Stochastic Systems
2011, Vol. 1, No. 2, 306–339
DOI: 10.1214/11-SSY026

ANALYSIS OF A SPLITTING ESTIMATOR FOR RARE EVENT PROBABILITIES IN JACKSON NETWORKS

By Jose Blanchet*, Kevin Leder† and Yixi Shi*

*Columbia University** *and University of Minnesota*†

We consider a standard splitting algorithm for the rare-event simulation of overflow probabilities in any subset of stations in a Jackson network at level n, starting at a fixed initial position. It was shown in [8] that a subsolution to the Isaacs equation guarantees that a subexponential number of function evaluations (in n) suffices to estimate such overflow probabilities within a given relative accuracy. Our analysis here shows that in fact $O(n^{2\beta_V+1})$ function evaluations suffice to achieve a given relative precision, where β_V is the number of bottleneck stations in the subset of stations under consideration in the network. This is the first rigorous analysis that favorably compares splitting against directly computing the overflow probability of interest, which can be evaluated by solving a linear system of equations with $O(n^d)$ variables.

1. Introduction. The development of rare-event simulation algorithms for overflow probabilities in stable open Jackson networks has been the subject of a substantial amount of papers in the literature during the last decades (see Section 2 for the specification of an open Jackson network). A couple of early references on the subject are [22] and [1]. Subsequent work which has also been very influential in the development of efficient algorithms for overflows of Jackson networks include [24, 13, 14, 18, 16, 10, 21, 12] and [8]. The survey papers of [17] and [7] provide additional references on this topic.

The two most popular approaches that are applied to the construction of efficient rare-event simulation algorithms are importance sampling and splitting (see [3]). Importance sampling involves simulating the system under consideration (in our case the Jackson network) according to a different set of probabilities in order to induce the occurrence of the rare event. Then, one attaches a weight to each simulation corresponding to the likelihood ratio of the observed outcome relative to the nominal/original distribution. In splitting, on the other hand, there is no attempt to bias the behavior of the system. Instead, the rare event of interest (in our case overflow in a Jackson network) is decomposed into a sequence of nested "milestone" events whose

Received February 2011.

subsequent occurrence is not rare. The rare event occurs when the last of the milestone events occurs. The idea is to keep splitting the particles as they reach subsequent milestones. Of course, each particle is associated with a weight corresponding to the total number of times it has split, so that the overall estimation (which is the sum of the weights corresponding to the particles that make it to the last milestone) provides an unbiased estimator of the probability of interest.

The most popular performance measure for efficiency analysis of rare-event simulation algorithms for Jackson networks corresponds to that of "asymptotic optimality" or "weak efficiency". In order to both explain the computational complexity implied by this notion and to put in perspective our contributions let us discuss the class of problems we are interested in: Starting from any fixed state, we consider the problem of computing the probability that the total number of customers in any fixed set of stations in the network reaches level n prior to reaching the origin. *In other words, we consider the probability that the sum of the queue lengths in any given subset of stations reaches level n within a busy period.* The number of stations in the whole network is assumed to be d and the number of bottleneck stations (i.e. stations with the maximum traffic intensity in equilibrium) is β.

Weak efficiency guarantees that a subexponential number of replications (as a function of the overflow level, say n) suffices for computing the underlying overflow probability of interest within a given relative accuracy. In contrast, as we shall explain in Section 2, overflow probabilities in the setting of Jackson networks can be computed by solving a linear system of equations with $O(n^d)$[1] unknowns. It is well known that Gaussian elimination then requires $O(n^{3d})$ operations (additions and multiplications) to find the exact solution. Moreover, since in our case the associated linear system has some sparsity properties the linear equations can be solved in at most $O(n^{3d-2})$ operations (see the discussion in Section 2). Our analysis for the solution of the associated linear system of equations is not intended to be exhaustive. Our objective is simply to make the point that naive Monte Carlo (which indeed takes an exponential number of replications in n to achieve a given relative accuracy) is not the natural benchmark that one should be using in order to test the performance of an efficient simulation estimator for overflows in Jackson networks. Rather, a more natural benchmark is the application of a straightforward method for solving the associated system of linear equations. It would be interesting to provide a detailed study of various methods for solving linear systems of equations (such as multi-grid

[1]Given two non-negative functions $f(\cdot)$ and $g(\cdot)$, we say $f(n) = O(g(n))$ if there exists $c, n_0 \in (0, \infty)$ such that $f(n) \leq cg(n)$ all $n \geq n_0$.

procedures) that are suitable for our environment and can even be combined with the ideas behind efficient simulation procedures. This, however, would be the subject of an entire paper and therefore is left as a topic for future research.

Our goal here is to analyze a class of splitting algorithms similar to those introduced in [24] for the evaluation of overflow probabilities at level n. Further analysis was given in [8], where the authors provide necessary and sufficient conditions for the design of the "milestone events" in order to achieve subexponential complexity in n.

Our contribution is to show that if the milestone events are properly placed as suggested by [8], the splitting algorithm requires $O(n^{2\beta+1})$ function evaluations (basically simple operations, see page 5 for a definition and discussion) to achieve a fixed relative error. Since clearly the number of bottleneck stations β is at most d, the complexity of splitting is $O(n^{2d+1})$, which is substantially smaller than that of the direct solution of the associated linear system. Our analysis therefore provides theoretical justification for the superior performance observed when applying splitting algorithms compared to directly solving the associated linear system. The precise statement of our main results is given in Theorem 1, at the end of Section 5.

We believe that our results shed light into the type of performance that can be expected when applying particle algorithms beyond the setting of Jackson networks. This feature should be emphasized, specially given the fact that a linear time algorithm for computing overflows in Jackson networks has been developed very recently (see [4]). Contrary to particle methods, which are versatile and that can in principle be applied in great generality, the algorithm in [4] takes advantage of certain properties of Jackson networks which are not shared by all classes of systems.

In addition, our results also provide interesting connections to recent performance analyses studied in the context of state-dependent importance sampling algorithms for a class of Jackson networks. These connections might eventually help guide the users of rare event simulation algorithms to decide when to apply importance sampling or splitting. For instance, consider the overflow at level n of the total population of a tandem network with d stations. The work of [10] proposes an importance sampling estimator based on the subsolution of an associated Isaacs equation. In particular, [10] shows that if exponential tiltings are applied using the gradient of the associated subsolution as the tilting parameter (depending on the current state), the corresponding algorithm is weakly efficient. It turns out that many subsolutions can be constructed by varying certain so-called "mollification parameters". A recent analysis based on Lyapunov inequalities given

in [6] shows that a natural selection of mollification parameters guarantees $O(n^{2(d-\beta)+1})$ function evaluations to achieve a given relative error. Our analysis here therefore guarantees that one can achieve a running time of order $O(n^{d+1})$ if one chooses importance sampling when there are more than $d/2$ bottleneck stations in the network and splitting if there are less than $d/2$ bottleneck stations. Although our analysis is still not sharp we believe that our results provide a significant step forward in understanding the connections between splitting and importance sampling.

The rest of the paper is organized as follows. A brief discussion on complexity and efficiency considerations is given in Section 2. Then we discuss the necessary large deviations asymptotics for Jackson networks required for our analysis in Section 3. The introduction of the splitting algorithm as well as connections to the theory developed in [8] is given in Section 4. Our complexity analysis is finally given in Section 5.

2. Complexity and efficiency. We shall review concepts of efficiency and complexity in rare event simulation. We start our discussion in the context of a generic class of rare event simulation problems. Consider a sequence of events $\{E_n, n = 1, 2 \ldots\}$ with $p_n \triangleq \mathbb{P}(E_n) \to 0$ as $n \nearrow \infty$ (Without loss of generality, we might assume that $p_n \to 0$ exponentially fast as $n \nearrow 0$.) The design of an efficient rare-event simulation algorithm is typically associated with the construction of an unbiased estimator, say $\hat{p}_n$, such that $p_n = \mathbb{E}[\hat{p}_n]$. A number of m i.i.d. replications $\{\hat{p}_n^{(1)}, \ldots, \hat{p}_n^{(m)}\}$ is produced, the average of which forms an estimate of p_n, namely

$$\hat{p}_n(m) = \frac{1}{m} \sum_{j=1}^{m} \hat{p}_n^{(j)}.$$

By virtue of Chebyshev's inequality we obtain the following property for the relative error, $|\hat{p}_n(m) - p_n|/p_n$, of the estimate

$$\mathbb{P}(|\hat{p}_n(m) - p_n|/p_n > \epsilon) \leq \frac{\mathrm{Var}(\hat{p}_n)}{m p_n^2 \epsilon^2}. \tag{1}$$

Hence, for a pre-determined upper bound ϵ of relative error, if we choose the number of replications m such that

$$m \geq \epsilon^{-2} \delta^{-1} (cv_n)^2, \tag{2}$$

where $cv_n^2 = \mathrm{Var}(\hat{p}_n)/p_n^2$ is the squared coefficient of variation of $\hat{p_n}$, we can guarantee that the relative error is no larger than ϵ with probability at least $1 - \delta$.

Equation (2) stipulates that m needs to grow at least at the same rate as cv_n^2 does in order to keep the relative error within a desirable threshold. If cv_n^2 grows at a subexponential rate (i.e. if $\log\left(cv_n\right)^2 = o\left(1/\log p_n\right)$, as $n \nearrow \infty$) the estimator is said to have *asymptotic optimality, logarithmic efficiency* or *weak efficiency*. In this case, the number of replications needs to increase subexponentially in n to achieve a prescribed level of relative accuracy. The name "asymptotic optimality" is derived from the fact that weak efficiency implies that the exponential rate of decay to zero of the $\mathbb{E}\hat{p}_n^2$ coincides with that of p_n^2 and therefore is maximal (by virtue of Jensen's inequality).

Obviously, one has to keep in mind that weak efficiency measures the optimality of the estimator for a given level of computational budget. For the splitting algorithm, it is apparent that the computational effort varies drastically with the degree of splitting performed; one must therefore take into account the cost involved in generating each replication of $\hat{p}_n$. We measure such cost in terms of the number of elementary function evaluations which we will take to be simple addition, multiplication, comparison, and the generation of a single uniform random variable. Ultimately, setting up the splitting algorithm requires the evaluation of just two logarithms of quantities which are readily available from the problem's input. We do not count these evaluations in the cost per replication. When we incorporate the computational cost per replication of the estimator, (2) says that the total number of function evaluations needed has to keep pace with the work-normalized squared coefficient of variation, i.e., $cv_n^2 \cdot N_n$, where N_n is the cost per replication of $\hat{p}_n$. We will show in Section 5 that N_n is closely related to the expected total number of the survival particles in a single run of the splitting algorithm.

In the setting of Jackson networks, it is important to recognize that overflow probabilities can be obtained by solving a system of linear equations. Therefore, a reasonable benchmark procedure for testing "efficiency" in any simulation based algorithm is to compare costs with those associated with directly solving the linear system. Jackson networks are basically multidimensional simple random walks with constrained behavior on the boundaries. In particular, they are Markov chains living on a countable state-space. The overflow probabilities can be conveniently expressed as first passage time probabilities, which in turn can be characterized as the solution to certain linear system of equations thanks to its countable state-space Markov chain structure. We shall quickly review how to obtain such linear system for a generic Markov chain $Q = \{Q_k : k \geq 0\}$ living on a countable state-space $\mathcal{S}$ with transition matrix $\{K(x,y) : x, y \in \mathcal{S}\}$. Let A, B be two disjoint subsets of $\mathcal{S}$, define $\sigma_A \triangleq \inf\{k \geq 0 : X \in A\}$, $\sigma_B \triangleq \inf\{k \geq 0 : X \in B\}$

and put $p(x) = \mathbb{P}_x(\sigma_A \leq \sigma_B)$. A simple conditioning argument on the first transition leads to

$$p(x) = \sum_{y \in \mathcal{S}} K(x,y)\, p(y) \tag{3}$$

subject to the boundary conditions

$$p(x) = 1 \text{ for } x \in A, \quad p(x) = 0 \text{ for } x \in B.$$

In fact, $p(\cdot)$ is the minimum non-negative solution to the above system (see [5]).

