm{Cov}\Big(\sum_{\ell=1}^{A_k(t)} N_{i_1,j_1|k}(\ell), \sum_{\ell=1}^{A_k(t)} N_{i_2,j_2|k}(\ell)\Big)$$

$$= \sum_{k=1}^{K} \Big(\mathbb{E}[A_k(t)]\,\mathrm{Cov}(N_{i_1,j_1|k}, N_{i_2,j_2|k}) + \mathrm{Var}\big(A_k(t)\big)\,\mathbb{E}[N_{i_1,j_1|k}]\mathbb{E}[N_{i_2,j_2|k}]\Big)$$

where in the second step we use the conditional covariance rule,

$$\mathrm{Cov}(X,Y) = \mathbb{E}\big[\mathrm{Cov}(X,Y|Z)\big] + \mathrm{Cov}\big(\mathbb{E}[X|Z], \mathbb{E}[Y|Z]\big).$$

Dividing by t and taking $t \to \infty$ yields the result.

Moving onto the asymptotic variability of $\breve{E}$ (this time treating the variance and the other covariance terms together) we expand and get:

$$\begin{aligned}\mathrm{Cov}\big(\breve{E}_{j_1}(t), \breve{E}_{j_2}(t)\big) &= \sum_{i_2=1}^{K} \mathrm{Cov}\big(A_{j_1}(t), \breve{D}_{i_2,j_2}(t)\big) + \sum_{i_1=1}^{K} \mathrm{Cov}\big(A_{j_2}(t), \breve{D}_{i_1,j_1}(t)\big) \\ &\quad + \sum_{i_1=1}^{K}\sum_{i_2=1}^{K} \mathrm{Cov}\big(\breve{D}_{i_1,j_1}(t), \breve{D}_{i_2,j_2}(t)\big) \qquad (22)\end{aligned}$$

To rewrite the first sum on the righthand side we can use

$$\begin{aligned}\mathrm{Cov}\big(A_{j_1}(t), \breve{D}_{i_2,j_2}(t)\big) &= \sum_{k=1}^{K} \mathrm{Cov}\big(A_{j_1}(t), \sum_{\ell=1}^{A_k(t)} N_{i_2,j_2|k}(\ell)\big) \\ &= \mathrm{Cov}\big(A_{j_1}(t), \sum_{\ell=1}^{A_{j_1}(t)} N_{i_2,j_2|j_1}(\ell)\big) \\ &= \mathbb{E}[\mathrm{Cov}\big(A_{j_1}(t), \sum_{\ell=1}^{A_{j_1}(t)} N_{i_2,j_2|j_1}(\ell) \;\big|\; A_{j_1}(t)\big) \\ &\quad + \mathrm{Cov}\big(\mathbb{E}[A_{j_1}(t) \mid A_{j_1}(t)], \mathbb{E}[\sum_{\ell=1}^{A_{j_1}(t)} N_{i_2,j_2|j_1}(\ell) \;\big|\; A_{j_1}(t)]\big) \\ &= \mathrm{Cov}\big(A_{j_1}(t), A_{j_1}(t)\,\mathbb{E}[N_{i_2,j_2|j_1}(\ell)]\big) \\ &= \mathrm{Var}\big(A_{j_1}(t)\big)\,\mathbb{E}[N_{i_2,j_2|j_1}(\ell)]\end{aligned}$$

with a similar expression holding for the second term, while the third term on the right hand side of (22) can be rewritten using

$$\begin{aligned}
\mathrm{Cov}\big(\breve{D}_{i_1,j_1}(t), \breve{D}_{i_2,j_2}(t)\big) &= \sum_{k_1=1}^{K}\sum_{k_2=1}^{K} \mathrm{Cov}\Big(\sum_{\ell_1=1}^{A_{k_1}(t)} N_{i_1,j_1|k_1}(\ell_1), \sum_{\ell_2=1}^{A_{k_2}(t)} N_{i_2,j_2|k_2}(\ell_2),\Big) \\
&= \sum_{k=1}^{K} \mathrm{Cov}\Big(\sum_{\ell_1=1}^{A_k(t)} N_{i_1,j_1|k}(\ell_1), \sum_{\ell_2=1}^{A_k(t)} N_{i_2,j_2|k}(\ell_2)\Big) \\
&= \sum_{k=1}^{K} \mathbb{E}_{A_k(t)}\ \mathrm{Cov}\Big(\sum_{\ell_1=1}^{A_k(t)} N_{i_1,j_1|k}(\ell_1), \sum_{\ell_2=1}^{A_k(t)} N_{i_2,j_2|k}(\ell_2) \ \Big|\ A_k(t)\Big) \\
&\quad + \sum_{k=1}^{K} \mathrm{Cov}\Big(\mathbb{E}[\sum_{\ell_1=1}^{A_k(t)} N_{i_1,j_1|k}(\ell_1)|A_k(t)], \mathbb{E}[\sum_{\ell_2=1}^{A_k(t)} N_{i_2,j_2|k}(\ell_2) \ \Big|\ A_k(t)]\Big) \\
&= \sum_{k=1}^{K} \mathbb{E}_{A_k(t)} \sum_{\ell=1}^{A_k(t)} \mathrm{Cov}\big(N_{i_1,j_1|k}(\ell), N_{i_2,j_2|k}(\ell)\big) \\
&\quad + \sum_{k=1}^{K} \mathrm{Cov}\big(A_k(t)\ \mathbb{E}[N_{i_1,j_1|k}(\ell)],\ A_k(t)\ \mathbb{E}[N_{i_2,j_2|k}(\ell)]\big) \\
&= \sum_{k=1}^{K} \mathbb{E}[A_k(t)] \mathrm{Cov}\big(N_{i_1,j_1|k}(\ell),\ N_{i_2,j_2|k}(\ell)\big) \\
&\quad + \mathrm{Var}\big(A_k(t)\big)\ \mathbb{E}[N_{i_1,j_1|k}(\ell)]\ \mathbb{E}[N_{i_2,j_2|k}(\ell)].
\end{aligned}$$

where the second and fourth equalities are due do independence of different customers in the absence of queuing, while the third equality is again the conditional covariance formula.

