Revisiting Stochastic Loss Networks: Structures and Algorithms

Kyomin Jung*
Department of Mathematics
MIT
Cambridge, MA 02139, USA
kmjung@mit.edu

Yingdong Lu
Mathematical Sciences Dept.
IBM Watson Research Center
Yorktown, NY 10598, USA
yingdong@us.ibm.com

Devavrat Shah†
Department of EECS
MIT
Cambridge, MA 02139, USA
devavrat@mit.edu

Mayank Sharma
Mathematical Sciences Dept.
IBM Watson Research Center
Yorktown, NY 10598, USA
mxsharma@us.ibm.com

Mark S. Squillante
Mathematical Sciences Dept.
IBM Watson Research Center
Yorktown, NY 10598, USA
mss@watson.ibm.com

ABSTRACT

This paper considers structural and algorithmic problems in stochastic loss networks. The very popular Erlang approximation can be shown to provide relatively poor performance estimates, especially for loss networks in the critically loaded regime. This paper proposes a novel algorithm for estimating the stationary loss probabilities in stochastic loss networks based on structural properties of the exact stationary distribution, which is shown to always converge, exponentially fast, to the asymptotically exact results. Using a variational characterization of the stationary distribution, an alternative proof is provided for an important result due to Kelly, which is simpler and may be of interest in its own right. This paper also determines structural properties of the inverse Erlang function characterizing the region of capacities that ensures offered traffic is served within a set of loss probabilities. Numerical experiments investigate various issues of both theoretical and practical interest.

Categories and Subject Descriptors

G.3 [**Probability & Statistics**]: Stochastic processes, Markov processes, Queueing theory; F.2.2 [**Nonnumerical Algorithms & Problems**]: Computations on discrete structures, Geometrical problems and computations; G.1.6 [**Optimization**]: Nonlinear programming

General Terms

Theory, Algorithms, Performance

*This work is partially carried out while the author was visiting the IBM T.J. Watson Research Center.

†Work of Shah was supported in parts by NSF CCF 0728554 and NSF CNS 0546590.

SIGMETRICS'08, June 2–6, 2008, Annapolis, Maryland, USA.

Keywords

Loss networks, multidimensional stochastic processes, stochastic approximations, Erlang loss formula and fixed-point approximation

1. INTRODUCTION

As the complexities of computer and communication systems continue to grow at a rapid pace, performance modeling, analysis and optimization are playing an increasingly important role in the design and implementation of such complex systems. Central to these research studies are the models of the various applications of interest. For almost a century, starting with the seminal work of Erlang [5], stochastic loss networks have been widely studied as models of many diverse computer and communication systems in which different types of resources are used to serve various classes of customers involving simultaneous resource possession and non-backlogging workloads. Examples include telephone networks, mobile cellular systems, ATM networks, broadband telecommunication networks, optical wavelength-division multiplexing networks, wireless networks, distributed computing, database systems, data centers, and multi-item inventory systems; see, e.g., [7, 10–13, 19–22, 25, 26]. Loss networks have even been used recently for resource planning within the context of workforce management in the information technology (IT) services industry, where a collection of IT service products are offered each requiring a set of resources with certain capabilities [2, 17]. In each case, the stochastic loss network is used to capture the dynamics and uncertainty of the computer/communication application being modeled.

One of the most important objectives in analyzing such loss networks is to determine performance measures of interest, most notably the stationary loss probability for each customer class. The classical Erlang formula, which has been thoroughly studied and widely applied in many fields of research, provides the probabilistic characterization of these loss probabilities. More specifically, given a stochastic network and a multiclass customer workload, the formula renders the stationary probability that a customer will be lost due to insufficient capacity for at least one resource type. While the initial results of Erlang [5] were for the particular case of Poisson arrivals and exponential service times, Sevastyanov [23] demonstrates that the Erlang formula holds under general finite-mean distributions for the customer service times. The results are also known to hold in the presence of dependencies among service times for a specific class [4]. Recent results [3] suggest that relaxations can be made to the customer arrival process, merely requiring that customers generate sessions according to a Poisson process and, within each session, blocked customers may retry with a fixed probability after an idle period of random length. A multi-period version of the Erlang loss model has also been recently studied [2].

Unfortunately, the computational complexity of the exact Erlang formula and related measures is known to be $\sharp P$-complete in the size of the network [16], thus rendering the exact formula of limited use for many networks in practice. (Refer to [24] for details on $\sharp P$-complete complexity.) The well-known Erlang fixed-point approximation has been developed to address this problem of complexity through a set of product-form expressions for the blocking probabilities of the individual resources that map the blocking probability of each resource to the blocking probabilities of other resources. In other words, it is as if customer losses are caused by independent blocking events on each of the resources used by the customer class based on the one-dimensional Erlang function. The Erlang fixed-point approximation has been frequently used and extensively studied as a tractable approach for calculating performance measures associated with the stochastic loss network, including estimates for stationary loss probabilities. Moreover, the Erlang fixed-point approximation has been shown to be asymptotically exact in two limiting regimes, one based on increasing the traffic intensities and resource capacities in a proportional manner [11, 12], and the other based on increasing the number of resource types and number of customer classes [25, 27].

Despite being asymptotically exact in certain limiting regimes, it is equally well known that the Erlang fixed-point approximation can provide relatively poor performance estimates in various cases. The stochastic loss networks that model many computer and communication systems often operate naturally in a so-called "critically loaded" regime [6]. Somewhat surprisingly, we find that, even though the Erlang fixed-point approximation can perform quite well in underloaded and overloaded conditions, the fixed-point approximation can provide relatively poor loss probability estimates when the network is critically loaded. We establish such qualitative results by means of estimating the convergence rate of the Erlang fixed-point approximation toward the exact solution under large network scalings. This motivates the need to design better algorithms for estimating loss probabilities.

In this paper we propose a novel algorithm for computing the stationary loss probabilities in stochastic loss networks, which we call the "slice method" because the algorithm exploits structural properties of the exact stationary distribution along "slices" of the polytope over which it is defined. Our algorithm is shown to always converge and to do so exponentially fast. Through a variational characterization of the stationary distribution, we establish that the results from our algorithm are asymptotically exact. We further estimate the convergence rate of our algorithm, where comparisons between the convergence rates of the Erlang fixed-point approximation and the slice method favors our approach. Using this variational characterization, we also provide an alternative proof of the main theorem in [11], which is much simpler and may be of interest in its own right. A collection of numerical experiments further investigates the effectiveness of our algorithm where it convincingly outperforms the Erlang fixed-point approximation for loss networks in the critically loaded regime.

Another important objective in analyzing stochastic loss networks concerns characterizing the fundamental relationships among the capacities for every resource type and the loss probabilities for every customer class. In particular, the exact Erlang formula provides the loss probabilities for all customer classes given the capacity of each resource and the workload of each class. Berezner et al. [1] consider the inverse of this function in the one-dimensional, single-resource case and provide bounds for the capacity required to satisfy the given workload and loss probability constraint. The corresponding bounds for the multidimensional version of this inverse function is a much more difficult problem. The main reason is that, depending upon the structure of the problem and in particular the resource requirements of each customer class, there can be infinitely many possible capacities that satisfy the given vector of loss probabilities for the customer classes. Furthermore, the set that contains all of these possible capacities can be unbounded.

In this paper, by exploiting large network scalings, our results for various approximation algorithms, and previous results for the one-dimensional problem, we establish structural properties for the multidimensional region of capacities that ensures offered traffic will be served within a given set of loss probabilities. This region of capacities is defined in terms of a system of polynomial equations and inequalities, such that the capacities which correspond to the given loss probability vector lie with this region. These results provide a probabilistic characterization of the theoretical relationships between the link capacity and loss probability vectors. Our results also can be exploited to efficiently search the feasible region of various optimization problems involving loss networks, including many resource allocation and capacity planning applications.

We make several important contributions in this paper. A new slice method for estimating the stationary loss probabilities in stochastic loss networks is proposed and shown to provide asymptotically exact results. The convergence rates of different approximation algorithms are obtained under large network scalings. A simpler proof is provided for a classical result of Kelly which should be of independent interest. The structural properties of the capacity vector region that achieves a given loss probability vector are obtained under large network scalings. While the problems we consider are of fundamental importance from the theoretical perspectives of stochastic loss networks in general and Erlang loss model approximations in particular, our analysis and results can support a wide range of practical applications involving loss networks.