Now, if Q describes the state of the embedded discrete time Markov chain corresponding to a Jackson network with d stations then $\mathcal{S} = \mathcal{Z}_+^d$. The transition dynamics of a Jackson network are specified as follows (see [23] p. 92). Inter-arrival times and service times are all independent and exponentially distributed random variables. The arrival rates are given by the vector $\lambda = (\lambda_1, \ldots, \lambda_d)^T$ and service rates are given by $\mu = (\mu_1, \ldots, \mu_d)^T$. (By convention all of the vectors in this paper are taken to be column vectors and ^{T} denotes transposition.) A job that leaves station i joins station j with probability $P_{i,j}$ and it leaves the system with probability

$$P_{i,0} \triangleq 1 - \sum_{j=1}^{d} P_{i,j}.$$

The matrix $P = \{P_{i,j} : 1 \leq i, j \leq d\}$ is called the routing matrix. We shall consider open Jackson networks, which satisfy the following conditions:

i) $\forall i$, either $\lambda_i > 0$ or $\lambda_{j_1} P_{j_1 j_2} \ldots P_{j_k i} > 0$ for some $j_1, \ldots, j_k$.
ii) $\forall i$, either $P_{i0} > 0$ or $P_{i j_1} P_{j_1 j_2} \ldots P_{j_k 0} > 0$ for some $j_1, \ldots, j_k$.
iii) The network is stable (i.e. a stationary distribution exists).

These conditions simply require that each station will receive jobs either directly from the outside or routed from other stations, and each job will leave the system eventually. Our main interest lies in the evaluation of $p_n(x)$ assuming that $B = \{0\}$ and $A_n = \{y : v^T y = n\}$ where v is a binary vector which encodes a particular subset of the network (i.e., the i-th position of the vector v is 1 if station i falls in the subset of interest, and 0 otherwise). We shall denote by $V(x) = x^T v$ the mapping recording the total population in the stations corresponding to the vector v. The case in which $v = \mathbf{1} = (1, 1, \ldots, 1)^T$ corresponds to the total population of the system. So, $p_n(x)$, or more precisely $p_n^V(x)$, corresponds to the overflow probability in the subset encoded by v within a busy period starting from x. In this

setting, it follows (as we shall review in the next section) that $p_n^V(x) \longrightarrow 0$ exponentially fast in n as $n \nearrow \infty$ and the system of equations (3) has $O(n^d)$ unknowns. Gaussian elimination requires $O(n^{3d})$ function evaluations to find the solution of such system. But since each state of the Markov chain in this case has possible interactions with only a small fraction of the entire state-space, it is therefore possible to permutate the states (say in lexicographic order) so that the system is banded (i.e. the associated matrix is sparse in the sense that its non-zero entries fall to a diagonal band.) One can show that the bandwidth is $O(n^{d-1})$, and therefore solving such a banded linear system requires $O(n^d \cdot (n^{d-1})^2) = O(n^{3d-2})$ operations (see, e.g., [2]).

Estimators that possess weak efficiency (in a work-normalized sense) are guaranteed to run at subexponential complexity. When comparing to the above *polynomial* algorithms of solving systems of linear equations, the efficiency analysis of such estimators appears to be insufficient. We will show in later analysis that the multilevel splitting algorithm suggested by Dean and Dupuis [8], applied to estimate the overflow probabilities in Jackson networks, requires fewer function evaluations than directly solving the associated system of linear equations.

3. Jackson networks: Notation and properties. As we mentioned in the previous section, a Jackson network is encoded by two vectors of arrival and service rates, $\lambda = (\lambda_1, \ldots, \lambda_d)^T$ and $\mu = (\mu_1, \ldots, \mu_d)^T$, together with a routing matrix $P = \{P_{i,j} : 1 \leq i, j \leq d\}$. Without loss of generality, we assume that $\sum_{i=1}^d (\lambda_i + \mu_i) = 1$. The network is assumed to be open and stable so conditions i), ii), and iii) described in the previous section are in place.

Given the stability assumption, the system of equations given by

$$\phi_i = \lambda_i + \sum_{j=1}^{d} \phi_j P_{ji}, \quad \forall i = 1, 2, \ldots, d \tag{4}$$

admits a unique solution $\phi^T = \lambda^T (I - P)^{-1}$ (see [3]). The traffic intensity at station i in the system in equilibrium is given by ρ_i which is defined by

$$\rho_i = \frac{\phi_i}{\mu_i} = \frac{[\lambda^T (I - P)^{-1}]_i}{\mu_i}, \tag{5}$$

and satisfies $\rho_i \in (0, 1)$ for all $i = 1, 2, \ldots, d$. Define $\rho_* = \max_{1 \leq i \leq d} \rho_i$ and let β be the cardinality of the set $\{i : \rho_i = \rho_*\}$.

We shall study the queueing network by means of the embedded discrete time Markov chain $Q = \{Q(k) : k \geq 0\}$, where $Q(k) = (Q_1(k), \ldots, Q_d(k))$.

For each k, $Q_i(k)$ represents the number of customers in station i immediately after the k-th transition epoch of the system. As mentioned before, the process Q lives in the space $\mathcal{S} = \mathcal{Z}_+^d$.

Let $V(x) = x^T v$ be the total population in the stations corresponding to the binary vector v. We are interested in the overflow probability in any given subset of the Jackson network. More precisely, we wish to estimate

$$p_n^V = \mathbb{P}\{\text{total population in stations encoded by } v \text{ reaches } n \text{ before returning to 0, starting from 0}\}. \tag{6}$$

In turn, p_n^V can be expressed in terms of the following stopping times,

$$T_{\{x\}} \triangleq \inf\{k \geq 1 : Q(k) = x\},$$
$$T_n^V \triangleq \inf\{k \geq 1 : V(Q(k)) \geq n\}.$$

Indeed, if we use the notation $\mathbb{P}_x(\cdot) \triangleq \mathbb{P}(\cdot|Q(0) = x)$ then we can rewrite p_n^V as

$$p_n^V = \mathbb{P}_0(T_n^V \leq T_{\{0\}}). \tag{7}$$

Similarly,

$$p_n^V(x) = \mathbb{P}_x(T_n^V \leq T_{\{0\}}). \tag{8}$$

The asymptotic analysis of $p_n^V(x)$ can be studied by means of large deviations theory. We shall indicate how this theory can be applied to specify an efficient splitting algorithm in the next section. In the mean time, let us provide a representation for the dynamics of the queue length process that will be convenient in order to motivate the elements of the efficient splitting algorithm that we shall analyze.

As mentioned earlier, Jackson networks are basically constrained random walks. The constraints arise because the number of customers in each station must be non-negative. Thinking about Jackson networks as constrained random walks facilitates the introduction and motivation of the necessary large deviations elements behind the description of the splitting algorithm. In order to specify the dynamics of the embedded discrete time Markov chain in terms of a random walk type representation we need to introduce notations which will be useful to specify the transitions at the boundaries induced by the non-negativity constraints.

The state-space $\mathcal{Z}_+^d$ can be partitioned into 2^d different regions which are indexed by all the subsets $E \subseteq \{1, \ldots, d\}$. The region encoded by a given subset E is defined as

$$\partial_E = \{z \in \mathbb{Z}_+^d : z_i = 0, i \in E, z_i > 0, i \notin E\}.$$

The interior of the domain is given by $\partial_\emptyset$ and the origin is represented by $\partial_{\{1,2,\ldots,d\}}$. Subsets other than the empty set represent the "boundaries" of the state-space and correspond to system configurations in which at least one station is empty. The collection of all possible values that the increments of the process Q can take depends on the current region at which Q is positioned. However, in any case, such collection is a subset of

$$\mathbb{V} \triangleq \{e_i, -e_i + e_j, -e_j : i, j = 1, 2, \ldots, d\},$$

where e_i is the vector whose i-th component is one and the rest are zero. An element of the form e_i represents an arrival at station i, an element of the form $-e_i + e_j$ represents a departure from station i that flows to station j and an element of the form $-e_j$ represents a departure from station j out of the system. The set of all possible departures from station i is a subset of

$$\mathbb{V}_i^- \triangleq \{w : w = -e_i \text{ or } w = -e_i + e_j \text{ for some } j = 1, \ldots, d\}.$$

Because of the non-negativity constraints on the boundaries of the system we have to be careful when specifying the transition dynamics. First we define a sequence of i.i.d. random variables $\{Y(k) : k \geq 1\}$ so that for each $w \in \mathbb{V}$

$$\mathbb{P}(Y(k) = w) = \begin{cases} \lambda_i & \text{if } w = e_i, \\ \mu_i P_{ij} & \text{if } w = -e_i + e_j, \\ \mu_i P_{i0} & \text{if } w = -e_i. \end{cases}$$

The dynamics of the queue-length process admit the random walk type representation given by

$$Q(k+1) = Q(k) + \zeta(Q(k), Y(k+1)), \tag{9}$$

where $\zeta(\cdot)$ is the constrained mapping and it is defined for $x \in \partial_E$ via

$$\zeta(x, w) \triangleq \begin{cases} 0 & \text{if } w \in \cup_{i \in E} \mathbb{V}_i^-, \\ w & \text{otherwise.} \end{cases}$$

The large deviations theory associated with Jackson networks is somewhat similar (at least in form) to that of random walks, technical results can be found in [9, 15] and [19]. One has to recognize, of course, that the non-smoothness of the constrained mapping as a function of the state of the system creates substantial technical complications, but we will leave aside this issue in our discussion because our objective is simply to describe the form of the necessary large deviations results for our purposes. An extremely

important role behind the development of large deviations theory for light-tailed random walks is played by the log-moment generating function of the increment distribution. So, given the similarities suggested by the dynamics of (9) and those of a simple random walk it is not surprising that the log-moment generating function of the increments, namely,

$$\psi(x,\theta) \triangleq \log \mathbb{E}\left[\exp\left(\theta^T \zeta(x, Y(k))\right)\right] \tag{10}$$

also plays a crucial role in the large deviations behavior of $p_n^V(x)$ as $n \nearrow \infty$.