Substituting in (22) and using $\sum_{i=1}^{K} N_{i,j|k}(\ell) = N_{j|k}(\ell)$ we arrive at

$$\begin{aligned}
\mathrm{Cov}\big(\breve{E}_{j_1}(t), \breve{E}_{j_2}(t)\big) &= \mathrm{Var}\big(A_{j_1}(t)\big)\, \mathbb{E}[N_{j_2|j_1}(\ell)] + \mathrm{Var}\big(A_{j_2}(t)\big)\, \mathbb{E}[N_{j_1|j_2}(\ell)] \\
&\quad + \sum_{k=1}^{K} \mathbb{E}[A_k(t)]\, \mathrm{Cov}\big(N_{j_1|k}(\ell), N_{j_2|k}(\ell)\big) \\
&\quad + \mathrm{Var}\big(A_k(t)\big)\, \mathbb{E}[N_{j_1|k}(\ell)] \mathbb{E}[N_{j_2|k}(\ell)].
\end{aligned}$$

Now dividing by t and letting $t \to \infty$, the result is immediate. □

We now represent $\mathbb{E}[N_{i,j|k}]$ and $\mathrm{Cov}(N_{i_1,j_1|k}, N_{i_2,j_2|k})$ in terms of the routing matrix P. It is an elementary application of "first step analysis" to calculate the desired moments (c.f. [18] and/or [20]), yet we have not seen this specific calculation elsewhere, so we spell out the details. Define:

$$m(i,j) := \begin{bmatrix} \mathbb{E}[N_{i,j|1}] \\ \vdots \\ \mathbb{E}[N_{i,j|K}] \end{bmatrix}, \quad m(i_1,j_1,i_2,j_2) := \begin{bmatrix} \mathbb{E}[N_{i_1,j_1|1}\, N_{i_2,j_2|1}] \\ \vdots \\ \mathbb{E}[N_{i_1,j_1|K}\, N_{i_2,j_2|K}] \end{bmatrix},$$

$$c(i_1,j_1,i_2,j_2) := \left[\begin{array}{c} \mathrm{Cov}\big(N_{i_1,j_1|1}, N_{i_2,j_2|1}\big) \\ \vdots \\ \mathrm{Cov}\big(N_{i_1,j_1|K}, N_{i_2,j_2|K}\big) \end{array} \right].$$

Lemma 5. *The definition of* $m(i,j)$ *in* (12) *agrees with the above, namely*

$$m(i,j) = (I-P)^{-1} e_{i,i} P_{\cdot,j}.$$

Further, let $i_1 \to j_1$ *and* $i_2 \to j_2$ *be distinct flows (i.e.,* $i_1 \neq i_2$*, or* $j_1 \neq j_2$*, or both), then*

$$\begin{aligned} m(i_1,j_1,i_2,j_2) &= m(i_1,j_1) m_{j_1}(i_2,j_2) + m(i_2,j_2) m_{j_2}(i_1,j_1), \qquad (23) \\ m(i,j,i,j) &= m(i,j)\big(1+2m_j(i,j)\big), \end{aligned}$$

and thus,

$$\begin{aligned} c(i_1,j_1,i_2,j_2) &= m(i_1,j_1)\, m_{j_1}(i_2,j_2) + m(i_2,j_2)\, m_{j_2}(i_1,j_1) - m(i_1,j_1) \bullet m(i_2,j_2), \qquad (24) \\ c(i,j,i,j) &= m(i,j)\big(1+2m_j(i,j)\big) - m(i,j) \bullet m(i,j). \end{aligned}$$

Proof. It is well-known that $\mathbb{E}[N_{i|k}]$ is the (k,i)th element of $(I-P)^{-1}$, and clearly $\mathbb{E}[N_{i,j|k}] = \mathbb{E}[N_{i|k}]\, p_{i,j}$, from which the first statement follows. For $\mathbb{E}[N_{i_1,j_1|k}\, N_{i_2,j_2|k}]$ we condition on the first transition from the initial node k, as follows (let $i_1 \neq i_2$, and/or $j_1 \neq j_2$).

$$\begin{aligned} \mathbb{E}[N_{i_1,j_1|k}\ N_{i_2,j_2|k}] &= \sum_{k'=1, k' \notin \{j_1,j_2\}}^{K} p_{k,k'}\ \mathbb{E}[N_{i_1,j_1|k'} N_{i_2,j_2|k'}] + \\ &\quad p_{k,j_1} \mathbb{E}[(\delta_{k,i_1} + N_{i_1,j_1|j_1}) N_{i_2,j_2|j_1}] + p_{k,j_2} \mathbb{E}[N_{i_1,j_1|j_2} (\delta_{k,i_2} + N_{i_2,j_2|j_2})] \\ &= \sum_{k'=1}^{K} p_{k,k'}\ \mathbb{E}[N_{i_1,j_1|k'}\, N_{i_2,j_2|k'}] + \\ &\quad p_{k,j_1} \delta_{k,i_1} \mathbb{E}[N_{i_2,j_2|j_1}] + p_{k,j_2} \delta_{k,i_2} \mathbb{E}[N_{i_1,j_2|j_2}]. \qquad (25) \end{aligned}$$

The equations (25) can be represented as,

$$m(i_1,j_1,i_2,j_2) = P\, m(i_1,j_1,i_2,j_2) + e_{i_1,i_1} P_{\cdot,j_1}\, m_{j_1}(i_2,j_2) + e_{i_2,i_2} P_{\cdot,j_2}\, m_{j_2}(i_1,j_1).$$

or rearranged to,

$$m(i_1,j_1,i_2,j_2) = (I-P)^{-1} \big(e_{i_1,i_1}\, P_{\cdot,j_1}\, m_{j_1}(i_2,j_2) + e_{i_2,i_2}\, P_{\cdot,j_2}\, m_{j_2}(i_1,j_1) \big),$$

which yields (23). In a similar way, we can show that

$$\mathbb{E}[N^2_{i,j|k}] = \sum_{k'=1}^{K} p_{k,k'}\ \mathbb{E}[N^2_{i,j|k'}] + p_{k,j} \delta_{k,i} \big(1 + 2\mathbb{E}[N_{i,j|j}]\big),$$

which gives

$$m(i,j,i,j) = (I-P)^{-1} e_{i,i} P_{\cdot,j} \big(1 + 2m_j(i,j)\big).$$

□

Proof of Theorem 1, (iii): Proposition 1 indicates that under assumption (A3), the variability parameters are the same as those of the zero service time processes. Now the combination of Proposition 2 and Lemma 5 yield the result. □