This paper is organized as follows. The next section contains some technical preliminaries. Section 3 describes three approximation algorithms for computing stationary loss probabilities. Our main results are presented in Section 4, with most of their proofs considered in Section 5. Some numerical experiments are provided in Section 6, and concluding remarks can be found in Section 7. Additional technical details can be found in [8].

2. PRELIMINARIES

2.1 Model

We investigate general stochastic loss networks with fixed routing, using the standard terminology in the literature based on routes (customer classes) and links (resource types); see, e.g., [13]. Consider a network with J links, labeled $1, 2, \ldots, J$. Each link j has C_j units of capacity. There is a set of K distinct (pre-determined) routes, denoted by $\mathcal{R} = \{1, \ldots, K\}$. A call on route r requires A_{jr} units of capacity on link j, $A_{jr} \geq 0$. Calls on route r arrive according to an independent Poisson process of rate ν_r, with $\underline{\nu} = (\nu_1, \ldots, \nu_K)$ denoting the vector of these rates. The dynamics of the network are such that an arriving call on route r is admitted to the network if sufficient capacity is available on all links used by route r; else, the call is dropped. To simplify the exposition, we will assume that the call service times are i.i.d. exponential random variables with unit mean. It is important to note, however, that our results are not limited to these service time assumptions since the quantities of interest remain unchanged in the stationary regime under general service time distributions due to the well-known *insensitivity* property of this class of stationary loss networks.

Let $\underline{n}(t) = (n_1(t), \dots, n_K(t)) \in \mathbb{N}^K$ be the vector of the number of active calls in the network at time t. By definition, we have that $\underline{n}(t) \in \mathcal{S}(C)$ where

$$\mathcal{S}(C) = \left\{ \underline{n} \in \mathbb{Z}^K : \underline{n} \geq 0,\ A\underline{n} \leq \underline{C} \right\},$$

and $\underline{C} = (C_1, \dots, C_J)$ denotes the vector of link capacities. Within this framework, the network is Markov with respect to state $\underline{n}(t)$. It has been well established that the network is a *reversible* multidimensional Markov process with a product-form stationary distribution [9]. Namely, there is a unique stationary distribution π on the state space $\mathcal{S}(C)$ such that for $\underline{n} \in \mathcal{S}(C)$

$$\pi(\underline{n}) = G(C)^{-1} \prod_{r \in \mathcal{R}} \frac{\nu_r^{n_r}}{n_r!},$$

where $G(C)$ is the normalizing constant (or partition function)

$$G(C) = \sum_{\underline{n} \in \mathcal{S}(C)} \prod_{r \in \mathcal{R}} \frac{\nu_r^{\underline{n}_r}}{\underline{n}_r!}.$$

2.2 Problems

A primary performance measure in loss networks is the per-route stationary loss probability, the fraction of calls on route r in equilibrium that are dropped or lost, denoted by L_r. It can be easily verified that L_r is well-defined in the above model. This model can be thought of as a stable system where admitted calls experience an average delay of 1 (their service requirement) and lost calls experience a delay of 0 (their immediate departure). Therefore, the average delay experienced by calls on route r is given by

$$D_r = (1 - L_r) \times 1 + L_r \times 0 \ = \ (1 - L_r).$$

Upon applying Little's law [15] to this stable system (with respect to route r), we obtain

$$\nu_r D_r = \mathbb{E}[n_r]$$

which yields

$$1 - L_r = \frac{\mathbb{E}[n_r]}{\nu_r}. \tag{1}$$

Thus, computing L_r is equivalent to computing the expected value of the number of active calls on route r with respect to the stationary distribution of the network. Even though we have an explicit formula, the computational complexity of the exact stationary distribution, known to be $\sharp P$-complete in general [16], renders its direct use of limited value in practice. We therefore need simple, efficient and (possibly) approximate algorithms for computing the stationary loss probabilities. One of our goals in this paper is to design a family of such iterative algorithms that also have provably good accuracy properties.

In the one-dimensional version of the foregoing stochastic loss network, it is known that the capacity C for the single resource must satisfy the inequalities

$$\nu(1 - L) \ < \ C \ < \ \nu(1 - L) + 1/L \tag{2}$$

in order to ensure that the arrivals at rate ν are served with a loss probability of at most L [1]. The corresponding problem in the multidimensional stochastic loss network of interest is much more difficult, however, and very limited results are known. On the other hand, understanding the fundamental relationships among the link capacities and the loss probabilities is critical to solving resource allocation problems in stochastic loss networks. We therefore need an effective characterization of these relationships, which can be exploited to improve the efficiency and quality of solutions to a wide variety of optimization problems [2,12,17,25]. Another of our goals in this paper is to determine the capacity region that ensures a given vector of loss probabilities will be satisfied.

2.3 Scaling

We consider a scaling of the stochastic loss network to model the type of large networks that arise in various applications. Although it has been well studied (see, e.g., [11]), we will use this scaling both to evaluate analytically the performance of different approximation algorithms for computing loss probabilities and to obtain the capacity region for satisfying a set of loss probabilities.

Given a stochastic loss network with parameters $\underline{C}$, A and $\underline{\nu}$, a scaled version of the system is defined by the scaled capacities

$$\underline{C}_N = N\underline{C} = (NC_1, \dots, NC_K)$$

and the scaled arrival rates

$$\underline{\nu}_N = N\underline{\nu} = (N\nu_1, \dots, N\nu_K),$$

where $N \in \mathbb{N}$ is the system scaling parameter. The corresponding feasible region of calls is given by $\mathcal{S}(N\underline{C})$. Now consider a normalized version of this region defined as

$$\mathcal{S}_N(C) = \left\{ \frac{1}{N}\underline{n} : \underline{n} \in \mathcal{S}(N\underline{C}) \right\}.$$

Then the following *continuous approximation* of $\mathcal{S}_N(C)$ emerges in the large N limit:

$$\bar{\mathcal{S}}(C) = \{\underline{x} : A\underline{x} \leq \underline{C},\ \underline{x} \in \mathbb{R}_+^K\}.$$

3. ALGORITHMS

We now describe three algorithms for computing the stationary loss probabilities $\underline{L} = (L_r) \in [0,1]^K$. The well-known Erlang fixed-point approximation is presented first, followed by a "1-point approximation" based on the concentration of the stationary distribution around its mode in large networks. The third algorithm is our new family of "slice methods" that attempts to compute the average number of active calls on different routes via an efficient exploration through "slices" of the admissible polytope $\mathcal{S}(C)$.

3.1 Erlang fixed-point approximation

The well-known Erlang formula [5] for a single-link, single-route network with capacity C and arrival rate ν states that the loss probability, denoted by $E(\nu, C)$, is given by

$$E(\nu, C) = \frac{\nu^C}{C!} \left[\sum_{i=0}^{C} \frac{\nu^i}{i!} \right]^{-1}.$$

Based on this simple formula, the Erlang fixed-point approximation for multi-link, multi-route networks arose from the hypothesis that calls are lost due to *independent* blocking events on each link in the route. More formally, this hypothesis implies that the loss probabilities of routes $\underline{L} = (L_1, \dots, L_K)$ and blocking probabilities of links $\underline{E} = (E_1, \dots, E_J)$ satisfy the set of *fixed-point* equations

$$\begin{aligned} E_j &= E(\rho_j, C_j), \\ \rho_j &= \frac{1}{1 - E_j} \left[\sum_r \nu_r A_{jr} \prod_i (1 - E_i)^{A_{ir}} \right], \\ 1 - L_r &= \prod_j (1 - E_j)^{A_{jr}}, \end{aligned} \tag{3}$$

for $j = 1, \ldots, J$ and $r \in \mathcal{R}$.

A natural iterative algorithm that attempts to obtain a solution to the above fixed-point equations is as follows:

ERLANG FIXED-POINT APPROXIMATION.

---

1. Denote by t the iteration of the algorithm, with $t = 0$ initially. Start with $E_j^{(0)} = 0.5$ for all $1 \le j \le J$.