In order to understand the large deviations behavior of p_n^V it is useful to scale space by $1/n$, thereby introducing a scaled queue length process $\{Q_n(k) : k \geq 0\}$ which evolves according to

$$Q_n(k+1) = Q_n(k) + \frac{1}{n}\zeta(Q_n(k), Y(k+1)).$$

Suppose that $Q_n(0) = y = x/n$ and note that $T_{\{0\}}$ and T_n^V can also be written as

$$T_{\{0\}} = \inf\{k \geq 1 : Q_n(k) = 0\}, T_n^V = \inf\{k \geq 1 : V(Q_n(k)) \geq 1\}.$$

Note that using the scaled queue length process one can write

$$p_n^V(y) = \mathbb{E}\left[p_n^V(y + \frac{1}{n}\zeta(y, Y(1)))\right]. \tag{11}$$

Here with a slight abuse of notation we use $p_n^V(y)$ to mean

$$\mathbb{P}\left(T_n^V \leq T_{\{0\}} | Q_n(0) = y\right).$$

Large deviations theory dictates that

$$p_n^V(y) = \exp\left(-nW_V(y) + o(n)\right) \tag{12}$$

as $n \nearrow \infty$ for some non-negative function $W_V(\cdot)$. In order to characterize $W_V(\cdot)$ we can combine the previous expression together with (11) and a formal Taylor expansion to obtain

$$\begin{aligned}
1 &= \frac{1}{p_n^V(y)}\mathbb{E}\left[p_n^V(y + \frac{1}{n}\zeta(y, Y(1)))\right] \\
&\approx \mathbb{E}\exp\{-nW_V[y + \frac{1}{n}\zeta(y, Y(1))] + nW_V(y)\} \\
&= \mathbb{E}\exp\{-\partial W_V(y)^T \zeta(y, Y(1)) + o(1)\} \\
&= \exp\left(\psi(y, -\partial W_V(y)) + o(1)\right).
\end{aligned}$$

Sending $n \nearrow \infty$ we formally arrive at the equation

$$\psi\left(y, -\partial W_V\left(y\right)\right) = 0 \tag{13}$$

together with the boundary condition $W_V\left(y\right) = 0$ if $V\left(y\right) \geq 1$. The previous equation is the so-called Isaacs equation which characterizes the large deviations behavior of $p_n^V\left(\cdot\right)$ and it was introduced together with a game theoretic interpretation by Dupuis and Wang in [11]. The solution to (13) is understood in a weak sense (as viscosity solution) because the function $W_V\left(\cdot\right)$ is typically not differentiable everywhere. Nevertheless, it coincides with a certain calculus of variations representation which can be obtained out of the local large deviations rate function for Jackson networks (see [19]).

An asymptotic lower bound for $W_V\left(y\right)$ can be obtained by finding an appropriate subsolution to the Isaacs equation, in which the equality signs in (13) are appropriately replaced by inequalities thereby obtaining a so-called subsolution to the Isaacs equation. In particular, $\overline{W}_V\left(\cdot\right)$ is said to be a subsolution to the Isaacs equation if

$$\psi(y, -\partial\overline{W}_V\left(y\right)) \leq 0 \tag{14}$$

subject to $\overline{W}_V\left(y\right) \leq 0$ if $V\left(y\right) \geq 1$. The subsolution property guarantees $\overline{W}_V\left(y\right) \leq W_V\left(y\right)$, which translates to an asymptotic logarithmic upper bound of $p_n^V\left(y\right)$. The subsolution is said to be maximal at zero if $\overline{W}_V\left(0\right) = W_V\left(0\right)$. Not surprisingly, subsolutions are easier to construct than solutions and, as we shall discuss in the next section, beyond their use in the development of asymptotic upper bounds they can be applied to the design of efficient simulation procedures. The use of subsolutions to the Isaacs equation for the design of efficient simulation algorithms was introduced in [11]. A derivation of the subsolution equation (14) following the same spirit leading to (13) using Lyapunov inequalities is given in [6].

As we mentioned in Section 2, the efficiency analysis of a rare-event simulation estimator depends on the growth rate of its coefficient of variation. We are interested in an asymptotic analysis that goes beyond the error term $\exp(o\left(n\right))$ given by the large deviations approximation (12). So, we must enhance the large deviations approximations in order to provide a more precise estimate for p_n^V. Developing such an estimate is the aim of the following proposition which follows as a consequence of Proposition 3 in Section 5 of this paper (see also Proposition 1 and the analysis in Section 5 in [4]).

PROPOSITION 1. *There exists $K > 0$ (independent of x and n) such that*

$$p_n^V\left(x\right) \leq KP\{V\left(Q\left(\infty\right)\right) = n\}/P\{Q\left(\infty\right) = x\},$$

where Q_∞ is the steady state queue length. Moreover, if $\|x\| \leq c$ for some $c \in (0,\infty)$ then[2]

$$p_n^V(x) = \Omega[P\{V(Q(\infty)) = n\}/P\{Q(\infty) = x\}] \tag{15}$$

as $n \nearrow \infty$.

REMARK. It is important to keep in mind that we shall mostly work with the process $Q(\cdot)$ directly, as opposed to the scaled version $Q_n(\cdot)$ which is used in the analysis of [8].

The previous proposition provides the necessary means to estimate p_n^V up to a constant; we just need to recall that the distribution of $Q(\infty)$ is computable in closed form (see [23] p. 95). In particular, we have that

$$\begin{aligned}\pi(m_1,\ldots,m_d) &= \prod_{j=1}^{d} \mathbb{P}(Q_j(\infty) = m_j) \\ &= \prod_{j=1}^{d} (1-\rho_j)\rho_j^{m_j}, \quad j = 1,\ldots,d, \text{ and } m_j \geq 0.\end{aligned}$$

We shall use $\pi(\cdot)$ to denote the stationary measure of Q. In simple words, the previous equation says that the steady state queue length process has independent components which are geometrically distributed. In particular, $P(Q_j(\infty) = m) = \rho_j^m(1-\rho_j)$ for $m \geq 0$. The next proposition follows directly from standard properties of the geometric distribution (see Proposition 3 in [4]). Before we proceed, it's useful to look at $V(Q(\infty))$ in the following way. Without loss of generality, we assume

$$V(Q(\infty)) = v^T Q(\infty) = Q_{j_1}(\infty) + \cdots + Q_{j_s}(\infty),$$

i.e., $\{j_1, j_2, \ldots, j_s\}$ are the stations encoded by the vector v. Further suppose that we can group these s stations into k groups by their traffic intensities. In other words, stations in $\{i_1^{\{1\}},\ldots,i_{m_1}^{\{1\}}\}$ have traffic intensity equal to ρ_{t_1}, $\ldots$, stations in $\{i_1^{\{k\}},\ldots,i_{m_k}^{\{k\}}\}$ have traffic intensity equal to ρ_{t_k}; and we have $m_1 + \cdots + m_k = s$. Now if we define

$$M_i = Q_{j_1^{\{i\}}}(\infty) + \cdots + Q_{j_{m_i}^{\{i\}}}(\infty),$$

[2]Given two non-negative functions $f(\cdot)$ and $g(\cdot)$, we say $f(n) = \Omega(g(n))$ if there exists $c, n_0 \in (0,\infty)$ such that $f(n) \geq cg(n)$ for all $n \geq n_0$.

then it's clear that the M_i's are negative binomially distributed with parameters m_i and $p_i = 1 - \rho_{t_i}$. Therefore,

$$V(Q(\infty)) = M_1 + \cdots + M_k,$$

is the sum of negative binomial random variables.

PROPOSITION 2. *$P[V(Q(\infty)) = n] = \Theta(e^{-n\gamma_V} n^{\beta_V - 1})$,[3] where $\gamma_V = -\log \rho_*^V$, in which $\rho_*^V = \max\{\rho_i : v_i = 1\}$; and $\beta_V = \sum_i I\{\rho_i = \rho_*^V, v_i = 1\}$ is the number of bottleneck stations in the target subset corresponding to v.*

PROOF. We have just showed that $V(Q(\infty))$ is the sum of negative binomial random variables, so it suffices to show that if $M_1, \ldots, M_k$ are independent random variables so that M_i is negative binomial with parameters (m_i, p_i) and $p_1 < \cdots < p_k$, then

$$\mathbb{P}(M_1 + \cdots + M_k = n) = \Theta(\mathbb{P}(M_1 = n)) \tag{16}$$

as $n \nearrow \infty$; that is, the tail of the probability mass function of the sum of independent negative binomials has the same behavior as the tail of the heaviest terms in the sum (in this case M_1 has the heaviest tail among the M_j's). In turn, it is easy to verify that $\mathbb{P}(M_1 = n) = \Theta((1-p_1)^n n^{m_1 - 1})$, so to show the proposition we just need to verify (16). We proceed by induction in k. First, let us treat the case $k = 2$. Assume that $p_1 < p_2$ and note that

$$\begin{aligned}
& \mathbb{P}(M_1 + M_2 = n) \\
= & \sum_{j=0}^{n} \mathbb{P}(M_1 = n - j)\, \mathbb{P}(M_2 = j) \\
= & \sum_{j=0}^{n} (1-p_1)^{n-j} p_1^{m_1} \binom{m_1 + n - j - 1}{m_1 - 1} (1-p_2)^j p_2^{m_2} \binom{m_2 + j - 1}{m_2 - 1} \\
= & \sum_{j=0}^{n} (1-p_1)^{n-j} (1-p_2)^j \Theta((n-j)^{m_1 - 1} j^{m_2 - 1}) \\
= & (1-p_1)^n n^{m_1 - 1} \sum_{j=0}^{n} \left(\frac{1-p_2}{1-p_1}\right)^j \Theta(j^{m_2 - 1}).
\end{aligned}$$

Since $(1-p_2)/(1-p_1) \in (0,1)$ it follows that the previous sum converges as $n \nearrow \infty$ and therefore we conclude that (16) for $k = 2$. Now we assume that

[3] Given two positive functions $f(\cdot)$ and $g(\cdot)$, recall that $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$.

the claim is valid for some value $k > 2$, we need to verify the claim for $k+1$. Assume without loss of generality that $p_1 < \cdots < p_k < p_{k+1}$ (otherwise re-label the random variables so that the order of the probabilities is as stated). Note that, by induction hypothesis,

$$\begin{aligned}\mathbb{P}(M_1 + \cdots + M_{k+1} = n) &= \sum_{j=0}^{n} \mathbb{P}(M_1 + \cdots + M_k = n-j)\,\mathbb{P}(M_{k+1} = j) \\ &= \Theta\left(\sum_{j=0}^{n} \mathbb{P}(M_1 = n-j)\right)\mathbb{P}(M_{k+1} = j).\end{aligned}$$

The rest of the analysis then proceeds just as in the case of $k = 2$ analyzed earlier, therefore we conclude the proof of the proposition. □

4. The splitting algorithm. The previous section discussed some large deviations properties required to guide the construction of an efficient splitting scheme using the theory developed in the work of Dean and Dupuis [8]. In order to explain the construction suggested by Dean and Dupuis let us first discuss the general idea behind the splitting algorithm that we shall analyze; a variation of which was first applied to Jackson networks by Villen-Altamirano and Villen-Altamirano [20].

The strategy is to divide the state-space into a collection of regions $\{C_j^n : 0 \le j \le l_n(x)\}$ which are nested and that help define "milestone" events that interpolate between the initial position of the process and the target set, which corresponds to the region C_0^n. That is, in our setting we put $C_0^n \triangleq \{x \in \mathcal{S} : V(x) \ge n\}$ and the remaining C_j^n's are placed so that $C_0^n \subseteq C_1^n \subseteq \cdots \subseteq C_{M_n}^n$. How to construct the level sets C_j^n in order to induce efficiency will be discussed below. An observation that is intuitive at this point, however, is that one should have $M_n = \Theta(n)$ so that the next milestone event becomes accessible given the current level. For the moment, let us assume that the C_j^n's have been placed. The splitting algorithm proceeds as follows.

Algorithm SA.

1.– Initiate the simulation procedure with a single particle starting from position $x \in C_k^n$ for a given $k \ge 1$. Let $w_1 = 1$ be the initial weight associated with such particle.
2.– Evolve the initial particle until either it hits $\{0\}$ or it hits level C_{k-1}^n. If the particle hits $\{0\}$, then the particle is said to die. If the particle reaches level C_{k-1}^n then it is *replaced* by r identical particles (for a given integer $r > 1$). The replacing particles are called the immediate

descendants or children of the initial particle, which in turn is said to be their parent. The children are positioned precisely at the place where the parent particle reached level C_{k-1}^n. The weight w_j associated with the j-th children (enumerate the children arbitrarily) has a value equal to the weight of the parent particle multiplied by $1/r$.