5 Asymptotic Variance and Uniform Integrability

As stated at onset our original goal is to obtain expressions for $\sigma_{i,j}$ and $\sigma_{i_1\to j_1,i_2\to j_2}$. As we state in Theorem 1, (ii) these can now be read off from the matrices $\Sigma^{(E)}$ and $\Sigma^{(D)}$ respectively. The presentation in this section is for the $\sigma_{i_1\to j_1,i_2\to j_2}$ terms; analogous results for the terms associated with $E(\cdot)$ can be proved in the exact same manner.

Proving Theorem 1, (ii) requires establishing suitable uniform integrability (UI) conditions for the following families:

$$\begin{aligned}
\mathcal{D}_{i,j}^{(1)} &= \Big\{\frac{D_{i,j}(t)-\nu_{i,j}t}{\sqrt{t}},\ t\geq t_0\Big\},\\
\mathcal{D}_{i,j}^{(2)} &= \Big\{\frac{\big(D_{i,j}(t)-\nu_{i,j}t\big)^2}{t},\ t\geq t_0\Big\},\\
\mathcal{D}_{(i_1,j_1),(i_2,j_2)} &= \Big\{\frac{\big(D_{i_1,j_1}(t)-\nu_{i_1,j_1}t\big)\big(D_{i_2,j_2}(t)-\nu_{i_2,j_2}t\big)}{t},\ t\geq t_0\Big\},
\end{aligned}$$

where $t_0>0$ is arbitrary. Note that while each of the families $\mathcal{D}_{i,j}^{(2)}$ is a special case of $\mathcal{D}_{(i_1,j_1),(i_2,j_2)}$, we treat it separately in this section for clarity. See for example [14] for properties of UI sequences and families, and relations to weak convergence.

The following proposition relates the diffusion parameters to the asymptotic variance parameters.

Proposition 3. *If $\mathcal{D}_{i,j}^{(1)}$ and $\mathcal{D}_{i,j}^{(2)}$ are UI then,*

$$\sigma_{i\to j}^2=\Sigma^{(D)}_{(i-1)K+j,\,(i-1)K+j}\ .$$

If $\mathcal{D}_{i,j}^{(1)}$ and $\mathcal{D}_{(i_1,j_1),(i_2,j_2)}$ are UI then,

$$\sigma_{i_1\to j_1,i_2\to j_2}=\Sigma^{(D)}_{(i_1-1)K+j_1,(i_2-1)K+j_2}\ .$$

Proof. By the projection map at time $t=1$ (c.f. [28]) we have the convergence in distribution:

$$\frac{D_{i,j}(t)-\nu_{i,j}t}{\sqrt{t}}\Rightarrow\hat{D}_{i,j}(1).$$

Further, using the continuous mapping theorem we obtain,

$$\frac{\big(D_{i,j}(t)-\nu_{i,j}t\big)^2}{t}\Rightarrow\big(\hat{D}_{i,j}(1)\big)^2.$$

Similarly we have the convergence in distribution on $\mathbb{R}^2$:

$$\Big[\frac{D_{i_1,j_1}(t)-\nu_{i_1,j_1}t}{\sqrt{t}},\ \frac{D_{i_2,j_2}(t)-\nu_{i_2,j_2}t}{\sqrt{t}}\Big]\Rightarrow\Big[\hat{D}_{i_1,j_1}(1),\ \hat{D}_{i_2,j_2}(1)\Big],$$

and thus using the continuous mapping theorem,

$$\frac{D_{i_1,j_1}(t)-\nu_{i_1,j_1}t}{\sqrt{t}}\cdot\frac{D_{i_2,j_2}(t)-\nu_{i_2,j_2}t}{\sqrt{t}}\Rightarrow\hat{D}_{i_1,j_1}(1)\cdot\hat{D}_{i_2,j_2}(1).$$

Under the UI conditions established below the above weak convergences in distribution imply that,

$$\begin{aligned}\lim_{t\to\infty}\mathbb{E}\Big[\frac{D_{i,j}(t)-\nu_{i,j}t}{\sqrt{t}}\Big] &= \mathbb{E}\big[\hat{D}_{i,j}(1)\big],\\ \lim_{t\to\infty}\mathbb{E}\Big[\frac{(D_{i,j}(t)-\nu_{i,j}t)^2}{t}\Big] &= \mathbb{E}\big[\big(\hat{D}_{i,j}(1)\big)^2\big],\end{aligned}$$

as well as,

$$\lim_{t\to\infty}\mathbb{E}\Big[\frac{\big(D_{i_1,j_1}(t)-\nu_{i_1,j_1}t\big)}{\sqrt{t}}\cdot\frac{\big(D_{i_2,j_2}(t)-\nu_{i_2,j_2}t\big)}{\sqrt{t}}\Big]=\mathbb{E}[\hat{D}_{i_1,j_1}(1)\,\cdot\,\hat{D}_{i_2,j_2}(1)].$$

Combining this implies that,

$$\begin{aligned}\sigma^2_{i\to j} &= \lim_{t\to\infty}\frac{\mathrm{Var}\big(D_{i,j}(t)\big)}{t}=\lim_{t\to\infty}\frac{\mathrm{Var}\big(D_{i,j}(t)-\nu_{i,j}t\big)}{t}\\ &= \lim_{t\to\infty}\frac{\mathbb{E}[(D_{i,j}(t)-\nu_{i,j}t)^2]}{t}-\Big(\lim_{t\to\infty}\frac{\mathbb{E}[D_{i,j}(t)-\nu_{i,j}t]}{\sqrt{t}}\Big)^2\\ &= \mathbb{E}[\big(\hat{D}_{i,j}(1)\big)^2]-\big(\mathbb{E}[\hat{D}_{i,j}(1)]\big)^2=\mathrm{Var}\big(\hat{D}_{i,j}(1)\big)=\Sigma^{(D)}_{(i-1)K+j,\,(i-1)K+j}.\end{aligned}$$