2. In iteration $t+1$, update $E_j^{(t+1)}$ according to

$$E_j^{(t+1)} = E(\rho_j^{(t)}, C_j),$$

where

$$\rho_j^{(t)} = (1 - E_j^{(t)})^{-1} \sum_{r:j\in r} \nu_r A_{jr} \prod_{i:i\in r} (1 - E_i^{(t)})^{A_{ir}}.$$

3. Upon convergence per appropriate stopping conditions, denote the resulting values by $E_j^{\mathcal{E}}$ for $1 \le j \le J$. Compute the loss probabilities from the Erlang fixed-point approximation, $L_r^{\mathcal{E}}$, $r \in \mathcal{R}$, as

$$1 - L_r^{\mathcal{E}} = \prod_j (1 - E_j^{\mathcal{E}})^{A_{jr}}.$$

3.2 1-point approximation

Kelly [11] established the asymptotic exactness of the Erlang fixed-point approximation in a large network limiting regime by showing that the stationary distribution concentrates around its mode $\underline{n}^*$ given by

$$\underline{n}^* \in \arg \max_{\underline{n}\in\mathcal{S}(C)} \pi(\underline{n}).$$

Such concentration suggests the following approach which is the premise of the *1-point approximation*: Compute the mode $\underline{n}^* = (n_r^*)$ of the distribution and use n_r^* as a surrogate for $\mathbb{E}[n_r]$ in the computation of L_r via equation (1). Before presenting our specific iterative algorithm, we consider some related optimization problems upon which it is based.

The definition of the stationary distribution $\pi(\cdot)$ suggests that the mode $\underline{n}^*$ corresponds to a solution of the optimization problem

$$\begin{array}{ll} \text{maximize} & \sum_r n_r \log \nu_r - \log n_r! \\ \text{over} & \underline{n} \in \mathcal{S}(C). \end{array}$$

By Stirling's approximation, $\log n_r! = n_r \log n_r - n_r + O(\log n_r)$. Using this and ignoring the $O(\log n_r)$ term, the above optimization problem reduces to

$$\begin{array}{ll} \text{maximize} & \sum_r n_r \log \nu_r + n_r - n_r \log n_r \\ \text{over} & \underline{n} \in \mathcal{S}(C). \end{array}$$

A natural continuous relaxation of $\underline{n} \in \mathcal{S}(C)$ is

$$\bar{\mathcal{S}}(C) = \left\{ \underline{x} \in \mathbb{R}_+^K : \ A\underline{x} \le \underline{C} \right\},$$

which yields the following primal problem (P):

$$\begin{array}{ll} \text{maximize} & \sum_r x_r \log \nu_r + x_r - x_r \log x_r \\ \text{over} & \underline{x} \in \bar{\mathcal{S}}(C). \end{array}$$

The above relaxation becomes a good approximation of the original problem when all components of $\underline{C}$ are large. In order to design a simple iterative algorithm, we consider the Lagrangian dual (D) to the primal problem P where standard calculations yield

$$\begin{array}{ll} \text{minimize} & \sum_r \nu_r \exp\left[-\sum_j y_j A_{jr} \right] + \sum_j y_j C_j \\ \text{over} & \underline{y} \ge 0. \end{array}$$

Define the dual cost function $g(\underline{y})$ as

$$g(\underline{y}) = \sum_r \nu_r \exp\left[-\sum_j y_j A_{jr} \right] + \sum_j y_j C_j.$$

By Slater's condition, the strong duality holds and hence the optimal cost of P and D are the same. Standard Karush-Kuhn-Tucker conditions imply the following: Letting $(\underline{x}^*, \underline{y}^*)$ be a pair of optimal solutions to P and D, then

(a) For each link j,

$$\frac{\partial g(\underline{y}^*)}{\partial y_j} = 0 \ \text{ or } \ y_j^* = 0 \ \& \ \frac{\partial g(\underline{y}^*)}{\partial y_j} \le 0.$$

Equivalently,

$$\sum_r A_{jr} \nu_r \exp\left[-\sum_j y_j^* A_{jr} \right] = C_j \ \& \ y_j^* > 0,$$

$$\text{or,} \ \sum_r A_{jr} \nu_r \exp\left[-\sum_j y_j^* A_{jr} \right] \le C_j \ \& \ y_j^* = 0.$$

(b) For each route r,

$$x_r^* = \nu_r \exp\left[-\sum_j y_j^* A_{jr} \right].$$

The above conditions suggest the following approach: Obtain a dual optimal solution, say $\underline{y}^*$, use it to obtain $\underline{x}^*$, and then compute the loss probability as $1 - L_r = x_r^*/\nu_r$. Next, we describe an iterative, coordinate descent algorithm for obtaining $\underline{y}^*$. In what follows, we will use the transformation $z_j = \exp(-y_j)$ given its similarity with the Erlang fixed-point approximation. Note that z_j is 1 minus the blocking probability for link j, E_j.

1-POINT APPROXIMATION.

---

1. Denote by t the iteration of the algorithm, with $t = 0$ initially. Start with $z_j^{(0)} = 0.5$ for all $1 \le j \le J$.

2. In iteration $t+1$, determine $\underline{z}^{(t+1)}$ as follows:

   (a) Choose coordinates from $1, \ldots, J$ in a round-robin manner.

   (b) Update $z_j^{(t+1)}$ by solving the equation

   $$g_j^{(t)}(x) = \min\left\{ C_j, g_j^{(t)}(1) \right\},$$

   where $g_j^{(t)}(x) = \sum_r A_{jr} \nu_r \prod_i z_i^{A_{ir}}$ with

   $$z_i = \begin{cases} z_i^{(t+1)} & \text{for } i < j, \\ x & \text{for } i = j, \\ z_i^{(t)} & \text{for } i > j. \end{cases}$$

Thus, $g_j^{(t)}(x)$ is the evaluation of part of the function $g(\cdot)$ corresponding to the j^{th} coordinate with values of components $< j$ being from iteration $t+1$, values of components $> j$ from iteration t, and component j being the variable itself.

3. Upon convergence per appropriate stopping conditions, denote the resulting values by z_j^* for $1 \le j \le J$. Compute the loss probabilities from the 1-point approximation, L_r^*, $r \in \mathcal{R}$, as

$$1 - L_r^* = \prod_j (z_j^*)^{A_{jr}}.$$

3.3 Slice method

The Erlang fixed-point approximation and the 1-point approximation essentially attempt to use the mode of the stationary distribution as a surrogate for the mean, which works quite well when the distribution is concentrated (near its mode). While this concentration holds for asymptotically large networks, it otherwise can be an important source of error and therefore we seek to obtain a new family of methods that provide better approximations.

The main premise of our slice methods follows from the fact that computing the loss probability L_r is equivalent to computing the expected number of calls $\mathbb{E}[n_r]$ via equation (1). By definition,

$$\mathbb{E}[n_r] = \sum_{k=0}^{\infty} k \mathbb{P}[n_r = k]$$

and thus $\mathbb{E}[n_r]$ can be obtained through approximations of $\mathbb{P}[n_r = k]$ rather than by the mode value n_r^*. Note that $\mathbb{P}[n_r = k]$ corresponds to the probability mass along the "slice" of the polytope defined by $n_r = k$. An exact solution for $\mathbb{E}[n_r]$ can be obtained with our slice method by using the exact values of $\mathbb{P}[n_r = k]$, but obtaining the probability mass along a "slice" can be as computationally hard as the original problem. Hence, our family of slice methods is based on approximations for $\mathbb{P}[n_r = k]$. To do so, we will exploit similar insights from previous approaches: Most of the mass along each slice is concentrated around the mode of the distribution restricted to the slice. This approximation is better than the "1-point approximation" since it uses the "1-point approximation" many times (once for each slice) in order to obtain a more accurate solution. Next, we formally describe the algorithm, where the cost function of the primal problem P is denoted by

$$q(\underline{x}) = \sum_r x_r \log \nu_r + x_r - x_r \log x_r.$$

SLICE METHOD.

Compute L_r for route $r \in \mathcal{R}$ as follows:

1. For each value of $k \in \{n_r : \underline{n} \in \mathcal{S}(C)\}$, use the "1-point approximation" to compute $\underline{x}^*(k, r)$ as the solution of the optimization problem

$$\text{maximize } \ q(\underline{x}) \ \text{ over } \ \underline{x} \in \bar{\mathcal{S}}(C) \ \& \ x_r = k.$$

2. Estimate $\mathbb{E}[n_r]$ as

$$\mathbb{E}[n_r] = \frac{\sum_k k \exp(q(\underline{x}^*(k,r)))}{\sum_k \exp(q(\underline{x}^*(k,r)))}.$$

3. Generate $L_r = 1 - \frac{\mathbb{E}[n_r]}{\nu_r}$.

3.4 3-point slice method

In the general slice method, for each route r, we apply the 1-point approximation to each slice defined by $n_r = k$, $k \in \{n_r : \underline{n} \in \mathcal{S}(C)\}$. In the scaled system, this requires $O(N)$ applications of the "1-point approximation" for each route. Recall that, in contrast, the Erlang approximation (or 1-point approximation) requires only $O(1)$ applications of the iterative algorithm. To obtain a variation of the general slice method with similar computational complexity, we introduce another slice method "approximation" whose basic premise is as follows: Instead of computing $\underline{x}^*(k, r)$ for all $k \in \{n_r : \underline{n} \in \mathcal{S}(C)\}$, we approximate $\underline{x}^*(k, r)$ by linear interpolation between pairs of 3 points.