3.– The procedure starting from step 1 is replicated for each of the offspring particles in place; carrying over the value of each of the weights at each level for the surviving particles (the weights of the particles that die can be disregarded).

4.– Steps 1 to 3 are repeated until all the particles either die or reach level C_0^n.

Dean and Dupuis in [8] show how to apply large deviations theory to select the C_j^n's in order to obtain a weakly efficient splitting algorithm. One needs to balance the number of the C_j^n's so that it is not unlikely for a given particle to reach the next level while keeping the total number of particles controlled. We now provide a formal motivation for the use of large deviations for constructing the C_j^n's in a balanced way.

It is convenient, as we did in our formal large deviations discussion in the previous section, to consider the scaled process $Q_n(\cdot)$. Let us assume that the splitting mechanism indicated in Algorithm SA is in place and that our initial position is set at level $Q(0) = x$, so that $Q_n(0) = y = x/n$. The C_j^n's are typically constructed in terms of the level sets of a so-called importance function which we shall denote by $U(\cdot)$. In particular, put $D_n \triangleq \{y \in n^{-1}\mathcal{S} : V(y) < 1\}$ and set $C_j^n = nL_{z_n(j)}$, where

$$L_z \triangleq \{y \in D_n : U(y) \leq z\}, \tag{17}$$

and the $z_n(j)$'s are appropriately chosen momentarily. Then, define

$$l_n(x) = \min\{j \geq 0 : x \in C_j^n\} = \min\{j \geq 0 : y \in L_{z_n(j)}\}. \tag{18}$$

The total weight corresponding to a particle that reaches level C_0^n given that it started at level $l_n(x)$ is $r^{-l_n(x)}$. In order to have at least a weakly efficient algorithm we wish to achieve two constraints. The first one imposes the aggregate weight of a particle reaching level C_0^n to be $p_n^V(x)\exp(-o(n))$; this would guarantee that the second moment of the resulting estimator achieves asymptotic optimality. The second constraint dictates that the expected number of particles that make it to C_0^n, which is roughly $r^{l_n(x)}p_n^V(x)$ exhibits subexponential growth (i.e. $\exp(o(n))$); this would guarantee a cost per replication that is subexponential. Note that both constraints lead to

the requirement of $r^{l_n(x)}p_n^V(x) = \exp(o(n))$. So, given a subsolution $\overline{W}_V(\cdot)$ to the corresponding Isaacs equation, which implies that

$$p_n^V(x) \leq \exp\left(-n\bar{W}_V(x/n) + o(n)\right),$$

it suffices to ensure that

$$l_n(x)\log(r) - n\overline{W}_V(x/n) = o(n). \tag{19}$$

The behavior of $l_n(x)$ as $n \nearrow \infty$ only relates to the properties of the function $U(\cdot)$ and it is really independent of the large deviations behavior of the system. In particular, picking $z_n(j) = \Delta j/n, \Delta > 0$ yields $l_n(x) = \lceil nU(x/n)/\Delta \rceil$ and therefore, equation (19) suggests that one should select $U(y) = \Delta\overline{W}_V(y)/\log(r)$ with $\overline{W}_V(0) = W_V(0)$ in order to obtain a weakly efficient estimator for p_n^V. This is precisely the conclusion obtained in the work of [8] who present a rigorous analysis that justifies the previous heuristic discussion. Our development in the next section will sharpen the efficiency properties of the sampler proposed in [8] when applied to Jackson networks. So, we content ourselves with the previous heuristic motivation for the splitting method that we will analyze in the next section and which in turn is based on the viscosity subsolution given by

$$\bar{W}_V(y) = \varrho^T y - \log\rho_*^V, \tag{20}$$

where $\varrho_i = \log\rho_i$ for $i = 1, \ldots, d$, see e.g., [12] and [8].

We close this section with a precise definition of the estimator that we will analyze. First, given a constant $\Delta > 0$ (the level size) define $\bar{W}_V(\cdot)$ as indicated in (20) for each $y = x/n$ with $x \in \mathcal{S}$. Then, select an integer $r > 1$ and define $U(y) = \Delta\overline{W}_V(y)/\log(r)$. Given the initial position x define the sets $\{C_j^n : 1 \leq j \leq l_n(x)\}$ as indicated above (see equation (18)). Run Algorithm SA and let N_n be the number of particles that survive up to C_0^n; their corresponding final weight is $1/r^{l_n(x)}$. Our estimator for $p_n^V(x)$ is simply

$$R_n(x) = N_n(x)/r^{l_n(x)}. \tag{21}$$

Now, for the sake of analytical convenience, when analyzing the second moment of $R_n(x)$ we will adopt the so-called *fully branching* representation of the previous estimator (see [8]). Such fully branching representation is obtained by splitting death particles at level zero. In particular, we modify *Algorithm SA* to obtain the following algorithm:

Algorithm SFB.

1.– Initiate the simulation procedure with a single particle starting from position $x \in C_k^n$ for a given $k \geq 1$. Let $w_1 = 1$ be the initial weight associated with such particle.
2.– Evolve the initial particle until it either hits $\{0\}$ (and die) or hits level C_{k-1}^n (remain active or alive), *in either case* the particle becomes the parent and is replaced by r descendants, positioned where the parent is located (either $\{0\}$ or the location where it enters level C_{k-1}^n). The weight of the j-th particle is set to equal the weight of its parent multiplied by $1/r$.
3.– For each *living* offspring particle, the procedure starting from step 1 is replicated. For each *dead* offspring particle, replace it by r descendants, set the weight of each child to be that of the parent multiplied by $1/r$.
4.– Steps 1 to 3 are repeated until all the particles either die or reach level C_0^n.

In other words, after $l_n(x)$ iterations we have $r^{l_n(x)}$ total particles labeled $1, 2, \ldots, r^{l_n(x)}$, each with weight $1/r^{l_n(x)}$. We define I_j as the indicator function of the event that the j-th particle is in C_0^n so that $N_n(x) = \sum_{j=1}^{r^{l_n(x)}} I_j$. The fully branching representation of $R_n(x)$ is simply

$$R_n(x) = r^{-l_n(x)} \sum_{j=1}^{r^{l_n(x)}} I_j. \tag{22}$$

5. Analysis of splitting estimators. We are now in a good position to perform a refined efficiency analysis for the estimator $R_n(x)$. We shall break our analysis into two parts. The first part corresponds to the expected number of particles generated per run and the second part deals with the second moment of $R_n(x)$. We establish upper bounds on both quantities that enable us to reach the conclusion that this multilevel splitting algorithm substantially outperforms the direct polynomial time algorithm for solving the associated system of linear equations.

Our analysis takes advantage of the time reversed process associated with the underlying Jackson network which we shall now define. Given the transition matrix $\{K(x,y) : x, y \in \mathcal{S}\}$ of the process Q, we define the reversed Markov chain $\tilde{Q} = \{\tilde{Q}(k) : k \geq 0\}$ via the transition matrix $\tilde{K}(\cdot)$:

$$\tilde{K}(y,x) = K(x,y)\,\pi(x)/\pi(y),$$

for $x, y \in \mathcal{S}$. It turns out that $\tilde{Q}$ also describes the queue length process of an open stable Jackson network with stationary distribution equal to $\pi(\cdot)$,

(see [23] p. 95). We will use $\tilde{P}_x(\cdot)$ to denote the probability measure in path space associated with $\tilde{Q}$ given that $\tilde{Q}(0) = x$.

The following result is similar to that of Proposition 1 in [4]. However, our representation in (23) is slightly more useful for our purposes.

PROPOSITION 3.

$$p_n^V(x) = \frac{\tilde{\mathbb{P}}_\pi\left(\tilde{Q}(0) \in C_0^n, \tilde{T}_{\{x\}} \leq \tilde{T}_{\{0\}}, \tilde{T}_{\{x\}} < \tilde{T}_n^V\right)}{\pi(x) P_x(T_{\{x\}} \geq T_n^V \wedge T_{\{0\}})} \tag{23}$$

$$= \frac{\tilde{\mathbb{P}}_\pi\left(\tilde{Q}(0) \in C_0^n, \tilde{\sigma}_{\{x\}} < \tilde{T}_{\{0\}} < \tilde{T}_n^V\right)}{\pi(0) P_0\left(\sigma_{\{x\}} < T_n^V \wedge T_{\{0\}}\right)} \tag{24}$$

where $\tilde{T}_n^V = \inf\{k \geq 1 : V(\tilde{Q}(k)) \geq n\} = \inf\{k \geq 1 : \tilde{Q}(k) \in C_0^n\}$, $\tilde{T}_{\{x\}} = \inf\{k \geq 1 : \tilde{Q}(k) = x\}$, $\sigma_{\{x\}} \triangleq \inf\{k \geq 0 : Q(k) = x\}$ *and* $\tilde{\sigma}_{\{x\}} \triangleq \inf\{k \geq 0 : \tilde{Q}(k) = x\}$. *Moreover, there exists* $\delta > 0$ *(independent of* $x \neq 0$ *and* n*) such that*

$$P_x(T_{\{x\}} \geq T_n^V \wedge T_{\{0\}}) \geq \delta. \tag{25}$$

PROOF. We assume that $x \neq 0$. The case $x = 0$ is included in the analysis of (24). First, we observe that

$$\begin{aligned} p_n^V(x) &= \mathbb{P}_x\left(T_n^V < T_{\{0\}}, T_{\{x\}} < T_n^V \wedge T_{\{0\}}\right) + \mathbb{P}_x\left(T_n^V < T_{\{0\}}, T_{\{x\}} \geq T_n^V \wedge T_{\{0\}}\right) \\ &= p_n^V(x)\, \mathbb{P}_x\left(T_{\{x\}} < T_n^V \wedge T_{\{0\}}\right) + \mathbb{P}_x\left(T_n^V < T_{\{0\}}, T_{\{x\}} \geq T_n^V \wedge T_{\{0\}}\right). \end{aligned}$$

Therefore,

$$p_n^V(x) = \frac{\mathbb{P}_x\left(T_n^V < T_{\{0\}}, T_{\{x\}} \geq T_n^V \wedge T_{\{0\}}\right)}{\mathbb{P}_x\left(T_{\{x\}} \geq T_n^V \wedge T_{\{0\}}\right)}.$$

Following the same technique as in Proposition 1 in [4] we have that

(26)

$$\begin{aligned} &\pi(x)\, \mathbb{P}_x\left(T_n^V < T_{\{0\}}, T_{\{x\}} \geq T_n^V \wedge T_{\{0\}}\right) \\ &= \sum_{k=0}^{\infty} \pi(x)\, \mathbb{P}_x\left(T_n^V < T_{\{0\}}, T_{\{x\}} \geq T_n^V \wedge T_{\{0\}}, T_n^V = k\right) \\ &= \sum_{k=1}^{\infty} \pi(x) \sum_{y_0 = x, y_1, .., y_{k-1} \in \mathcal{S} \setminus \left(\{0, x\} \cup C_0^n\right), y_k \in C_0^n} K(y_0, y_1) \times \cdots \times K(y_{k-1}, y_k) \\ &= \sum_{k=1}^{\infty} \sum_{y_0 = x, y_1, .., y_{k-1} \in \mathcal{S} \setminus \left(\{0, x\} \cup C_0^n\right), y_k \in C_0^n} \tilde{K}(y_1, y_0) \times \cdots \times \tilde{K}(y_k, y_{k-1})\, \pi(y_k). \end{aligned}$$