Similarly,

$$\begin{aligned}\sigma_{i_1\to j_1,i_2\to j_2} &= \lim_{t\to\infty}\frac{\mathrm{Cov}\big(D_{i_1,j_1}(t),D_{i_2,j_2}(t)\big)}{t}=\lim_{t\to\infty}\frac{\mathrm{Cov}\big(D_{i_1,j_1}(t)-\nu_{i_1,j_1}t,D_{i_2,j_2}(t)-\nu_{i_2,j_2}t\big)}{t}\\ &= \lim_{t\to\infty}\frac{\mathbb{E}[(D_{i_1,j_1}(t)-\nu_{i_1,j_1}t)(D_{i_2,j_2}(t)-\nu_{i_2,j_2}t)]}{t}\\ &- \Big(\lim_{t\to\infty}\frac{\mathbb{E}[D_{i_1,j_1}(t)-\nu_{i_1,j_1}t]}{\sqrt{t}}\Big)\Big(\lim_{t\to\infty}\frac{\mathbb{E}[D_{i_2,j_2}(t)-\nu_{i_2,j_2}t]}{\sqrt{t}}\Big)\\ &= \mathbb{E}[\hat{D}_{i_1,j_1}(1)\,\hat{D}_{i_2,j_2}(1)]-\mathbb{E}[\hat{D}_{i_1,j_1}(1)]\,\mathbb{E}[\hat{D}_{i_2,j_2}(1)]\\ &= \mathrm{Cov}\Big(\hat{D}_{i_1,j_1}(1),\,\hat{D}_{i_2,j_2}(1)\Big)=\Sigma^{(D)}_{(i_1-1)K+j_1,\,(i_2-1)K+j_2}.\end{aligned}$$

□

In establishing the UI, we make use of the following useful inequality:

Lemma 6. *For $r>1$ and arbitrary real values $z_1,\ldots,z_K$,*

$$\Big|\sum_{k=1}^K z_k\Big|^r\le K^{r-1}\sum_{k=1}^K\big|z_k\big|^r. \tag{26}$$

Proof. For positive z_k the function f, defined by $f(z_1,\ldots,z_K)=(\sum_{k=1}^K z_k)^r/\sum_{k=1}^K z_k^r$, has a maximum K^{r-1} at $z_1=\ldots=z_K=\big(\sum_{k=1}^K z_k^r/\sum_{k=1}^K z_k\big)^{1/(r-1)}$. So for general real z_k we have

$$\big|\sum_{k=1}^K z_k\big|^r\le\big(\sum_{k=1}^K|z_k|\big)^r\le K^{r-1}\sum_{k=1}^K|z_k|^r.$$

□

We now establish the required UI.

Proposition 4. *If assumption (A1) holds then the families of random variables $\mathcal{D}^{(1)}_{i,j}$, $\mathcal{D}^{(2)}_{i,j}$ and $\mathcal{D}_{(i_1,j_1),(i_2,j_2)}$ are UI.*

Proof. We first note that UI of $\mathcal{D}^{(2)}_{i,j}$ implies UI of the other two types of families as well, due to Theorem 4.7 (with $p = q = 2$) in Chapter 5 of [14]. To establish UI of $\mathcal{D}^{(2)}_{i,j}$, recall from the previous section the representation $D_{i,j}(t) = \breve{D}_{i,j}(t) - \breve{N}_{i,j}(t)$, where $\breve{D}_{i,j}(t)$ is the number of instantaneous passes on flow $i \to j$, and $\breve{N}_{i,j}(t)$ is the number of future passes on that flow. Together with the triangle inequality and Lemma 6 we have:

$$\begin{aligned}
\Big|\frac{D_{i,j}(t) - \nu_{i,j}t}{\sqrt{t}}\Big| &\leq \Big|\frac{\breve{D}_{i,j}(t) - \nu_{i,j}t}{\sqrt{t}}\Big| + \Big|\frac{\breve{N}_{i,j}(t)}{\sqrt{t}}\Big|, \\
\Big|\frac{\big(D_{i,j}(t) - \nu_{i,j}t\big)^2}{t}\Big| = \Big|\frac{D_{i,j}(t) - \nu_{i,j}t}{\sqrt{t}}\Big|^2 &\leq 2\Big(\Big|\frac{\breve{D}_{i,j}(t) - \nu_{i,j}t}{\sqrt{t}}\Big|^2 + \Big|\frac{\breve{N}_{i,j}(t)}{\sqrt{t}}\Big|^2\Big).
\end{aligned}$$

It thus suffices to show that,

$$\breve{\mathcal{D}}^{(2)}_{i,j} := \Big\{\frac{\big(\breve{D}_{i,j}(t) - \nu_{i,j}t\big)^2}{t},\ t \geq t_0\Big\}, \quad \text{and} \quad \breve{\mathcal{N}}^{(2)}_{i,j} := \Big\{\frac{\big(\breve{N}_{i,j}(t)\big)^2}{t},\ t \geq t_0\Big\},$$

are UI.