For a given route r, first apply the 1-point approximation for the entire polytope $\bar{\mathcal{S}}(C)$ to obtain the mode of distribution $\underline{x}^*$. Define

$$n^{\max}(r) \triangleq \max\{n_r : \underline{n} \in \mathcal{S}(C)\}.$$

Next, obtain $\underline{x}^*(n^{\max}(r), r)$, the mode of distribution in the slice $n_r = n^{\max}(r)$, using the 1-point approximation as in the general slice method. Finally, obtain $\underline{x}^*(0, r)$, the mode of distribution in the slice $n_r = 0$, using the 1-point approximation. Now for $k \in \{n_r : \underline{n} \in \mathcal{S}(C)\}$, unlike in the general slice method, we will use an interpolation scheme to compute $x^*(k, r)$ as follows:

(a) If $k \le x_r^*$, then

$$\underline{x}^*(k, r) = \underline{x}^* \cdot \frac{k}{x_r^*} + \underline{x}^*(0, r) \cdot \frac{x_r^* - k}{x_r^*}.$$

That is, $x^*(k, r)$ is the point of intersection (in the space $\mathbb{R}^K$) of the slice $x_r = k$ with the line passing through the two points $\underline{x}^*$ and $\underline{x}^*(0, r)$.

(b) For $x_r^* < k \le n_r^{\max}$, let

$$\underline{x}^*(k, r) = \underline{x}^*(n^{\max}(r), r) \cdot \frac{k - x_r^*}{n^{\max}(r) - x_r^*} + \underline{x}^* \cdot \frac{n^{\max}(r) - k}{n^{\max}(r) - x_r^*}.$$

Note that due to the convexity of the polytope $\bar{\mathcal{S}}(C)$, the interpolated $x^*(k, r)$ are inside the polytope. Now, as in the general slice method, we use these $x^*(k, r)$ to compute the approximation of $\mathbb{E}[n_r]$ and subsequently L_r. A pseudo-code for the 3-point slice method can be found in [8].

4. OUR RESULTS

In this section we present our main results, most of the proofs of which are postponed until the next section.

4.1 Recovering an old result

Consider a stochastic loss network with parameters A, $\underline{C}$ and $\underline{\nu}$ that is scaled by N as defined in Section 2. Kelly [11] obtained a fundamental result which shows that, in the scaled system, the stationary probability distribution concentrates around its mode. Therefore, the results of the 1-point approximation are asymptotically exact. We reprove this result using a variational characterization of the stationary distribution, which yields a much simpler (and possibly more insightful) set of arguments.

THEOREM 1. *Consider a loss network scaled by parameter N. Let L_r^N be the exact loss probability of route $r \in \mathcal{R}$. Then*

$$\left| (1 - L_r^N) - \frac{x_r^*}{\nu_r} \right| = O\left(\sqrt{\frac{\log N}{N}} \right). \quad (4)$$

Kelly established the asymptotic exactness of the Erlang fixed-point approximation by using the above result together with the fact that the Erlang fixed-point approximation for a scaled system essentially solves the dual D as N increases.

4.2 Error in Erlang fixed-point approximation

The Erlang fixed-point approximation is quite popular due to its natural iterative solution algorithm and its asymptotic exactness in the limiting regime. However, it is also well known that the Erlang fixed-point approximation can perform poorly in various cases. This is especially true when the load vector $\underline{\nu}$ is such that it falls on the boundary of $\bar{\mathcal{S}}(C)$, i.e., the stochastic loss network is in the critically loaded regime. More precisely, this means $\underline{\nu}$ is such that at least one of the constraints in $A\underline{\nu} \leq \underline{C}$ is tight. It can be readily verified (at least for simple examples) that, when $\underline{\nu}$ is strictly inside or strictly outside $\mathcal{S}(C)$, then the error in the Erlang fixed-point approximation for the scaled network is $O(1/N)$. However, for the boundary, the qualitative error behavior changes, and in particular we prove the following result.

THEOREM 2. *When the vector ν lies on the boundary of $\bar{\mathcal{S}}(C)$,*

$$\|\underline{L}^{\mathcal{E},N} - \underline{L}^N\|_2 = \Omega\left(\sqrt{\frac{1}{N}}\right), \tag{5}$$

where $\underline{L}^{\mathcal{E},N} = (L_r^{\mathcal{E},N})$ is the vector of loss probabilities from the Erlang fixed-point approximation and $\underline{L}^N = (L_r^N)$ is the vector of exact loss probabilities, both for a loss network scaled by N.

4.3 Accuracy of the slice method

The drastically poorer accuracy of the Erlang fixed-point approximation at the boundary (i.e., in the critical regime) from Theorem 2 strongly motivates the need for new and better loss probability approximations. This led to our development of the general "slice method" described in Section 3.3, for which we establish its asymptotic exactness using the variational characterization of the stationary distribution.

THEOREM 3. *For each route $r \in \mathcal{R}$, let $L_r^{S,N}$ be the loss probability estimate obtained from the general slice method for the system scaled with parameter N. Let L_r^N be the corresponding exact loss probability. Then, for any system parameter values, we have*

$$\left|L_r^{S,N} - L_r^N\right| = O\left(\sqrt{\frac{\log N}{N}}\right). \tag{6}$$

This result establishes the asymptotic exactness of the slice method over all ranges of parameters. The proven error bound, which essentially scales as $O(1/\sqrt{N})$, does not imply that it is strictly better than the Erlang fixed-point approximation. We are unable to establish strict dominance of the slice method, but numerical results in Section 6 illustrate that the slice method can convincingly outperform the Erlang fixed-point approximation under critical loading.

4.4 Convergence of algorithms

So far, certain accuracy properties have been established for the iterative algorithms. We now establish the exponential convergence of the iterative algorithm for the general slice method. It is sufficient to state the convergence of the 1-point approximation, since this is used as a subroutine in our slice methods.

THEOREM 4. *Given a loss network with parameter $A, \underline{C}$ and $\underline{\nu}$, let $\underline{z}^{(t)}$ be the vector produced by the 1-point approximation at the end of iteration t. Then, there exists an optimal solution $\underline{y}^*$ of the dual problem D such that*

$$\left\|\underline{z}^{(t)} - \underline{z}^*\right\| \leq \alpha \exp\left(-\beta t\right),$$

where $\underline{z}^* = (z_j^*)$ with $z_j^* = \exp(-y_j^*)$ and α, β positive constants which depend on the problem parameters.

The proof of Theorem 4 is provided in [8].

4.5 Capacity region of inverse function

Suppose a desired vector of loss probabilities $\underline{L} = (L_r)$ is given, either directly or through constraints. We seek to identify the region of capacities $\underline{C}$ that ensures the arrival rate vector $\underline{\nu}$ will be served with loss probabilities of at most $\underline{L}$. This region is of theoretical importance because it characterizes fundamental properties between the link capacity and loss probability vectors. It also can be exploited to efficiently search the feasible region in various optimization problems involving stochastic loss networks. Obviously, the exact Erlang loss formula can not serve this purpose. In fact, even the Erlang fixed-point equations, which still have the basic structure of the Poisson distribution function, turn out to be too complicated for this purpose. We instead determine the region of interest through the following result.