Letting $\tilde{y}_i = y_{k-i}$ for $i = 1, \ldots k$ we see that the summation in each of the terms above ranges over paths $\tilde{y}_0, \ldots, \tilde{y}_k$ satisfying that $\tilde{y}_0 \in C_0^n$, $\tilde{T}_{\{x\}} = k$ (so in particular $\tilde{y}_k = x$) and also that $\tilde{T}_{\{0\}} \geq k, \tilde{T}_n^V > k$. So, we can interpret the previous sum as

$$\tilde{\mathbb{P}}_\pi \left(\tilde{Q}\left(0\right) \in C_0^n, \tilde{T}_{\{x\}} \leq \tilde{T}_{\{0\}}, \tilde{T}_{\{x\}} < \tilde{T}_n^V \right).$$

This yields part (23). Part (24) corresponds to Proposition 1 of [4]; it follows using the same trick as in the analysis of display (26), after multiplying and dividing by $\pi\left(0\right)$ when computing the probability of going from zero to the target set via the point x. The most interesting part is the bound (25), which is essentially the argument in Proposition 7 of [4], but we discuss it here to make our exposition self contained. We need to show that there exists $\delta > 0$, such that $\mathbb{P}_x(T_{\{x\}} \geq T_n^V \wedge T_{\{0\}}) \geq \delta$ uniformly over $x \neq 0$. The strategy follows the following steps: 1) Argue first that the probability is positive if $x \neq 0$ and, therefore, bounded away from zero over compact sets in x, 2) Now consider the case in which x is outside a suitably defined compact set, then argue that by intersecting with an event involving finitely many service times and routing events inside the network, we can reach a system configuration with m_1 *fewer* customers in the system than the total number initially present in configuration x, 3) Finally, once we have m_1 fewer customers, argue, using the stability of the Jackson network, that with high probability, the system will eventually empty before coming back to *any* configuration with as many customers as the initial configuration x. Thus, effectively our plan is to show that

$$\inf_{x: x \neq 0} \mathbb{P}_x(T_{\{x\}} \geq T_n^V \wedge T_{\{0\}}) \geq \delta.$$

We now proceed to carry over the previous program. First, if $x \neq 0$, we must clearly have that $\mathbb{P}_x(T_{\{x\}} \geq T_{\{0\}}) > 0$ (i.e. for each $x \neq 0$, the event $T_{\{x\}} > T_{\{0\}}$ is a possible event). To see this, we argue as follows. Note that we have an open Jackson network, so each customer in the system must eventually leave the system if no arrivals are allowed to enter the network. So, if we intersect with the event that the next inter-arrival time into the system is sufficiently large (which clearly is an event with positive probability), we can work *only* with the current customers inside the network, which are distributed in each of the stations according to the state of the system x. Let us use $||x||$ to denote the L_1 norm of x (since the components of x are non-negative, $||x||$ is just the sum of the components of x). If $||x|| \leq m_0$ for some constant m_0, we can always construct an event with the property that, given the initial configuration of the system x, everybody leaves the

network prior to an arrival *and* before we find the network once again in the initial configuration x. Observe that if we are forced to cycle back to the initial configuration x with probability one assuming that no arrivals are allowed into the system, then it would *not* be true that each customer must eventually leave the system and this violates the condition that the network is open. Therefore, since the set of configurations x such that $||x|| \leq m_0$ is finite we can find $\delta_0 > 0$ (possibly depending on m_0) such that

$$\inf_{x:x\neq 0,||x||\leq m_0} \mathbb{P}_x(T_{\{x\}} \geq T_{\{0\}}) \geq \delta_0. \tag{27}$$

Now, we proceed with part 2) of the program. Let us assume that $||x|| > m_0$ for $m_0 > 0$ chosen momentarily. Following the same type of reasoning described earlier we have that if $m_1 < m_0$, then we can find $\delta_1 > 0$ (possibly depending on m_1) such that

$$\inf_{||x||\geq m_0} \mathbb{P}_x(T_{\{x\}} \geq T_{||x||-m_1}) > \delta_1,$$

where $T_{||x||-m_1} = \inf\{k \geq 1 : ||Q(k)|| = ||x|| - m_1\}$. In simple words, we can make sure that m_1 customers leave the system prior to an arrival and prior to cycling back to configuration x, regardless of the initial configuration x; this is done by intersecting with an event that depends on the order in which finitely many services are completed and jobs are routed through the network. Therefore, we have that

$$\begin{aligned}\mathbb{P}_x(T_{\{x\}} \geq T_{\{0\}}) &\geq \mathbb{P}_x(T_{\{x\}} \geq T_{\{0\}}, T_{\{x\}} \geq T_{||x||-m_1}) \\ &\geq \delta_1 \inf_{\xi:||\xi||=||x||-m_1} \mathbb{P}_\xi(T_{||x||} \geq T_{\{0\}}).\end{aligned}$$

Finally, we proceed with step 3) of the program, namely, arguing that if m_1 is chosen sufficiently large, then one can actually find $\varepsilon > 0$ such that

$$\sup_{\xi:||\xi||=||x||-m_1} \mathbb{P}_\xi(T_{||x||} < T_{\{0\}}) < 1 - \varepsilon. \tag{28}$$

Let $\widetilde{N} = ||x||$ and assume that ξ is such that $||\xi|| = \widetilde{N} - m_1$. We observe that if $\delta_2 > 0$ is chosen small enough, then

$$\mathbb{P}_\xi(T_{||x||} < T_{\{0\}}) = \mathbb{P}_\xi(T_{||x||} < T_{\{0\}}, T_{||x||} \leq \widetilde{N}\delta_2) + \mathbb{P}_\xi(T_{||x||} < T_{\{0\}}, T_{||x||} > \widetilde{N}\delta_2). \tag{29}$$

Now, note that

$$\mathbb{P}_\xi(T_{||x||} < T_{\{0\}}, T_{||x||} > \widetilde{N}\delta_2) = \mathbb{E}_\xi[I(T_{||x||} > \widetilde{N}\delta_2)\mathbb{P}_{Q(\widetilde{N}\delta_2)}(T_{||x||} < T_{\{0\}})]. \tag{30}$$

Given the initial configuration ξ, large deviation results for Jackson networks (see [15]) guarantee that for any $\epsilon_0 > 0$,

$$\mathbb{P}_\xi\left(||Q(\tilde{N}\delta_2) - \tilde{N}q(\delta_2)|| > \tilde{N}\epsilon_0\right) = \exp\left(-\tilde{N}I(\epsilon_0) + o(\tilde{N})\right),$$

as $\tilde{N} \nearrow \infty$ for some $I(\epsilon_0) > 0$ and some $q(\delta_2)$ (which corresponds to the fluid limit evaluated at δ_2). In the language of large deviations, the fluid limit corresponds to the zero-cost trajectory. And trajectories outside of the band that centers on the fluid limit have probabilities that decay exponentially fast. Moreover, since the network is stable and open, we have that $||q(\delta_2)|| < 1 - \delta_3$ for some $\delta_3 > 0$. Therefore, once again appealing to the large deviations results of [15], we obtain that if $\epsilon_0 < \delta_3$, then

$$\begin{aligned}\sup_{\{q:||q-q(\delta_2)||<\epsilon_0\}} \mathbb{P}_{\tilde{N}q}\left(T_{||x||} < T_{\{0\}}\right) \leq & \sup_{\{q:||q||\leq 1-\delta_3+\epsilon_0<1\}} \mathbb{P}_{\tilde{N}q}\left(T_{\tilde{N}} < T_{\{0\}}\right) \\ =& O\left(e^{-\delta\tilde{N}}\right),\end{aligned}$$

for some $\delta > 0$. Consequently,

$$\begin{aligned}&\mathbb{E}_\xi\left(I\big(T_{||x||} > \tilde{N}\delta_2\big)\mathbb{P}_{Q(\tilde{N}\delta_2)}\left(T_{||x||} < T_{\{0\}}\right)\right) \\ &\leq \mathbb{P}\left(||Q\big(\tilde{N}\delta_2\big) - \tilde{N}q(\delta_2)|| > \epsilon_0\tilde{N}\right) + \sup_{\{q:||q||\leq 1-\delta_3+\epsilon_0<1\}} \mathbb{P}_{\tilde{N}q}\left(T_{||x||} < T_{\{0\}}\right) \\ &= O\left(e^{-\delta\tilde{N}}\right),\end{aligned}$$

for some $\delta > 0$. Therefore the right hand side of (30) decreases exponentially fast in $\tilde{N}$. It suffices then to study the first term in (29). Note that

$$\begin{aligned}\mathbb{P}_\xi(T_{||x||} < T_{\{0\}}, T_{||x||} \leq \widetilde{N}\delta_2) &\leq \mathbb{P}_\xi(\cup_{k\leq\widetilde{N}\delta_2}\{||Q(k)|| \geq \widetilde{N}\}) \\ &\leq \sum_{k\leq\widetilde{N}\delta_2} \mathbb{P}_\xi(||Q(k)|| \geq \widetilde{N}).\end{aligned} \tag{31}$$

We will apply a Chernoff-bound argument to bound the right hand side of the previous display. Fix an integer $m_3 > 0$ and write $k = m_3 s + l$ for some integer $s \geq 0$ and $l \in \{0, 1, \ldots, m_3 - 1\}$. Let $Q(0) = \xi$ and note that

$$\begin{aligned}||Q(k)|| - ||\xi|| &= ||Q(m_3 s + l)|| - ||Q(m_3 s)|| \\ &\quad + \sum_{j=0}^{s-1}[||Q(m_3(j+1))|| - ||Q(m_3 j)||].\end{aligned}$$

Because the network is stable it follows that one can choose $m_3 > 0$ (depending only on the characteristics of the network) so that if $||z|| \geq \widetilde{N}(1-2\delta_2) > m_3$, then

$$\mathbb{E}_z[||Q\,(m_3)\,|| - ||z||] \leq -\varepsilon_1.$$

In simple words, if the initial population is very large, on average we shall expect more customers to leave than those who arrive. Clearly, one also has that $||Q\,(m_3)\,|| - ||z|| \leq m_3$ (at most m_3 people leave or arrive in m_3 transitions of the network), so we have that one can compute a constant $m_4 > 0$, uniform in z as long as $||z|| \geq \widetilde{N}(1-2\delta_2) > m_3$ such that

$$\log \mathbb{E}_z \exp(\theta[||Q\,(m_3)\,|| - ||z||]) \leq -\varepsilon_1\theta + m_4\theta^2.$$

So, selecting $\theta^* > 0$ sufficiently small we obtain that

$$\log \mathbb{E}_z \exp(\theta^*[||Q\,(m_3)\,|| - ||z||]) \leq -\varepsilon_1\theta^*/2. \tag{32}$$

Now we are in good shape to apply the Chernoff-bound argument. Note that

$$\begin{aligned}\mathbb{P}_\xi(||Q\,(k)\,|| \geq \widetilde{N}) &\leq \mathbb{P}_\xi(||Q\,(k)\,|| - ||\xi|| \geq m_1)\\ &\leq \exp\,(-\theta^* m_1)\exp(\theta^* m_3)\\ &\quad\cdot \mathbb{E}_\xi\left(\theta^* \exp\left(\sum_{j=0}^{s-1}[||Q\,(m_3(j+1))\,|| - ||Q\,(m_3 j)\,||]\right)\right).\end{aligned}$$

Note that we can apply (32) repeatedly to estimate the exponential of the the expectation in the previous display given that $||\xi|| = \widetilde{N} - m_1$ and that $k \leq \widetilde{N}\delta_2$, which in particular (because Jackson networks increase or decrease by at most one unit in each transition, and recall that $\widetilde{N}$ is large, so that $m_1 < \widetilde{N}\delta_2$), implies that $||Q\,(k)|| \geq \widetilde{N}(1-2\delta_2)$ if $k \leq \widetilde{N}\delta_2$. Therefore, we obtain that

$$\begin{aligned}\mathbb{P}_\xi(||Q\,(k)\,|| \geq \widetilde{N}) &\leq \exp\,(-\theta^* m_1)\exp(\theta^* m_3)\exp(-s\varepsilon_1\theta^*/2)\\ &= \exp\,(-\theta^*(m_1 - m_3))\exp(-[k/m_3]\varepsilon_1\theta^*/2).\end{aligned}$$

Adding over k and choosing m_1 sufficiently large we conclude that the right hand side of (31) can be made arbitrarily small. (Note that having selected m_1, we then choose $m_0 > m_1$ in the discussion following (27)). This combined with our analysis for (30) allows us to conclude (28) and therefore we conclude our result. □

Proposition 1 and 2 from Section 3 follow as a consequence of this result, the rest of the details are given in Section 5 of [4]. Nevertheless, in the interest

of making this paper as self-contained as possible, without compromising its length, we mention that the most difficult part remaining in Proposition 1 involves the lower bound in equation (15). For this part, one can use identity (24) combined with a similar analysis behind (25) to show that there exists $\delta > 0$ such that for all n large enough

$$\tilde{\mathbb{P}}_\pi \left(\tilde{\sigma}_{\{x\}} < \tilde{T}_{\{0\}} < \tilde{T}_n^V | \tilde{Q}\left(0\right) \in C_0^n \right) \geq \delta.$$

The rest of the argument behind Proposition 1 and 2 from Section 3 then follows from elementary properties of the steady-state distribution $\pi\left(\cdot\right)$.