To see $\breve{\mathcal{D}}^{(2)}_{i,j}$ is UI it is useful to denote,

$$\breve{D}_{i,j|k}(t) := \sum_{\ell=1}^{A_k(t)} N_{i,j|k}(\ell) \quad \text{and} \quad \nu_{i,j|k} := \alpha_k \mathbb{E}[N_{i,j|k}].$$

Note that since, $\breve{D}_{i,j}(t) = \sum_{k=1}^K \breve{D}_{i,j|k}(t)$, we have $\sum_{k=1}^K \nu_{i,j|k} = \nu_{i,j}$. We now get,

$$\begin{aligned}
\Big|\frac{\big(\breve{D}_{i,j}(t) - \nu_{i,j}t\big)^2}{t}\Big| &= \Big|\frac{\big(\sum_{k=1}^K \breve{D}_{i,j|k}(t) - (\sum_{k=1}^K \nu_{i,j|k})t\big)^2}{t}\Big| \\
&= \Big(\frac{\big|\sum_{k=1}^K \big(\breve{D}_{i,j|k}(t) - \nu_{i,j|k}t\big)\big|}{\sqrt{t}}\Big)^2 \\
&\leq K \sum_{k=1}^K \Big|\frac{\breve{D}_{i,j|k}(t) - \nu_{i,j|k}t}{\sqrt{t}}\Big|^2.
\end{aligned}$$

In the above we again used the triangle inequality and Lemma 6. We now need to show that the families

$$\Big\{\frac{\big(\breve{D}_{i,j|k}(t) - \nu_{i,j|k}t\big)^2}{t},\ t \geq t_0\Big\},$$

are UI:

$$\begin{aligned}
\frac{\big(\breve{D}_{i,j|k}(t)-\nu_{i,j|k}t\big)^2}{t} &= \frac{\big(\sum_{\ell=1}^{A_k(t)} N_{i,j|k}(\ell)-\nu_{i,j|k}t\big)^2}{t}\\
&= \frac{\big(\sum_{\ell=1}^{A_k(t)+1} N_{i,j|k}(\ell)-\nu_{i,j|k}t - N_{i,j|k}(A_k(t)+1)\big)^2}{t}\\
&= \frac{\big(\sum_{\ell=1}^{A_k(t)+1} \big(N_{i,j|k}(\ell)-\frac{\nu_{i,j|k}}{\alpha_k}\big) + (A_k(t)+1)\frac{\nu_{i,j|k}}{\alpha_k} - \nu_{i,j|k}t - N_{i,j|k}(A_k(t)+1)\big)^2}{t}\\
&= \frac{\big(\sum_{\ell=1}^{A_k(t)+1} \big(N_{i,j|k}(\ell)-\frac{\nu_{i,j|k}}{\alpha_k}\big) + ((A_k(t)+1)-\alpha_k t)\frac{\nu_{i,j|k}}{\alpha_k} - N_{i,j|k}(A_k(t)+1)\big)^2}{t}\\
&\le 3\Big(\frac{\big(\sum_{\ell=1}^{A_k(t)+1} \big(N_{i,j|k}(\ell)-\frac{\nu_{i,j|k}}{\alpha_k}\big)\big)^2}{t} + \frac{\big(((A_k(t)+1)-\alpha_k t)\frac{\nu_{i,j|k}}{\alpha_k}\big)^2}{t} + \frac{\big(N_{i,j|k}(A_k(t)+1)\big)^2}{t}\Big)
\end{aligned}$$

The first term is a stopped random walk with zero mean increments where $A_k(t)+1$ is UI by (10). Thus due to Theorems 6.1–6.3 in [15], the first term is UI. The second term is UI again by (10). The third term is obviously UI since the family $N_{i,j|k}(\cdot)$ is i.i.d.

To show that $\breve{\mathcal{N}}^{(2)}_{i,j}$ is UI, we need to show that the second moment of $\breve{N}_{i,j}(t)/\sqrt{t}$ converges (to zero). This 'reverse approach' is due to Remark 5.4 in Chapter 5 of [14]. Define $\breve{N}^{Q}_{i,j|k}(t) := \sum_{\ell=1}^{Q_k(t)} N_{i,j|k}(\ell)$, where $Q_k(t)$ is the queue length at node k at time t. Then the expectation and variance of the random sums $\breve{N}^{Q}_{i,j|k}(t)$, and hence also (by (19)) of $\breve{\mathcal{N}}_{i,j}(t)$, can be expressed in the expectations and variances of $Q_k(t)$ and $N_{i,j|k}(\ell)$, all of which are $o(t)$ (by Assumption (A3)) and finite respectively. Thus the result follows.

□

Proof of Theorem 1, (ii): Proposition 4 establishes UI of the families needed for Proposition 3. □

6 A Numerical Example

Consider the 6 node network illustrated on Figure 1 with parameters,

$$P = \begin{bmatrix} 0 & 0 & 1/2 & 1/2 & 0 & 0\\ 1/2 & 0 & 0 & 1/2 & 0 & 0\\ 1/2 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1/2 & 1/2\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mu = \begin{bmatrix} 8.25\\ 8.25\\ 5\\ 8.25\\ 5\\ 5 \end{bmatrix},$$

$$\alpha = \begin{bmatrix} 1\\ 4\\ 0\\ 0\\ 0\\ 0 \end{bmatrix}, \quad v^2 = \begin{bmatrix} 2\\ 2\\ 0\\ 0\\ 0\\ 0 \end{bmatrix}. \tag{27}$$

For this network,

$$\nu = (I - P')^{-1}\alpha = \begin{bmatrix} 4 & 4 & 2 & 8 & 4 & 4 \end{bmatrix}' < \mu,$$

hence there exist settings under which it can be stabilized and thus assumptions (A1)–(A3) hold (for example a generalized Jackson network with a non-idling policy and light-tailed inter-arrival and service times). Note that besides verification of the above inequality, the values of μ do not play a further role in the calculation of the variability parameters. Nevertheless, we use them in a simulated example below.

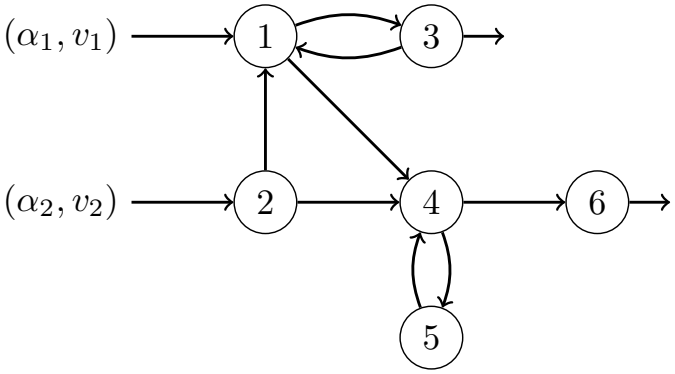

Figure 1: An example network.