THEOREM 5. *For a loss system scaled by parameter N, there exists a $\delta(N)$ such that for any given feasible loss probabilities L_r^N and any small positive number $\epsilon \ll 1$, the capacity vectors $\underline{C}^N$ that achieve these loss probabilities fall within the region defined by the system of polynomial equations and inequalities*

$$\begin{aligned} \log(1 - L_r^N - \delta(N)N^{-1/2+\epsilon}) &\leq -\sum_j A_{jr}E_j, \\ \rho_j &= \sum_r \nu_r A_{jr} \prod_{i \neq j}(1 - E_i)^{A_{ir}}, \\ \rho_j(1 - E_j) < C_j^N &< \rho_j(1 - E_j) + 1/E_j. \end{aligned}$$

The problem of linear optimization over a region defined by polynomial equations and inequalities is known to be NP-hard, which improves upon the $\sharp P$ complexity for calculating loss probabilities (refer to [24]). Moreover, there exist standard nonlinear optimization methods for studying the geometry of such regions (see, e.g., [14]), as well as polynomial approximations for solving various optimization problems whose feasible region is defined as above. Theorem 5 therefore can be instrumental in improving the efficiency for solving a wide variety of optimization problems involving stochastic loss networks. For example, the above polynomial equations and inequalities can be easily incorporated into the optimization problems considered in [2] by adding the corresponding constraints on L_r and using the methodologies developed in [14] to obtain a near-optimal solution.

5. PROOFS

We now consider the proofs of most of our main results above.

5.1 Proof: Theorem 1

Variational characterization of π. Recall that the stationary distribution π is represented as

$$\pi(\underline{n}) := \frac{1}{G(C)}\exp(Q(\underline{n})), \quad \exp(Q(\underline{n})) = \prod_{r \in \mathcal{R}} \frac{\nu_r^{n_r}}{n_r!}$$

for $\underline{n} \in \mathcal{S}(C)$. Define $\mathcal{M}(C)$ as the space of distributions on $\mathcal{S}(C)$. Clearly, $\pi \in \mathcal{M}(C)$. For $\mu \in \mathcal{M}(C)$, define

$$\begin{aligned} F(\mu) &\triangleq \sum_{\underline{n}\in\mathcal{S}(C)} \mu(\underline{n})Q(\underline{n}) - \sum_{\underline{n}\in\mathcal{S}(C)} \mu(\underline{n}) \log \mu(\underline{n}) \\ &= \mathbb{E}_{\mu}(Q) + H(\mu). \end{aligned}$$

Next, we state a variational characterization of π, which will be extremely useful throughout. This characterization essentially states that π is characterized uniquely as the maximizer of $F(\cdot)$ over $\mathcal{M}(C)$. See [8] for the proof of Lemma 6.

LEMMA 6. *For all $\mu \in \mathcal{M}(C)$,*

$$F(\pi) \geq F(\mu).$$

The equality holds iff $\mu = \pi$. Further, $F(\pi) = \log G(C)$.

Scaled system: A useful approximation. Now, consider the scaled system with parameter N. For any $\underline{n} \in \mathcal{S}(NC)$, this is equivalent to considering $\frac{1}{N}\underline{n} \in \mathcal{S}_N(C)$. Then, π for a scaled system is equivalent to the distribution π_N on $\mathcal{S}_N(C)$ defined, for $\underline{x} \in \mathcal{S}_N(C)$, as

$$\pi_N(\underline{x}) = \pi(N\underline{x}) = \frac{1}{G(NC)} \exp(Q(N\underline{x})).$$

Upon considering $Q(N\underline{x})$, we have

$$\begin{aligned} &\exp(Q(N\underline{x})) = \prod_r \frac{(N\nu_r)^{Nx_r}}{(Nx_r)!} \\ = \;&\exp\left(\sum_r Nx_r \log N\nu_r - \sum_r \log(Nx_r)!\right) \\ = \;&\exp\left(N\log N \sum_r x_r + N\sum_r x_r \log \nu_r - \sum_r \log(Nx_r)!\right) \\ = \;&\exp\Bigg(N\log N \sum_r x_r + N\sum_r x_r \log \nu_r \\ &\quad - N\sum_r x_r \log Nx_r + \sum_r Nx_r + \sum_r O(\log Nx_r)\Bigg) \\ = \;&\exp\Bigg(N\sum_r x_r \log \nu_r - N\sum_r x_r \log x_r \\ &\quad + N\sum_r x_r + \sum_r O(\log Nx_r)\Bigg), \end{aligned}$$

where the above calculations make use of Stirling's approximation:

$$\log M! = M\log M - M + O(\log M).$$

It then follows from these calculations that

$$\begin{aligned} \frac{1}{N}Q(N\underline{x}) &= \sum_r x_r \log \frac{\nu_r e}{x_r} + \frac{1}{N}\left[\sum_r \log(Nx_r)\right] \\ &= q(\underline{x}) + O\left(\frac{\log N}{N}\right), \end{aligned}$$

where

$$q(\underline{x}) = \sum_r x_r \log \frac{\nu_r e}{x_r}. \tag{7}$$

Concentration of π_N. Given the above calculations, we further obtain the following concentration for the distribution π_N, which will be crucial in proving Theorem 1. Refer to [8] for the proofs of Lemmas 7 and 8.

LEMMA 7. *Given any $\varepsilon > 0$, define the set*

$$A_\varepsilon = \left\{\underline{y} \in \mathcal{S}_N(C) : \|\underline{y} - \underline{x}^*\| > \varepsilon\right\}$$

where $\underline{x}^* = \arg\max_{\underline{x}\in\bar{\mathcal{S}}(C)} q(\underline{x})$. Then

$$\pi_N(A_\varepsilon) = O\left(\varepsilon^{-2}\frac{\log N}{N}\right). \tag{8}$$

LEMMA 8. *For any $y \in \bar{\mathcal{S}}(C)$,*

$$q(\underline{y}) \leq q(\underline{x}^*) - \frac{1}{C_*}\left\|\underline{y} - \underline{x}^*\right\|^2.$$

Completing proof of Theorem 1. Using $\varepsilon_k = k\sqrt{\frac{\log N}{N}}$ for the value of ε in the conclusion of Lemma 7, then from (8) we obtain

$$\pi_N\left(|x_r - x_r^*| > \varepsilon_k\right) = O\left(\frac{1}{k^2}\right), \tag{9}$$

which immediately implies

$$\mathbb{E}\left[|x_r - x_r^*|\right] = O\left(\sqrt{\frac{\log N}{N}}\right) \times O\left(\sum_k \frac{1}{k^2}\right) = O\left(\sqrt{\frac{\log N}{N}}\right).$$

Thus,

$$\mathbb{E}[|x_r - x_r^*|] = O\left(\sqrt{\frac{\log N}{N}}\right),$$

and since $L_r = 1 - \frac{\mathbb{E}[x_r]}{\nu_r}$, we have

$$|L_r - L_r^*| = \frac{\mathbb{E}[|x_r - x_r^*|]}{\nu_r} = O\left(\frac{1}{\nu_r}\sqrt{\frac{\log N}{N}}\right).$$

Additional result: Value of $\log G(NC)$. The above results (specifically, Lemma 6 and Lemma 7), lead to a sharp characterization of $\log G(NC)$ for the scaled system as expressed in the following lemma. See [8] for the proof of Lemma 9.

LEMMA 9.

$$\left|\frac{1}{N}\log G(NC) - \max_{x\in\bar{\mathcal{S}}(C)} q(\underline{x})\right| = O\left(\frac{\log N}{N}\right).$$

5.2 Proof: Theorem 2

Kelly proves in [11] that, for any route r,

$$\|(1 - L_r^{\varepsilon,N}) - \frac{x_r^*}{\nu_r}\| = O\left(\frac{1}{N}\right). \tag{10}$$

Hence, the following lemma together with (1) and (10) establishes Theorem 2.

LEMMA 10. *When the vector ν lies on the boundary of $\bar{\mathcal{S}}(C)$,*

$$\left\|\left(\frac{\mathbb{E}_N[x_r]}{\nu_r}\right)_r - \left(\frac{x_r^*}{\nu_r}\right)_r\right\|_2 = \Omega\left(\sqrt{\frac{1}{N}}\right), \tag{11}$$

where $(\frac{\mathbb{E}_N[x_r]}{\nu_r})_r = (\frac{\mathbb{E}_N[x_1]}{\nu_1}, \frac{\mathbb{E}_N[x_2]}{\nu_2}, \ldots)$ is the vector consisting of the expectations for the routes in the scaled (discrete) system with parameter N, and $(\frac{x_r^*}{\nu_r})_r = (\frac{x_1^*}{\nu_1}, \frac{x_2^*}{\nu_2}, \ldots)$.