Given the subsolution we proposed in Section 4, the importance function can be written as

$$\begin{aligned} U\left(x/n\right) = \bar{W}_V\left(x/n\right) \frac{\Delta}{\log r} &= \left(\frac{1}{n} \varrho^T x - \log \rho_*^V \right) \frac{\Delta}{\log r} \\ &= C \left(\Delta - \frac{1}{n} \alpha^T x \Delta \right), \end{aligned} \tag{33}$$

where $C = -\log \rho_*^V / \log r$, and $\alpha = \varrho / \log \rho_*^V$. The level index function also simplifies to

$$l_n\left(x\right) = \left\lceil \frac{nU\left(x/n\right)}{\Delta} \right\rceil = \left\lceil nC \left(1 - \frac{1}{n} \alpha^T x \right) \right\rceil = \lceil C(n - \alpha^T x) \rceil. \tag{34}$$

We shall first look at the expected number of surviving particles of the splitting algorithm which characterizes the stability of the algorithm. One shall keep in mind that when the complexity of the splitting algorithm is concerned, what actually matters is the total function evaluation involved in each run. An upper bound is obtained for this quantity, as measured by the sum of all particles generated at interim levels weighted by the maximum remaining function evaluations associated with each of them. We first have the following result.

PROPOSITION 4. *The expected terminal number of particles for the splitting algorithm specified by* (Δ, U) *above satisfies*

$$\mathbb{E}\left[N_n\left(x\right)\right] = \Theta\left(n^{\beta_V - 1} \right) \tag{35}$$

where β_V*, introduced in Proposition 2, denotes the number of bottleneck stations corresponding to the vector* v*.*

PROOF. It can be seen from the *fully-branching* algorithm that

$$\mathbb{E}\left[N_n\left(x\right)\right] = r^{l_n(x)}\, p_n^V\left(x\right).$$

From Proposition 2 we know that $p_n^V(x) = \Theta(\pi^{-1}(x)e^{-\gamma_V n}n^{\beta_V-1})$. Since $e^{-\gamma_V} = e^{\log \rho_*^V} = e^{-C\log r} = r^{-C}$, we can write $p_n^V(x) = \Theta(\pi^{-1}(x)r^{-nC}n^{\beta_V-1})$. Hence, plug in $l_n(x) = \lceil C(n-\alpha^T x)\rceil$, and note that $\pi^{-1}(x) = \tilde{c}r^{C\alpha^T x}$ for some positive constant $\tilde{c}$, we have

$$\mathbb{E}\left[N_n\left(x\right)\right] = \Theta\left(r^{C\alpha^T x}r^{-nC}n^{\beta_V-1}r^{\lceil C\left(n-\alpha^T x\right)\rceil}\right) = \Theta\left(n^{\beta_V-1}\right). \qquad \square$$

As pointed out earlier, the number of terminal surviving particles, although a reasonable proxy to measure the stability of the algorithm, is not suitable for quantifying the complexity. We also need to take into account the number of function evaluations required to generate $R_n(x)$. The next result addresses precisely this issue.

PROPOSITION 5. *The expected computational effort per run required to generate a single replication of* $R_n(x)$ *is* $O(n^{\beta_V+1})$.

To prove this, we need the following result, which upper bounds the probability that a particle makes it to the level $C^n_{l_n(x)-m}$. We first state the result and postpone the proof until after the proof of Proposition 5.

PROPOSITION 6. *For a given generation* m*, denote by* $Q_{m,j}$ *the position of the* j*-th particle, then*

$$\mathbb{P}_x\left(Q_{m,1} \in C^n_{l_n(x)-m}\right) = O\left(\left(\frac{m-1}{C}\right)^{\beta_V-1}\left(\rho_*^V\right)^{\frac{m-1}{C}}\right). \tag{36}$$

Given this result, we now proceed to prove Proposition 5.

PROOF OF PROPOSITION 5. Let N_m^n, $m = 0,\ldots,l_n(x)$, be the number of particles that survive to level $C^n_{l_n(x)-m}$. Again fully-branching algorithm allows us to write

$$\mathbb{E}[N_m^n] = r^m\mathbb{P}_x\left(Q_{m,1} \in C^n_{l_n(x)-m}\right).$$

Thanks to Proposition 6, along with $\left(\rho_*^V\right)^{-1/C} = r$, we have

$$\mathbb{E}[N_m^n] = O\left(r^m\left(\frac{m-1}{C}\right)^{\beta_V-1}\left(\rho_*^V\right)^{\frac{m-1}{C}}\right) = O\left(r\left(\frac{m-1}{C}\right)^{\beta_V-1}\right). \tag{37}$$

Also let $\eta_{m,j}$ be the remaining computational effort of the j-th particle at the start of the m-th level until it either reaches the next level or it dies out. Put $\bar{\eta}_{m,j}(x_j)$ to be the expectation of $\eta_{m,j}$ given that the position of

the j-th particle at the start of level m is x_j. Note that the norm of the position of x_j is less than $c \cdot m$ for a given constant c that depends on the traffic intensities of the system but not on the position of the particle per-se. Therefore, it is easy to see that

$$\sup_{1\le j\le N_m^n} \bar{\eta}_{m,j}(x_j) \le c \cdot m, \tag{38}$$

for some $c \in (0,\infty)$. Intuitively, each particle at level m either advances to the next level, or it dies out by hitting the zero level before moving to the next one, since it takes $\Theta(1)$ work to cross one single layer, $\eta_{m,j}$ is dominated by the work required to die out, and hence its mean is bounded from above by $c \times m$ for some constant c. Using (37) and (38), we can bound the expected total work per run as follows

$$\begin{aligned}
\mathbb{E}\left[\sum_{m=0}^{l_n(x)-1}\sum_{j=1}^{N_m^n}\eta_{m,j}\right] &= \sum_{m=0}^{l_n(x)-1}\mathbb{E}\left[\sum_{j=1}^{N_m^n}\bar{\eta}_{m,j}(x_j)\right] \\
&\le \sum_{m=0}^{l_n(x)-1}\mathbb{E}[N_m^n]\cdot c\cdot m \\
&\le c'\cdot\sum_{m=0}^{l_n(x)-1}\left(\frac{m-1}{C}\right)^{\beta_V-1} m \\
&= O\big(n^{\beta_V+1}\big),
\end{aligned}$$

for some positive constant c and c' where in the last step we use the definition of $l_n(x)$ given in (34). □

It remains to prove Proposition 6.

Proof of Proposition 6. We begin the proof with an important property implied by the splitting algorithm:

$$\begin{aligned}
V(Q_{m,1}) > 0 &\Leftrightarrow Q_{m,1} \in C^n_{l_n(x)-m} = nL_{(l_n(x)-m)\Delta/n} \\
&\Leftrightarrow Q_{m,1} \in \{z \in nD_n : U(z/n) \le (l_n(x)-m)\Delta/n\} \\
&\Leftrightarrow Q_{m,1} \in \left\{z \in nD_n : C\left(1-\frac{1}{n}\alpha^T z\right)\right. \\
&\qquad\qquad \left. \le \frac{1}{n}\left(C\left(n-\alpha^T x\right)-m+1\right)\right\} \\
&\Leftrightarrow Q_{m,1} \in \{z \in nD_n : \alpha^T z \ge \alpha^T x + \frac{m-1}{C}\} \\
&\Leftrightarrow Q_{m,1} \in \{z \in nD_n : \varrho^T z \le \varrho^T x - (m-1)\log r\}
\end{aligned} \tag{39}$$

where we used the representations of $U(\cdot)$ and $l_n(x)$ in (33) and (34) and the definition of L_z in (17). In other words, if a particle survives m generations then its current position is beyond the mth level, which implies that the weighted sum of system population, with weight given by the vector ϱ, is bounded from above by that of the initial position adjusted by a linear function in m. If we define the stopping time $\hat{T}_{\frac{m}{C}} \triangleq \inf\{k \geq 1 : \alpha^T Q(k) \geq \alpha^T x + \frac{m-1}{C}\} = \inf\{k \geq 1 : \varrho^T Q(k) \leq \varrho^T x - (m-1)\log r\}$, the above property also implies that $Q_{m,1} \in C^n_{l_n(x)-m} \Leftrightarrow \hat{T}_{\frac{m}{C}} < T_0$. Following an argument similar to the proof of (23) in Proposition 3 (in fact easier because here we are interested in an upper bound only), it follows that there exists constant $\hat{c} > 0$, independent of x and m, such that

$$\begin{aligned}\mathbb{P}_x\left(Q_{m,1} \in C^n_{l_n(x)-m}\right) &= \mathbb{P}_x\left(\hat{T}_{\frac{m}{C}} < T_0\right)\\ &\leq \frac{\hat{c}}{\pi(x)}\mathbb{P}\left[\varrho^T Q(\infty) \leq \varrho^T x - (m-1)\log r\right]\\ &= \frac{\hat{c}}{\pi(x)}\mathbb{P}\left[\alpha^T Q(\infty) \geq \alpha^T x + \frac{(m-1)}{C}\right].\end{aligned}$$

To finish the proof we need the following Lemma.

LEMMA 1.

$$\begin{aligned}\mathbb{P}\left[\alpha^T Q(\infty) \geq \alpha^T x + \frac{(m-1)}{C}\right] &= \Theta\left[\mathbb{P}\left(Z\left(\beta_V, 1-\rho_*^V\right) \geq \alpha^T x + \frac{m-1}{C}\right)\right]\\ &= \Theta\left[\left(\frac{m-1}{C}\right)^{\beta_V - 1}\left(\rho_*^V\right)^{\frac{m-1}{C}}\right]\end{aligned}$$

where $Z(n,p)$ denotes a $NBin(n,p)$ (negative binomial) random variable.