It is now a straight forward matter to use (13) (or alternatively (14)–(16)) from our main theorem to obtain variability parameters. Note that in this process, the only matrix that requires inversion is $(I - P')$. The rest of the calculations follow from matrix composition, addition and multiplication operations.

The resulting matrix $\Sigma^{(D)}$ is of dimension 36×36. We present the diagonals of this matrix (which are $\sigma^2_{i\to j}$) in the following table:

$i\backslash j$	1	2	3	4	5	6
1	0	0	32/9	20/9	0	0
2	3/2	0	0	3/2	0	0
3	31/18	0	0	0	0	0
4	0	0	0	0	199/18	55/18
5	0	0	0	199/18	0	0
6	0	0	0	0	0	0

As a further illustration we present a few selected non-diagonal elements of $\Sigma^{(D)}$:

$$\sigma_{2\to1,2\to4} = -1/2, \quad \sigma_{4\to5,5\to4} = 199/18, \quad \sigma_{1\to3,4\to6} = 5/19, \quad \sigma_{1\to3,2\to4} = -1/3.$$

In discussing these values, it is good to consider the *asymptotic correlation coefficient*:

$$\rho_{i_1\to i_2, j_1\to j_2} := \frac{\sigma_{i_1\to i_2, j_1\to j_2}}{\sqrt{\sigma^2_{i_1\to i_2}\sigma^2_{j_1\to j_2}}}.$$

For these selected flow pairs it evaluates to

$$\rho_{2\to1,2\to4} = -\frac{1}{3}, \quad \rho_{4\to5,5\to4} = 1, \quad \rho_{1\to3,4\to6} \approx 0.16856, \quad \rho_{1\to3,2\to4} \approx -0.14434.$$

The first two values are easily explained in our example, the other two are not. For $\rho_{2\to1,2\to4}$ consider the Bernoulli splitting at the output of queue 2 and the fact there is no feedback to this queue. Recall that in this case $\sigma_{2\to1,2\to4} = (v_2^2 - \alpha_2)/4$ for $v_2^2 = 2, \alpha_2 = 4$. In this case the asymptotic correlation coefficient is $(v_2^2 - \alpha_2)/(v_2^2 + \alpha_2)$. In considering $\rho_{4\to5,5\to4}$ observe that there is no random routing in this part of the network: All jobs that enter 5 come from 4 and then return to 5.

We are not aware of an "easy" explanation of the values of $\rho_{1\to3,4\to6}$ and $\rho_{1\to3,2\to4}$. It is insightful to see that as in this case, some correlations between flows are positive while others are negative. We do not know of an a priori way of finding out the sign of these correlations without using our main result. In fact, evaluating $\Sigma^{(D)}$ with v_2 as free variable, we get,

$$\rho_{1\to3,2\to4} = \frac{v_2^2 - 4}{\sqrt{(v_2^2+4)(v_2^2+30)}}.$$

We thus see that the sign of the correlation between those two flows depends on the variability of the arrival process into 2. Observe that in the asymptotically uncorrelated case (i.e. when $v_2 = 4$),

$$\lim_{t\to\infty} \frac{\mathrm{Var}\big(A_2(t)\big)}{\mathbb{E}[A_2(t)]} = 1,$$

as is for a Poisson process. This is consistent with the fact that in the case of a classic Jackson network (Poisson arrival process and exponential processing times) case, since node 2 has no feedback its output is a Poisson process and splitting of departures from node 2 results in two independent Poisson flows, $2 \to 1$ and $2 \to 4$. The first of these flows affects $1 \to 3$ but not the second. Hence in such a case it is expected that $\rho_{1\to3,2\to4} = 0$.

Arrivals to Individual Queues

Moving onto arrival processes into individual queues, application of our main result yields:

$$\Sigma^{(E)} = \begin{bmatrix} 68/9 & 4/3 & 40/9 & 44/9 & 22/9 & 22/9 \\ & 2 & 2/3 & 10/3 & 5/3 & 5/3 \\ & & 32/9 & 10/9 & 5/9 & 5/9 \\ & & & 182/9 & 127/9 & 55/9 \\ & & & & 199/18 & 55/18 \\ & & & & & 55/18 \end{bmatrix}.$$

Observe that $\sigma_2^2 = 2$ as expected since there are only exogenous arrivals to this queue. Further since all jobs that pass through queue 5 eventually also pass through queue 6 we have,

$$\sigma_{k,5} = \sigma_{k,6}, \quad k = 1, 2, 3, 6.$$

It is the diagonal elements of $\Sigma^{(E)}$ that may be useful for network decomposition approximations (which we do not explore further in this paper). Normalizing the diagonals by ν we get,

$$c^2 = \begin{bmatrix} 1.89 & 0.5 & 1.78 & 2.53 & 2.76 & 0.76 \end{bmatrix}'. \tag{28}$$

Simulation Results

To further illustrate our result and explore the effect of different policies and constraints on the variance of flows, we carried out a Monte-Carlo simulation of the example network.

In the simulation we set the service distributions of queue k to be distributed as a sum of two i.i.d. exponential random variables, each with mean $(2\mu_k)^{-1}$. This results in a so-called Erlang 2 distribution (having a squared coefficient of variation of $1/2$) with mean μ_k^{-1}.

The arrival process, $A_1(\cdot)$, is the more variable of the two arrival processes. It is taken to be a renewal process of inter-arrival times that are distributed as a mixture of two independent exponential random variables (hyper-exponential): *w.p.* $1/3$ a mean 2 exponential and *w.p.* $2/3$ a mean $1/2$ exponential. This distribution has mean 1 and squared coefficient of variation 2 agreeing with α_1 and v_1^2 as specified in (27).

The arrival process, $A_2(\cdot)$, is less variable. It is taken to be a renewal process with inter-arrival times that are Erlang 2 distributed this time with mean $1/4$. This is in agreement with α_2 and v_2^2 as specified in (27).

We consider two settings:

Single-class: Each queue has a dedicated (separate) server. This is a generalized Jackson network.