PROOF. We shall briefly summarize a few of the technical details, referring to [8] for the complete proof of Lemma 10. Let us start with the following claim, the proof of which is provided in [8].

CLAIM 11. *For any $r \in \mathcal{R}$,*

$$|\mathbb{E}_{\bar{\pi}_N}[x_r] - \mathbb{E}_N[x_r]| = O\left(\frac{1}{N}\right).$$

Then from Claim 11, to prove Lemma 10, it suffices to show that

$$\|\mathbb{E}_{\bar{\pi}_N}[x_r] - \nu\|_2 = \Omega\left(\frac{1}{\sqrt{N}}\right). \quad (12)$$

Define

$$S \triangleq \{v \in S^K : \bar{\mathcal{S}}(C) \cap (\nu + tv) \neq \emptyset \text{ for some } t > 0\},$$

where S^K is the unit sphere in $\mathbb{R}^K$. Now, for a given $v \in S$ and $t \in [0, t_v)$ where $t_v = \sup\{t \in \mathbb{R}_+ : (\nu + tv) \in \bar{\mathcal{S}}(C)\}$, define

$$g_N(v,t) \quad \triangleq \quad \exp(q_N(N(\nu + tv))).$$

Then from spherical integration, we obtain

$$\begin{aligned}\mathbb{E}_{\bar{\pi}_N}[\underline{x}] &= \frac{\int_S \int_0^{t_v} (\nu + tv) g_N(v,t) t^{K-1} dt\, dv}{\int_S \int_0^{v_t} g_N(v,t) t^{K-1} dt\, dv} \\ &= \nu + \frac{\int_S v \int_0^{t_v} g_N(v,t) t^K dt\, dv}{\int_S \int_0^{v_t} g_N(v,t) t^{K-1} dt\, dv},\end{aligned}$$

and thus

$$\|\mathbb{E}_{\bar{\pi}_N}[\underline{x}] - \nu\|_2 = \left\| \frac{\int_S v \int_0^{t_v} g_N(v,t) t^K dt\, dv}{\int_S \int_0^{t_v} g_N(v,t) t^{K-1} dt\, dv} \right\|_2. \quad (13)$$

Next, we introduce the following lemma, which will be crucial in proving (12). See [8] for the proof of Lemma 12.

LEMMA 12. *Let the polytope $\bar{\mathcal{S}}(C)$ and an integer ℓ be given. If N is large enough, then for all $v \in S$*

$$\int_0^{t_v} g_N(v,t) t^\ell dt = \Theta\left(N^{-\frac{K}{2}} \exp(NK) \frac{\Gamma\left(\frac{\ell+1}{2}\right)}{N^{\frac{\ell+1}{2}}}\right),$$

where $\Gamma(\cdot)$ is the Gamma function, and the constant hidden in $\Theta(\cdot)$ is uniformly bounded over all $v \in S$.

Finally, let T be a tangent plane of $\bar{\mathcal{S}}(C)$ at the point ν and let $w \in S^K$ be a unit vector that is perpendicular to T and that satisfies $v \cdot w \geq 0$, for any $v \in S$. Then, from (13), we have

$$\begin{aligned}\|\mathbb{E}_{\bar{\pi}_N}[\underline{x}] - \nu\|_2 &= \left\| \frac{\int_S v \int_0^{t_v} g_N(v,t) t^K dt\, dv}{\int_S \int_0^{t_v} g_N(v,t) t^{K-1} dt\, dv} \right\|_2 \\ &\geq \left| w \cdot \frac{\int_S v \int_0^{t_v} g_N(v,t) t^K dt\, dv}{\int_S \int_0^{t_v} g_N(v,t) t^{K-1} dt\, dv} \right| \\ &= \left| \frac{\int_S w \cdot v \int_0^{t_v} g_N(v,t) t^K dt\, dv}{\int_S \int_0^{t_v} g_N(v,t) t^{K-1} dt\, dv} \right| \\ &= \frac{\Theta\left(N^{-\frac{K}{2}} \exp(NK) \frac{\Gamma\left(\frac{K+1}{2}\right)}{N^{\frac{K+1}{2}}}\right) \int_S v \cdot w\, dv}{\Theta\left(N^{-\frac{K}{2}} \exp(NK) \frac{\Gamma\left(\frac{K}{2}\right)}{N^{\frac{K}{2}}}\right) \int_S 1\, dv} \\ &= \Theta\left(\sqrt{\frac{1}{N}}\right), \quad (14)\end{aligned}$$

where we used Lemma 12 and the facts that $\frac{\int_S v \cdot w\, dv}{\int_S 1\, dv} = \Theta(1)$ and $\frac{\Gamma\left(\frac{K+1}{2}\right)}{\Gamma\left(\frac{K}{2}\right)} = \Theta(1)$. From (14) we obtain (12), which completes the proof of Lemma 10. □

5.3 Proof: Theorem 3

Theorem 1 implies that the actual loss probability L_r^N, $r \in \mathcal{R}$, is given by

$$L_r^N = 1 - \frac{x_r^*}{\nu_r} + O\left(\sqrt{\frac{\log N}{N}}\right).$$

Therefore, the proof of Theorem 3 will be implied by showing that for all $r \in \mathcal{R}$

$$L_r^{S,N} \quad = \quad 1 - \frac{x_r^*}{\nu_r} + O\left(\sqrt{\frac{\log N}{N}}\right). \quad (15)$$

This result is established next where the proof crucially exploits our concentration Lemma 7.

From the definition of the "slice method", the estimated loss probability $L_r^{S,N}$ is defined as

$$L_r^{S,N} \quad = \quad 1 - \frac{1}{\nu_r} \frac{\sum_k k \exp(q(\underline{x}^*(k,r)))}{\sum_k \exp(q(\underline{x}^*(k,r)))}. \quad (16)$$

Recall that $\underline{x}^*(k,r)$ is the solution of the optimization problem

$$\begin{aligned} &\text{maximize} && q(\underline{x}) \\ &\text{over} && \underline{x} \in \bar{\mathcal{S}}(C) \ \&\ x_r = k, \end{aligned}$$

further recalling the definition of the function $q(\cdot)$ as

$$q(\underline{x}) = \sum_r x_r \log \nu_r + x_r - x_r \log x_r.$$

Now, consider a route $r \in \mathcal{R}$. In the rest of the proof, we will use

$$\varepsilon = \sqrt{\frac{2C^* \log N}{N}},$$

where $C_* = \max_j C_j$. Further define the following useful subsets

$$\begin{aligned} S(r,N) &\triangleq \{n_r : \underline{n} \in \mathcal{S}_N(C)\}, \\ S_\varepsilon(r,N) &\triangleq \{k \in S(r,N) : \|\underline{x}^*(k,r) - \underline{x}^*\| \leq \varepsilon\}, \\ S_\varepsilon^c(r,N) &\triangleq \{k \in S(r,N) : \|\underline{x}^*(k,r) - \underline{x}^*\| > \varepsilon\}. \end{aligned}$$

Next, we note two facts that will be used to prove appropriate lower and upper bounds which yield the desired result (15). First, Lemma 8 and the above definitions imply that, for $k \in S_\varepsilon^c(r,N)$,

$$\exp\left(Nq(\underline{x}_r^k)\right) \quad \leq \quad \frac{1}{N^2} \exp\left(Nq(\underline{x}^*)\right). \quad (17)$$

Second, it is easy to see there exists $k \in S_\varepsilon(r,N)$ such that

$$\|\underline{x}^*(k,r) - \underline{x}^*\| = O\left(\frac{1}{N}\right).$$

For this k, we have

$$\exp\left(Nq(\underline{x}^*(k,r))\right) \quad = \quad \Theta\left(\exp(Nq(\underline{x}^*))\right). \quad (18)$$

Use of (17)-(18): Lower bound. Since $|S(r,N)| = O(N)$ in the scaled system, (17) and (18) imply that

$$\begin{aligned} &\sum_{k \in S_\varepsilon^c(r,N)} \exp(Nq(x^*(k,r))) \\ &= O\left(\frac{\exp(Nq(x^*))}{N}\right) \\ &\leq O\left(\frac{1}{N} \sum_{k \in S_\varepsilon(r,N)} \exp(Nq(x^*(k,r)))\right). \quad (19) \end{aligned}$$