PROOF OF LEMMA. Note that

$$\begin{aligned}\alpha^T Q(\infty) &= Q(\infty)^T \frac{\varrho}{\log \rho_*^V}\\ &= \sum_{i=1}^d Q_i(\infty) I\left(\rho_i = \rho_*^V\right) + \sum_{i=1}^d Q_i(\infty) I\left(\rho_i \neq \rho_*^V\right)\frac{\log \rho_i}{\log \rho_*^V}\\ &= Z\left(\beta_V, 1-\rho_*^V\right) + W.\end{aligned}$$

One direction is elementary, since $\alpha^T Q(\infty) \geq Z\left(\beta_V, 1-\rho_*^V\right)$, we clearly have

(40)
$$\mathbb{P}\left[\alpha^T Q(\infty) \geq \alpha^T x + \frac{(m-1)}{C}\right] \geq \mathbb{P}\left[Z\left(\beta_V, 1-\rho_*^V\right) \geq \alpha^T x + \frac{(m-1)}{C}\right].$$

For the other direction, note that there exists constants $c_4 > 0$, and $\tilde{\rho} < \rho_*^V$ such that

$$\begin{aligned} W &= \sum_{i=1}^{d} Q_i(\infty) I\left(\rho_i \neq \rho_*^V\right) \frac{\log \rho_i}{\log \rho_*^V} \\ &\leq c_4 \sum_{i=1}^{d} Q_i(\infty) I\left(\rho_i \neq \rho_*^V\right) \\ &\leq_{st} c_4 Z\left(d - \beta_V, 1 - \tilde{\rho}\right), \end{aligned}$$

where " $\leq_{st}$" denotes that the left hand side is stochastically dominated by the right hand side. As a result,

$$\alpha^T Q(\infty) \leq_{st} Z\left(\beta_V, 1 - \rho_*^V\right) + c_4 Z\left(d - \beta_V, 1 - \tilde{\rho}\right).$$

But since $1 - \rho_*^V < 1 - \tilde{\rho}$, a similar argument as given in the proof of Proposition 2 allows us to obtain

(41)
$$\mathbb{P}\left[\alpha^T Q(\infty) \geq \alpha^T x + \frac{(m-1)}{C}\right] \leq c_0 \mathbb{P}\left[Z\left(\beta_V, 1 - \rho_*^V\right) \geq \alpha^T x + \frac{(m-1)}{C}\right],$$

for some finite constant c_0 that is independent of m. Combining (40) and (41), we have

(42)
$$\begin{aligned} &\mathbb{P}\left[\alpha^T Q(\infty) \geq \alpha^T x + \frac{(m-1)}{C}\right] \\ = \;& \Theta\left[\mathbb{P}\left(Z\left(\beta_V, 1 - \rho_*^V\right) \geq \alpha^T x + \frac{(m-1)}{C}\right)\right]. \end{aligned}$$

Using again Proposition 3 of [4], we reach the conclusion that

$$\mathbb{P}\left[\alpha^T Q(\infty) \geq \alpha^T x + \frac{(m-1)}{C}\right] = \Theta\left[\left(\frac{m-1}{C}\right)^{\beta_V - 1} \left(\rho_*^V\right)^{\frac{m-1}{C}}\right]$$

□

The result of Proposition 6 directly follows. □

To facilitate the analysis of the second moment of $R_n(x)$ we add the following notations. We follow the analysis in [8] to make our exposition here self-contained. For a given generation m, denote by $Q_{m,j}$ the position of the j-th particle; recall that the accumulated weight up to the m-th stage of such a particle is r^m. Let $\chi_{m,j}$ be the disjoint grouping of particles in the

next generation (i.e., $m+1$) according to their "parents" in generation m. For $k \in \chi_{m,j}$, denote by d_k the offsprings of this particle at the final stage $l_n(x)$. We then have the following expansion of the second moment of $R_n(x)$:

(43)
$$\begin{aligned}
&\mathbb{E}_x\left[\left(\sum_{j=1}^{r^{l_n(x)}} I_j\, r^{-l_n(x)}\right)^2\right] \\
&= \sum_{m=0}^{l_n(x)-1} \mathbb{E}_x\left[\sum_{j=1}^{r^m} \sum_{k,l\in\chi_{m,j},k\neq l} \left(\sum_{m_k\in d_k} I_{m_k} r^{-l_n(x)}\right)\left(\sum_{m_l\in d_l} I_{m_l} r^{-l_n(x)}\right)\right] \\
&\quad + \mathbb{E}_x\left[\sum_{j=1}^{r^{l_n(x)}} I_j\, r^{-2l_n(x)}\right],
\end{aligned}$$

where we define I_{m_k} to be the indicator function of the event that particle m_k is in the set C_0^n. The second term above is essentially the diagonal terms of the second moment (43), and for the off-diagonal terms, for each generation, we categorize particles according to their common ancestors, a technique used by [8]. For the first term, we have

$$\begin{aligned}
&\sum_{m=0}^{l_n(x)-1} \mathbb{E}_x\left[\sum_{j=1}^{r^m} \sum_{k,l\in\chi_{m,j},k\neq l} \left(\sum_{m_k\in d_k} I_{m_k} r^{-l_n(x)}\right)\left(\sum_{m_l\in d_l} I_{m_l} r^{-l_n(x)}\right)\right] \\
&= \sum_{m=0}^{l_n(x)-1} \mathbb{E}_x\left[\sum_{j=1}^{r^m} I\left(V\left(Q_{m,j}\right)>0\right)\left(r^{-m}\right)^2\right. \\
&\left.\cdot \sum_{k,l\in\chi_{m,j},k\neq l} \left(\frac{1}{r}\sum_{m_k\in d_k} I_{m_k} r^{-(l_n(x)-m-1)}\right)\left(\frac{1}{r}\sum_{m_l\in d_l} I_{m_l} r^{-(l_n(x)-m-1)}\right)\right].
\end{aligned}$$

Conditioning on the whole genealogy up to step m, we obtain

$$\begin{aligned}
&\mathbb{E}_x\left[\sum_{j=1}^{r^m} I\left(V\left(Q_{m,j}\right)>0\right)\left(r^{-m}\right)^2\right. \\
&\left.\cdot \sum_{k,l\in\chi_{m,j},k\neq l} \left(\frac{1}{r}\sum_{m_k\in d_k} I_{m_k} r^{-(l_n(x)-m-1)}\right)\left(\frac{1}{r}\sum_{m_l\in d_l} I_{m_l} r^{-(l_n(x)-m-1)}\right)\right] \\
&= \mathbb{E}_x\left[\sum_{j=1}^{r^m} I\left(V\left(Q_{m,j}\right)>0\right)\left(r^{-m}\right)^2 \mathbb{E}_x\left(\sum_{k,l\in\chi_{m,j},k\neq l}\right.\right.
\end{aligned}$$

$$\left(\frac{1}{r}\sum_{m_k\in d_k} I_{m_k} r^{-(l_n(x)-m-1)}\right)\left(\frac{1}{r}\sum_{m_l\in d_l} I_{m_l} r^{-(l_n(x)-m-1)}\right)\Big|Q_{m,j}\Bigg)\Bigg]$$

$$= \mathbb{E}_x\left[\sum_{j=1}^{r^m} I\left(V\left(Q_{m,j}\right)>0\right) r^{-2m} \sum_{k,l\in\chi_{m,j},k\neq l}\right.$$

$$\left.\left(\frac{1}{r}\mathbb{E}_{Q_{m,j}}\left(\sum_{m_k\in d_k} I_{m_k} r^{-(l_n(x)-m-1)}\right)\frac{1}{r}\mathbb{E}_{Q_{m,j}}\left(\sum_{m_l\in d_l} I_{m_l} r^{-(l_n(x)-m-1)}\right)\right)\right].$$

Note that $\mathbb{E}_{Q_{m,j}}[\sum_{m_k\in d_k} I_{m_k} r^{-(l_n(x)-m-1)}] = p_n^V(Q_{m,j})$, and $\mathcal{W} = \sum_{k,l\in\chi_{m,j};k\neq l} r^{-2} = (r-1)/r$. Summing over m we obtain

$$\begin{aligned}
&\mathbb{E}_x\left[\left(\sum_{j=1}^{r^{l_n(x)}} I_j\, r^{-l_n(x)}\right)^2\right] - \mathbb{E}_x\left(\sum_{j=1}^{r^{l_n(x)}} I_j r^{-2l_n(x)}\right)\\
&= \mathcal{W}\sum_{m=0}^{l_n(x)-1}\mathbb{E}_x\left[\sum_{j=1}^{r^m} I\left(V\left(Q_{m,j}\right)>0\right) r^{-2m} p_n^V\left(Q_{m,j}\right)^2\right]\\
&= \mathcal{W}\sum_{m=0}^{l_n(x)-1} r^{-m}\mathbb{E}_x\left[I\left(V\left(Q_{m,1}\right)>0\right) p_n^V\left(Q_{m,1}\right)^2\right].
\end{aligned}$$

Combining this with the diagonal term in (43), which can be readily expressed as $r^{-l_n(x)}p_n^V(x)$, we arrive at the following expansion for the second moment of $R_n(x)$:

$$\begin{aligned}
\mathbb{E}_x\left[R_n(x)^2\right] &= \mathcal{W}\sum_{m=0}^{l_n(x)-1} r^{-m}\mathbb{E}_x\left[I\left(V\left(Q_{m,1}\right)>0\right) p_n^V\left(Q_{m,1}\right)^2\right]\\
&\quad + r^{-l_n(x)}p_n^V(x).
\end{aligned} \tag{44}$$

The next result takes advantage of expression (44) to obtain an upper bound for $\mathbb{E}_x[R_n(x)^2]$.

PROPOSITION 7. *The second moment of $R_n(x)$ satisfies*

$$\mathbb{E}\left[R_n(x)\right]^2 = p_n^V(x)^2\, O\left(n^{\beta_V}\right). \tag{45}$$

where β_V is the number of bottleneck stations in the subset corresponding to V.

In order to prove the previous result, we will show that the second moment of $R_n(x)$ is dominated by the first item on the right hand side of the equality in (44). In turn, the asymptotic behavior of such term hinges on the conditional distribution of the exact position of the particle in generation m, $Q_{m,1}$ in $C^n_{l_n(x)-m}$.