Multi-class: Queues 1 and 2 are served by the same server under a non-pre-emptive priority policy giving priority to queue 1. All other queues have their own server. Note that in this case the load on the server of queues 1 and 2 is $\nu_1/\mu_1 + \nu_2/\mu_2 \approx 0.97 < 1$. I.e. it is quite heavily loaded but is still stable. Note in general having a load of less than unity does not immediately imply that the system is stable yet for this simple case it can be shown that stability holds under such a priority policy (c.f. [2]).

Besides exemplifying the correctness of our theoretical results, the goal in this simulation set-up is to illustrate that while the asymptotic variability parameters do not depend on service times and scheduling policies, the shape of the variance curve is in general influenced by such factors.

We ran 2×10^5 simulation runs of each case (single-class and multi-class) each for $1,000$ time units, starting at time $t = 0$ with the system empty[1]. We then estimated $\mathrm{Var}\big(D_{5\to4}(t)\big)$ for each run over a grid of time points $t = 20, 40, 60, \ldots, 1000$, by taking the sample variance at each time point over 2×10^5 observations. Note that we purposely observe the flow $5 \to 4$ which is not directly adjacent to the multi-class server serving 1 and 2.

Our main theorem applied to this example implies that in both the single-class and multi-class case, for non-small t,

$$\mathrm{Var}\big(D_{5\to4}(t)\big) \approx \sigma^2_{5\to4} t = \frac{199}{18} t = 11.05\overline{5}\, t.$$

This is illustrated in Figure 2 (top) where we plot the variance curves versus the approximation $\sigma^2_{5\to4}t$. To take a closer look at the effect of single-class vs. multi-class we then plot the bias, $\sigma^2_{5\to4}t - \mathrm{Var}\big(D_{5\to4}(t)\big)$ on Figure 2 (bottom). It is indeed evident that different system characteristics yield different variance curves.

[1]The simulation was carried out using a simulation package written in C++: PRONETSIM. See [24], Appendix A, for details about this software.

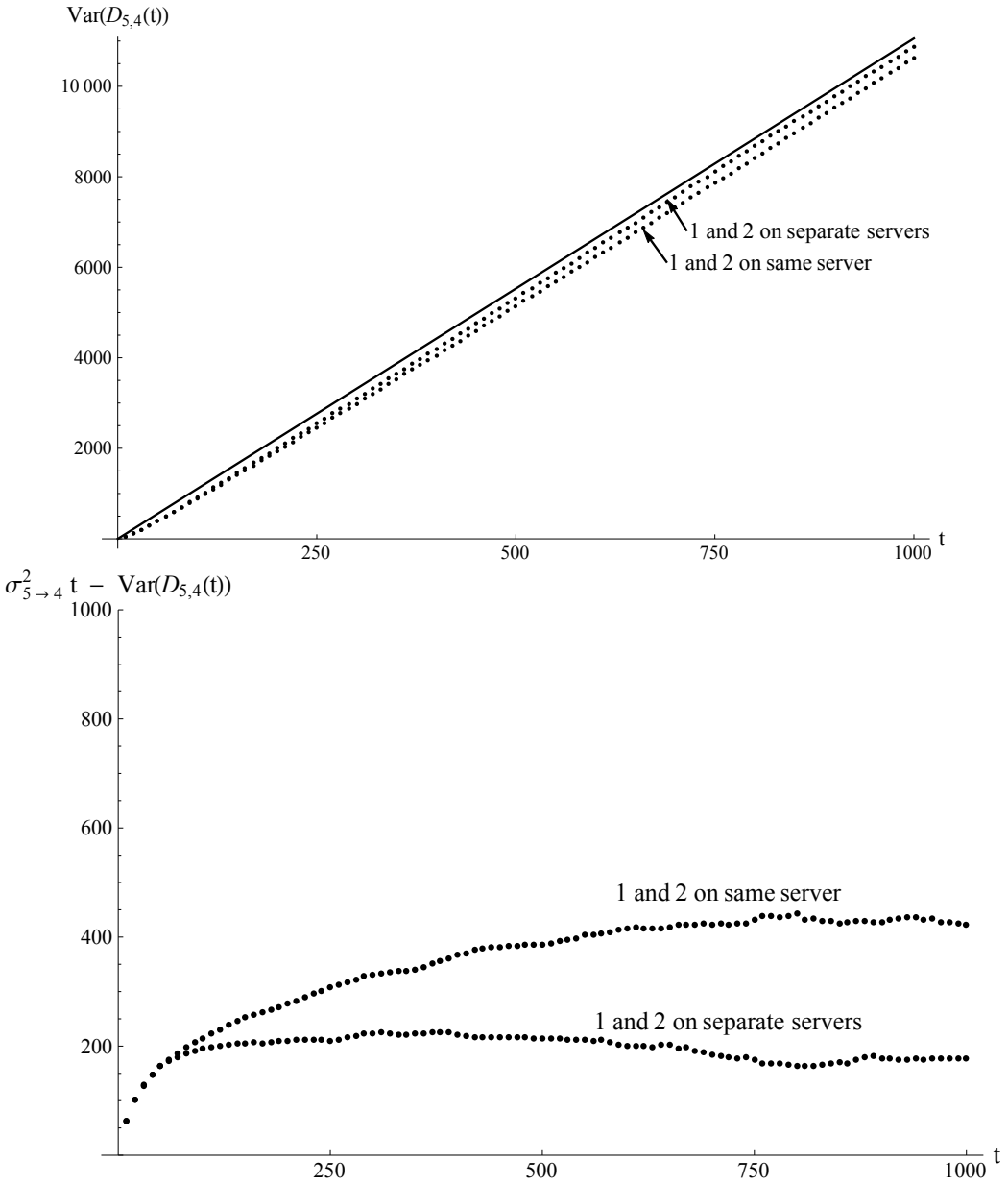

Figure 2: Simulation estimates of $\mathrm{Var}\big(D_{5\to 4}(t)\big)$ for two cases: single-class (1 and 2 on separate servers) and multi-class (1 and 2 on same server with a priority policy). The top graph illustrates the variance curve estimates (dotted) vs. the solid line $\sigma^2_{5\to 4}t$. The bottom graph shows the bias: $\sigma^2_{5\to 4}t - \mathrm{Var}\big(D_{5\to 4}(t)\big)$. As is illustrated, both systems have the same asymptotic variance for $D_{5\to 4}(t)$, yet their variance curves differ for finite t.