From (19), the value of ε and the above subset definitions, we obtain the following sequence of inequalities:

$$\begin{aligned}
&\frac{\sum_{k\in S(r,N)} k\exp(Nq(x^*(k,r)))}{\sum_{k\in S(r,N)}\exp(Nq(x^*(k,r)))}\\
&\quad\geq \frac{\sum_{k\in S_\varepsilon(r,N)} k\exp(Nq(x^*(k,r)))}{\sum_{k\in S(r,N)}\exp(Nq(x^*(k,r)))}\\
&\quad\geq \frac{\sum_{k\in S_\varepsilon(r,N)} k\exp(Nq(x^*(k,r)))}{(1+O(1/N))\left(\sum_{k\in S(r,N)}\exp(Nq(x^*(k,r)))\right)}\\
&\quad\geq \frac{1}{1+O(1/N)}(x_r^*-\varepsilon)\\
&\quad= x_r^* - O\left(\sqrt{\frac{\log N}{N}}\right). \qquad (20)
\end{aligned}$$

Use of (17)-(18): Upper bound. For all $k \in S_\varepsilon^c(r,N)$, $|k|$ is bounded by some constant, and therefore we have

$$\begin{aligned}
&\sum_{k\in S_\varepsilon^c(r,N)} k\exp(Nq(x^*(k,r)))\\
&\quad= O\left(\frac{\exp(Nq(x^*))}{N}\right)\\
&\quad\leq O\left(\frac{1}{N}\sum_{k\in S_\varepsilon(r,N)}\exp(Nq(x^*(k,r)))\right). \qquad (21)
\end{aligned}$$

From (21) and the definition of ε, we obtain

$$\begin{aligned}
&\frac{\sum_{k\in S(r,N)} k\exp(Nq(x^*(k,r)))}{\sum_{k\in S(r,N)}\exp(Nq(x^*(k,r)))}\\
&\quad\leq \frac{\sum_{k\in S(r,N)} k\exp(Nq(x^*(k,r)))}{\sum_{k\in S_\varepsilon(r,N)}\exp(Nq(x^*(k,r)))}\\
&\quad\leq (1+O(1/N))\frac{\sum_{k\in S_\varepsilon(r,N)} k\exp(Nq(x^*(k,r)))}{\sum_{k\in S_\varepsilon(r,N)}\exp(Nq(x^*(k,r)))}\\
&\quad\leq (1+O(1/N))(x_r^*+\varepsilon)\\
&\quad= x_r^* + O\left(\sqrt{\frac{\log N}{N}}\right). \qquad (22)
\end{aligned}$$

Finally, equations (20) and (22) together with (16) imply (15), thus completing the proof of Theorem 3.

5.4 Proof: Theorem 5

From Theorem 1 and (10), the error for the Erlang fixed-point approximation is $O(\sqrt{\log N/N})$ for a scaled system with parameter N. Namely, the unique set of blocking probabilities E_j satisfies

$$|1 - L_r^N - \prod_j (1-E_j)^{A_{jr}}| = O\left(\sqrt{\frac{1}{N}}\right).$$

Hence, for any small positive number $\epsilon > 0$, there exists $\delta(N) > 0$ such that

$$|1 - L_r^N - \prod_j (1-E_j)^{A_{jr}}| \leq \delta(N)N^{-1/2+\epsilon}.$$

We therefore have

$$\log(1 - L_r^N - \delta(N)N^{-1/2+\epsilon}) \leq \sum_j A_{jr}\log(1-E_j),$$

and then the inequality $\log(1-E_j) \leq -E_j$ yields

$$\log(1 - L_r^N - \delta(N)N^{-1/2+\epsilon}) \leq -\sum_j A_{jr}E_j. \qquad (23)$$

Meanwhile, we know that E_j is the solution to the Erlang fixed-point equations

$$E_j = E(\rho_j^N, C_j^N)$$

where

$$\rho_j^N = \sum_r N\nu_r A_{jr}\prod_{i\neq j}(1-E_i)^{A_{ir}}, \qquad (24)$$

which is a polynomial of E_j. Although the Erlang formula itself is in a complicated form, this connection enables us to apply the arguments for the one-dimensional relationship between blocking probability and capacity demonstrated in (2). Hence, we obtain

$$\rho_j^N(1-E_j) < C_j^N < \rho_j^N(1-E_j) + 1/E_j; \qquad (25)$$

see Theorem 2.1 in [1]. Note that, for each $j = 1, 2, \cdots, J$, (25) breaks into one linear inequality and one quadratic inequality of C_j^N and E_j. This completes the proof of Theorem 5.

6. EXPERIMENTS

The main contributions of this paper are the theoretical results presented in Sections 3 – 5. However, to illustrate and quantify the performance of our family of slice methods, we consider two different sets of numerical experiments. The first is based on a small canonical loss network topology that is used to investigate the fundamental properties of our slice methods and previous approaches with respect to the scaling parameter N. Then we turn to consider a large set of numerical experiments based on workforce management applications in the IT services industry using real-world data.

6.1 Small loss networks

We consider a small canonical loss network topology comprised of two routes and three links, as illustrated in Figure 1. Both routes share link 2 with links 1 and 3 dedicated to routes 1 and 2, respectively. More precisely, the network is defined by

$$A = \begin{bmatrix} 1 & 0\\ 1 & 1\\ 0 & 1\end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} 2\\ 3\\ 2\end{bmatrix}.$$

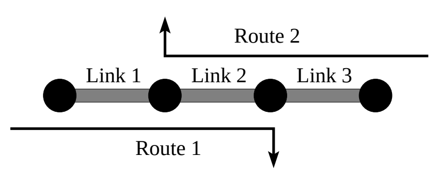

Figure 1: Illustration of the small canonical network model.

In our first collection of experiments, we set $\rho_1 = 2$, $\rho_2 = 1$. The loss probabilities for this network model instance are then computed using our general slice method, the Erlang fixed-point approximation, and the 1-point approximation, where the loss probabilities in each case are considered as a function of the scaling parameter N. Note that, in this small model, the results from the 3-point slice method are identical to those from the general slice method, since the trace of the maximizer point for each slice in the general slice method indeed forms a linear interpolation of the three points. We also directly compute the exact loss probability by brute force and then obtain the average error (over both routes) for each method. These results are presented in Figure 2.

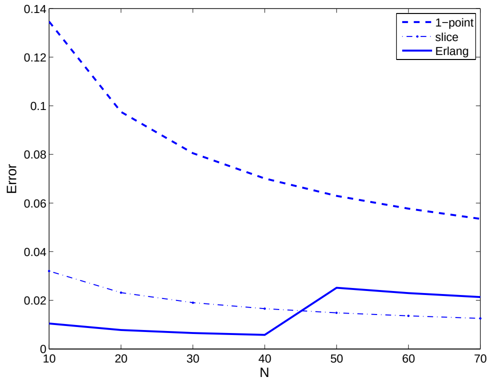

Figure 2: Average error of loss probabilities computed for each method as a function of the scaling parameter N.

First, we observe that the slice method performs better than the 1-point approximation method for every scaling value N. This result is as expected since the slice method utilizes more information about the probability distribution of the underlying polytope than the 1-point method. Second, it is quite interesting to observe that the Erlang fixed-point method initially performs better than the slice method in the small scaling region, whereas the slice method eventually provides the best performance among the approximation methods in the larger scaling region and the performance ordering among the methods shown for $N = 70$ continues to hold for $N > 70$. To understand this phenomena, note that as a function of the scaling with respect to N, the output of the Erlang fixed-point method converges to that of the 1-point approximation method on the order of $O\left(\frac{1}{N}\right)$ and the errors of the 1-point approximation method are given by $\Omega\left(\sqrt{\frac{1}{N}}\right)$, as established in Theorem 2. Moreover, when N becomes larger the error of the slice method becomes smaller than that of the Erlang fixed-point method because the error of the 1-point approximation method is roughly a constant times that of the slice method for every sufficiently large N (as seen in Figure 2). Finally, while the asymptotic exactness of the Erlang fixed-point approximation is associated with the 1-point approximation method, Figure 2 also illustrates some of the complex characteristics of the Erlang fixed-point approximation in the non-limiting regime.

We also consider a second collection of experiments representing the symmetric case of $\rho_1 = \rho_2 = 1.5$. These results exhibit the same trends as in the asymmetric case, and hence are omitted.