PROOF. Using the equivalence observed in (39), the expectation term in the sum of (44) can be expressed as

$$\begin{aligned}
&\mathbb{E}_x\left[I\left(V\left(Q_{m,1}\right)>0\right)p_n^V\left(Q_{m,1}\right)^2\right]\\
&=\mathbb{E}_x\left[I\left(\varrho^T Q_{m,1}\leq \varrho^T x-(m-1)\log r\right)p_n^V\left(Q_{m,1}\right)^2\right]\\
&=\mathbb{E}_x\left[p_n^V\left(Q_{m,1}\right)^2|\varrho^T Q_{m,1}\leq \varrho^T x-(m-1)\log r\right]\mathbb{P}_x\left(\hat{T}_{\frac{m}{C}}<T_0\right)
\end{aligned}\tag{46}$$

where we used the property derived in (39). Before we proceed, let us define the inverse mapping $V^{-1}:\mathcal{Z}_+\to\mathcal{Z}_+^d$ by

$$V^{-1}(n)=\{x\in\mathcal{Z}_+^d : V(x)=n\},$$

i.e., the configuration of the network such that the total population in stations encoded by v is n. For the first item in (46), we have

$$\begin{aligned}
&\mathbb{E}_x\left[p_n^V\left(Q_{m,1}\right)^2|\varrho^T Q_{m,1}\leq \varrho^T x-(m-1)\log r\right]\\
&\leq K\mathbb{E}\left[\frac{\pi^2\left(V^{-1}(n)\right)}{\pi^2\left(\{Q_{m,1}\}\right)}|\varrho^T Q_{m,1}\leq \varrho^T x-(m-1)\log r\right]\\
&=K\pi^2\left(V^{-1}(n)\right)c_1\mathbb{E}_\pi\left[e^{-2\varrho^T Q_{m,1}}|\varrho^T Q_{m,1}\leq \varrho^T x-(m-1)\log r\right]
\end{aligned}\tag{47}$$

where c_1, K are some constants independent of n. Here for the inequality we used Proposition 1. To reach the equality we used the fact that $\pi^{-1}\left(\{Q_{m,1}\}\right)=c_1e^{-\varrho^T Q_{m.1}}$ for some positive constant c_1. As for the expectation term in (47), since the process $Q(\cdot)$ has for each dimension an increment at most of unit size, we can write

(48)

$$\begin{aligned}
&\mathbb{E}_\pi\left[e^{-2\varrho^T Q_{m,1}}|\varrho^T Q_{m,1}\leq \varrho^T x-(m-1)\log r\right]\\
&=\mathbb{E}_\pi\left[e^{-2\varrho^T Q_{m,1}}|\varrho^T x-(m-1)\log r-\delta\leq \varrho^T Q_{m,1}\leq \varrho^T x-(m-1)\log r\right]\\
&\leq c_2\exp\left(-2\varrho^T x+2(m-1)\log r\right)\\
&=c_3\exp\left(-2\frac{m-1}{C}\log\rho_*^V\right)=c_3\left(\rho_*^V\right)^{-2\frac{m-1}{C}},
\end{aligned}$$

where c_2, c_3 and δ are some positive constants. Combining this with

$$\mathbb{P}_x\left(\hat{T}_{\frac{m}{C}} < T_0\right) = O\left(\left(\frac{m-1}{C}\right)^{\beta_V - 1}\left(\rho_*^V\right)^{\frac{m-1}{C}}\right)$$

according to Proposition 6, we obtain the following upper bound for the expectation term in the sum of expression (44):

$$\begin{aligned}
&\mathbb{E}_x\left[I\left(V\left(Q_{m,1}\right) > 0\right) p_n^V\left(Q_{m,1}\right)^2\right] \\
&= K\pi^2\left(V^{-1}(n)\right)\pi^{-2}(x)\left(\rho_*^V\right)^{-2\frac{m-1}{C}} O\left(\left(\frac{m-1}{C}\right)^{\beta_V - 1}\left(\rho_*^V\right)^{\frac{m-1}{C}}\right) \qquad (49)\\
&= O\left(p_n^V(x)^2 r^{m-1}\left(\frac{m-1}{C}\right)^{\beta_V - 1}\right)
\end{aligned}$$

where for the second equality we used again Proposition 1 and the fact that $\rho_*^V = r^{-C}$. Putting the bound in (49) back to the sum in the first item of (44), we have

$$\begin{aligned}
&\sum_{m=0}^{l_n(x)-1} r^{-m}\mathbb{E}_x\left[I\left(V\left(Q_{m,1}\right) > 0\right) p_n^V\left(Q_{m,1}\right)^2\right] \\
&= r^{-1}\sum_{m=0}^{l_n(x)-1} O\left(p_n^V(x)^2\left(\frac{m-1}{C}\right)^{\beta_V - 1}\right) \qquad (50)\\
&= p_n^V(x)^2\, O\left(n^{\beta_V}\right).
\end{aligned}$$

Finally, note that the second item of (44) is dominated by (50), and it follows immediately that

$$\mathbb{E}\left[R_n(x)\right]^2 = p_n^V(x)^2\, O\left(n^{\beta_V}\right).$$

□

Equipped with these results, we are ready to summarize our discussions in the statement of the following Theorem, which is the main result of this paper.

THEOREM 1. *To estimate the overflow probability $p_n^V(x)$ using $R_n(x)$, the number of function evaluations needed for a given level of relative error is $O(n^{2\beta_V+1})$.*

PROOF. Recall from Section 2 that the number of function evaluations sufficient to achieve a pre-determined level of relative accuracy for the splitting estimator is proportional to the work-normalized squared coefficient of variation. This is therefore immediate by combining the upper bound analysis of the computational effort per run in Proposition 5 along with the upper bound of the second moment of $R_n(x)$ available in Proposition 7. □

A direct comparison to the $O(n^{3d-2})$ complexity of solving a system of linear equations (see Section 2) yields the immediate conclusion that the splitting algorithm is "efficient" in the sense that it is an improvement over the "benchmark" polynomial algorithm. Even in the worst case scenario, when we look at the total population of the network and the network is totally symmetric, i.e., all stations are bottlenecks ($\beta_V = d > 3$), the number of function evaluations needed is a substantial reduction of n^{d-3}. In the case where $\beta_V = 1$, the algorithm only requires a number of function evaluations that at most grows cubically in the level of overflow n. Furthermore, if the number of bottlenecks is less than half of the total number of stations, i.e. $\beta_V < d/2$, the splitting algorithm enjoys a running time of order smaller than $O(n^d)$, which is not worse than storing the vector that encodes the solution to the associated linear system. If, on the other hand, more than half of the stations are bottlenecks, faster importance sampling based algorithms do exist at least for the case of tandem networks; see the analysis in [6], which implies that $O(n^{2(d-\beta)+1})$ function evaluations suffice to obtain an estimator with a given relative precision. Overall, the analysis thus provides some sort of guidance on the choice of simulation algorithms. It is meaningful to point out that the previous comparison is not based on the sharpest analysis. In fact we only resort to a rather crude upper bound in the analysis of the second moment of $R_n(x)$ in (47). A sharper result is possible by bounding the expectation term in (46) with more care. But as pointed out in the Introduction, even though there is still room for a more refined analysis, we believe our work provides substantial insights leading to a better understanding of the relations between these two classes of algorithms.

REMARK 1. Numerical experiments have been performed for this class of algorithms in [8]. We replicated some of their experiments and from the numerical evidence we could see that there is still room for a sharper bound. In particular, when studying overflow for the total population of the network, our experiments suggest a computational cost roughly similar to $O(n^{\beta_V})$ (as opposed to $O(n^{2\beta_V+1})$) for a fixed level of relative error. We have chosen not to present the numerical details in this paper since we think a sharper analysis is needed for a better interpretation of the results. The rough $O(n^{\beta_V+1})$

additional effort in our estimate, we believe, comes from the application of (36) in the proofs of both Proposition 5 and Proposition 7. Note that the bound becomes too loose when the position of the survival particle at level m satisfying $V(Q_{m,1}) > 0$ is no longer $O(1)$. Instead, conditional on a particle surviving at level $m = \Theta(n)$, the particle is with high probability in the most likely fluid trajectory to overflow. However, to account for its exact position, we would need a conditional local central limit theorem correction. This accounts for a factor of $n^{\beta_V/2}$ in both 1) expected computational effort per run for a single replication of the estimator and 2) the second moment of the estimator. Combining these two terms seems to explain most of the gap between our bound and what appears to be the actual empirical performance.

Acknowledgements. The authors are grateful to the referees for the careful review of the manuscript and the useful comments that greatly improved the exposition of the paper. Support from the NSF foundation through the grants DMS-0806145, DMS-0846816 and DMS-1069064 is gratefully acknowledged.

REFERENCES

[1] V. Anantharam, P. Heidelberger, and P. Tsoucas. Analysis of rare events in continuous time marked chains via time reversal and fluid approximation. *IBM Research Report, REC 16280*, 1990.

[2] P. Arbenz and W. Gander. A survey of direct parallel algorithms for banded linear systems. Technical Report 221, Department Informatik,ETH Zurich, 1994.

[3] S. Asmussen and P. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer-Verlag, New York, NY, USA, 2008. MR2331321

[4] J. Blanchet. Optimal sampling of overflow paths in Jackson networks. *To Appear in Math. of O.R.*, 2011.

[5] J. Blanchet and P. Glynn. Efficient rare-event simulation for the maximum of a heavy-tailed random walk. *Ann. of Appl. Probab.*, 18:1351–1378, 2008. MR2434174

[6] J. Blanchet, K. Leder, and P. Glynn. Lyapunov functions and subsolutions for rare event simulation. *Submitted*, 2011.

[7] J. Blanchet and M. Mandjes. Rare event simulation for queues. In G. Rubino and B. Tuffin, editors, *Rare Event Simulation Using Monte Carlo Methods*, pages 87–124. Wiley, West Sussex, United Kingdom, 2009. Chapter 5. MR2730763

[8] T. Dean and P. Dupuis. Splitting for rare event simulation: A large deviation approach to design and analysis. *Stochastic Processes and Its Applications*, (119):562–587. MR2494004

[9] P. Dupuis and R. S. Ellis. The large deviation principle for a general class of queueing systems I. *Trans. of the American Mathematical Society*, 347:2689–2751, 1995. MR1290716

[10] P. Dupuis, A. Sezer, and H. Wang. Dynamic importance sampling for queueing networks. *Ann. Appl. Probab.*, 17:1306–1346, 2007. MR2344308

[11] P. Dupuis and H. Wang. Importance sampling, large deviations, and differential games. *Stoch. and Stoch. Reports*, 76:481–508, 2004. MR2100018

[12] P. Dupuis and H. Wang. Importance sampling for Jackson networks. *Preprint*, 2008.

[13] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. Multilevel splitting for estimating rare event probabilities, 1999. MR1710951

[14] P. Glasserman and S. Kou. Analysis of an importance sampling estimator for tandem queues. *ACM TOMACS*, 5:22–42, 1995.

[15] I. Ignatiouk-Robert. Large deviations of Jackson networks. *Annals of Applied Probability*, 10:962–1001, 2000. MR1789985

[16] S. Juneja and V. Nicola. Efficient simulation of buffer overflow probabilities in Jackson networks with feedback. *ACM Trans. Model. Comput. Simul.*, 15(4):281–315, 2005.

[17] S. Juneja and P. Shahabuddin. Rare event simulation techniques: An introduction and recent advances. In S. G. Henderson and B. L. Nelson, editors, *Simulation*, Handbooks in Operations Research and Management Science, pages 291–350. Elsevier, Amsterdam, The Netherlands, 2006.

[18] D. Kroese and V. Nicola. Efficient simulation of a tandem Jackson network. *ACM Trans. Model. Comput. Simul.*, 12:119–141, 2002.

[19] K. Majewski and K. Ramanan. How large queues build up in a Jackson network. *To Appear in Math. of O.R.*, 2008.

[20] M.Villen-Altamirano and J. Villen-Altamirano. Restart: A method for accelerating rare even simulations. In J.W. Colhen and C.D. Pack, editors, *Proceedings of the 13th International Teletraffic Congress. In Queueing, performance and control in ATM*, pages 71–76. Elsevier Science Publishers, 1993.

[21] V. Nicola and T. Zaburnenko. Efficient importance sampling heuristics for the simulation of population overflow in Jackson networks. *ACM Trans. Model. Comput. Simul.*, 17(2), 2007.

[22] S. Parekh and J. Walrand. Quick simulation of rare events in networks. *IEEE Trans. Automat. Contr.*, 34:54–66, 1989. MR0970932

[23] P. Robert. *Stochastic Networks and Queues.* Springer-Verlag, Berlin, 2003. MR1996883

[24] M. Villén-Altamirano and J. Villén-Altamirano. Restart: a straightforward method for fast simulation of rare events. In *Winter Simulation Conference*, pages 282–289, 1994.

Department of Industrial Engineering
and Operations Research
Columbia University
in the City of New York
E-mail: jose.blanchet@columbia.edu
ys2347@columbia.edu

Department of Industrial
and System Engineering
University of Minnesota
Minneapolis, MN
E-mail: kevin.leder@me.umn.edu