It is somewhat expected that the multi-class case will have a higher bias, since in this case the server of 1 and 2 is under a heavier load (0.97). Further, in that case one can expect more "bursts" on the flow $2 \to 4$ since queue 2 is served with low-priority. These bursts perhaps "propagate" to flow $4 \to 5$ and ultimately to the flow which we measure: $5 \to 4$. Nevertheless, such phenomena are not captured by the asymptotic quantities found

in the current paper. It should be noted that in [16] second order properties of this sort are explored for elementary queueing systems such as the stable M/G/1 queue. It is not clear how to extend such an investigation to networks.

7 Conclusion

While stable queueing networks have been analysed for decades, up to now, exact expressions for the asymptotic variability of flows have not been known. In this paper we put forward easy computable expressions together with a simple diffusion limit theorem for the flows. It is interesting to see if and how the manufacturing queueing modeling (c.f. [5]) and queueing network decomposition community will adopt our results and incorporate them in heuristic decomposition schemes.

The queueing networks we considered in this paper are assumed to be open and stable. This stands in contrast with the more general case handled in [6] (where nodes are allowed to be either under-loaded, over-loaded or critical). It should be mentioned that our results easily carry over to the case where some nodes are over-loaded. In this case, the service times of over-loaded nodes contributes to the exogenous arrivals in a straightforward manner (see for example [13] for an early treatment of this idea). On the contrary, the case in which some nodes are critical is more challenging. In that case, the single-server queue was only recently handled with some difficulty in [1]. There the authors observed a BRAVO effect (Balancing Reduces Asymptotic Variance of Outputs). We do not handle this in the network context. Thus the challenge of finding the asymptotic variability of flows in critical queueing networks remains.

Acknowledgements: We thank Gideon Weiss for fruitful discussions at onset. This work began while Yoni Nazarathy was affiliated with Swinburne University of Technology, Melbourne. Yoni Nazarathy is supported by Australian Research Council (ARC) grants DP130100156 and DE130100291. Part of the work was carried out while Werner Scheinhardt was supported by an Ethel Raybould Visiting Fellowship to the University of Queensland.

References

[1] A. Al Hanbali, M. Mandjes, Y. Nazarathy, and W. Whitt. The asymptotic variance of departures in critically loaded queues. *Advances in Applied Probability*, 43:243–263, 2011.

[2] M. Bramson. *Stability of Queueing Networks*. Springer, 2008.

[3] M. Bramson and J. G. Dai. Heavy traffic limits for some queueing networks. *Annals of Applied Probability*, 11(1):49–90, 2001.

[4] P.J. Burke. The output of a queuing system. *Operations Research*, 4(6):699–704, 1956.

[5] J.A. Buzacott and J.G. Shanthikumar. *Stochastic Models of Manufacturing Systems*. Prentice Hall, 1993.

[6] H. Chen and A. Mandelbaum. Stochastic discrete flow networks: Diffusion approximations and bottlenecks. *The Annals of Probability*, 19(4):1463–1519, 1991.

[7] H. Chen and D.D. Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer, 2001.

[8] J.G. Dai and M. J. Harrison. Steady-state analysis of RBM in a rectangle: Numerical methods and a queueing application. *The Annals of Applied Probability*, 1(1):16–35, 1991.

[9] D.J. Daley. Queueing output processes. *Advances in Applied Probability*, 8:395–415, 1976.

[10] R. L. Disney and P. C. Kiessler. *Traffic Processes in Queueing Networks – A Markov Renewal Approach*. The Johns Hopkins University Press, 1987.

[11] R. L. Disney and D. Konig. Queueing networks: A survey of their random processes. *SIAM Review*, 27(3):335–403, 1985.

[12] P. W. Glynn. Diffusion approximations. *In Handbooks in Operations Research, Vol 2, D.P. Heyman and M.J. Sobel (eds.), North-Holland, Amsterdam*, pages 145–198, 1990.

[13] J.B. Goodman and W. Massey. Non-ergodic Jackson network. *Journal of Applied Probability*, 21(4):860–869, 1984.

[14] A. Gut. *Probability: A graduate course*. Springer, 2005.

[15] A. Gut. *Stopped random walks: Limit theorems and applications*. Springer, 2009.

[16] S. Hautphenne, Y. Kerner, Y. Nazarathy, and P.G. Taylor. The second order terms of the variance curves for some queueing output processes. *arXiv preprint arXiv:1311.0069*, 2013.

[17] J. R. Jackson. Jobshop-like queueing systems. *Management Science*, 10(1):131–142, 1963.

[18] S. Karlin and H.M. Taylor. *A First Course in Stochastic Processes*. Academic Press, 1975.

[19] F. Kelly. *Reversibility and Stochastic Networks*. John Wiley & Sons, 1979.

[20] J.G. Kemeny and J.L. Snell. *Finite Markov Chains*. Springer-Verlag, 1960.

[21] S. Kim. Modeling cross correlation in three-moment four-parameter decomposition approximation of queueing networks. *Operations research*, 59(2):480–497, 2011.

[22] P. Kuehn. Approximate analysis of general queuing networks by decomposition. *IEEE Transactions on Communications*, 27(1):113–126, 1979.

[23] S. P. Meyn. *Control Techniques for Complex Networks*. Cambridge University Press, 2008.

[24] Y. Nazarathy. *On control of queueing networks and the asymptotic variance rate of outputs*. PhD thesis, University of Haifa, 2008.

[25] Y. Nazarathy and G. Weiss. Positive Harris recurrence and diffusion scale analysis of a push pull queueing network. *Performance Evaluation*, 67(4):201–217, 2010.

[26] M.I. Reiman. Open queueing networks in heavy traffic. *Math. Oper. Res.*, 9(3):441–458, 1984.

[27] W. Whitt. Performance of the queueing network analyzer. *Bell System Technical Journal*, 62(9):2817–2843, 1983.

[28] W. Whitt. *Stochastic Process Limits*. Springer New York, 2002.

[29] R. J. Williams. Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems*, 30:27–88, 1998.