6.2 Larger real-world networks

Numerical experiments were also conducted for a large number of real-world loss network instances taken from various resource planning applications within the context of workforce management in the IT services industry. In each of these applications, the network routes represent various IT service products and the network links represent different IT resource capabilities. The various data sets comprising these model instances were obtained from actual IT service companies. First, we generally note that our results from such real-world model instances exhibit trends with respect to the scaling parameter N that are similar to those presented in Section 6.1 for a much simpler canonical model which captures fundamental properties of stochastic loss networks.

In the remainder of this section we shall focus on two representative model instances and present the details of our comparative findings among the slice methods and previous approaches. The first model instance consists of 37 routes and 84 links, whereas the second model instance consists 110 routes and 132 links. In both data sets, the arrival rate vector ν happens to lie on the boundary of $\mathcal{S}(C)$. Figure 3 depicts the sparsity plot for the A matrix of the first model together with the corresponding distributions for the number of routes per link and the number of links per route. Figures 4 and 5 present the same information for the second model instance.

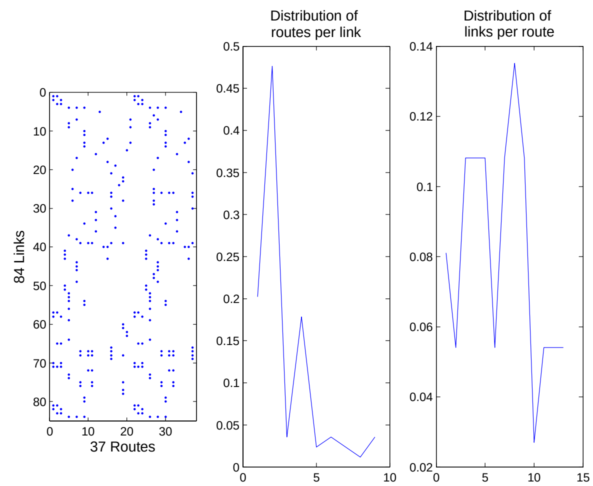

Figure 3: Sparsity plot and distributions of routes/link and links/route for the first model instance.

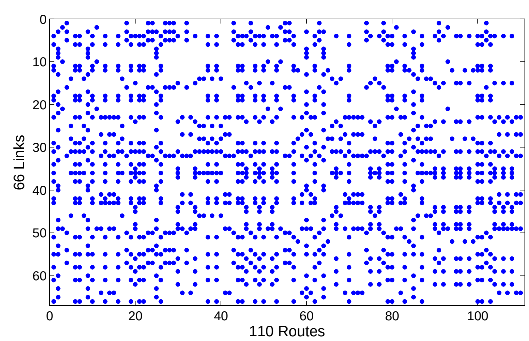

Figure 4: Sparsity plot for the second model instance.

The loss probabilities are computed for each loss network model instance using our general slice method, our 3-point interpolation slice method, and the Erlang fixed-point approximation. Since all of the real-world model instances are too large to numerically compute the exact solution, we use simulation of the corresponding loss network to estimate the exact loss probabilities within tight confidence intervals. The average error (over all routes) and the individual per-route errors are then computed for each method in comparison with the exact loss probabilities, where the former results are summarized in Table 1.

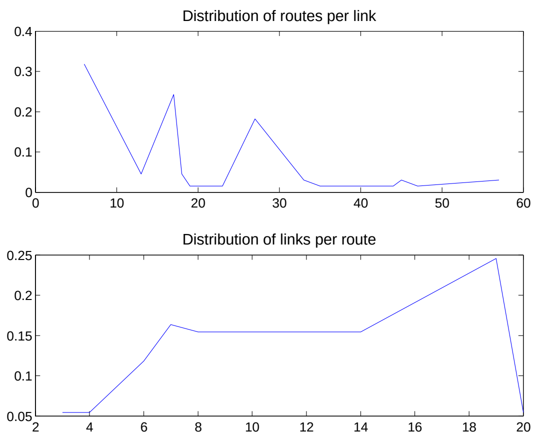

Figure 5: Distributions of routes/link and links/route for the second model instance.

	Erlang	slice method	3-point slice
Model instance 1	0.3357	0.1720	0.1720
Model instance 2	0.3847	0.0923	0.1148

Table 1: Average error of loss probabilities for each method.

The improvements in the approximation errors provided by the general slice method and the 3-point slice method over the Erlang fixed-point method are presented in Figures 6 and 7. Specifically, we plot the relative improvement $\mathcal{I}_r$ in the approximation error for each route r that is obtained with both slice methods, where

$$\mathcal{I}_r \ := \ \frac{|L_r^{\mathcal{E}} - L_r| \ - \ |L_r^S - L_r|}{L_r} \tag{26}$$

with $L_r^{\mathcal{E}}$ and L_r^S denoting the route-r loss probability from the Erlang fixed-point approximation and from one of the slice methods, respectively, and L_r denoting the exact loss probability for route r. Hence, a positive relative improvement $\mathcal{I}_r$ quantifies the benefits of the slice method, a negative relative improvement $\mathcal{I}_r$ quantifies the benefits of the Erlang fixed-point approximation, and $\mathcal{I}_r = 0$ indicates equivalent results from both methods. The average relative improvement (over all routes) of $\mathcal{I} = 0.49$ for the 3-point slice method in model instance 1 is shown by the horizontal line in Figure 6, and the average relative improvements (over all routes) of $\mathcal{I} = 0.76$ and $\mathcal{I} = 0.70$ for the general and 3-point slice methods in model instance 2, respectively, are shown by the points on the y-axis in Figure 7. We note that, in the first model instance, the general slice method and the 3-point slice method provide identical loss probabilities for all routes with one exception where the difference between the loss probabilities from the slice methods for this one route is quite small. Therefore, only the results for the 3-point slice method are presented in Figure 6.

It can be clearly observed from the results in Figures 6 and 7 that the average relative improvements of our slice methods over the Erlang fixed-point approximation are quite significant. Even more importantly, we observe that the relative improvements for the individual routes are consistently and significantly better under both slice methods. In particular, the general (respectively, 3-point) slice method provides the exact loss probabilities for 98 (respectively, 93) of the 110 routes, while the Erlang fixed-point approximation never provides exact results, in model instance 2 and the 3-point slice method provides the exact loss probabilities for 10 of the 37 routes, while the Erlang fixed-point approximation provides the exact results for 5 of these 10 routes, in model instance 1. Note that when $L_r^S = L_r$, the relative error for the Erlang fixed-point approximation, $L_r^{\mathcal{E}}/L_r$, is equal to $1 + \mathcal{I}_r$ (respectively, $1 - \mathcal{I}_r$) when $L_r^{\mathcal{E}} > L_r$ (respectively, $L_r^{\mathcal{E}} < L_r$). In all of the cases where $\mathcal{I}_r = 1.0$, which represents a considerable number of routes in model instance 1 and the overwhelming majority of routes in model instance 2, both slice methods provide the exact loss probability for route r while the Erlang fixed-point approximation yields $L_r^E = 0$, even though the exact loss probabilities for these routes span the full range of values in $(0, 1)$. The loss probability estimates for a few routes are better under the Erlang fixed-point approximation than under the slice methods, but such routes are clearly in the minority representing a single route in model instance 1 and less than 6.5% of the routes in model instance 2.

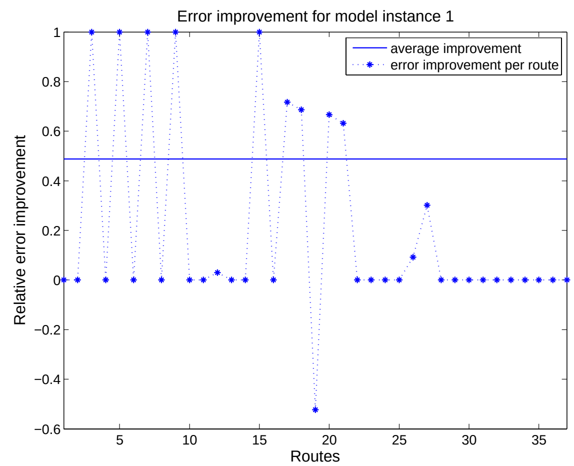

Figure 6: Relative improvement of the approximation errors in loss probabilities for model instance 1.

The above results for two representative examples of a large number of loss networks taken from real-world workforce management applications clearly illustrate and quantify the benefits of our family of slice methods over the classical Erlang fixed-point approximation, at least for the class of loss networks considered.

7. CONCLUSION

Stochastic loss networks have emerged in recent years as canonical models for a wide variety of multi-resource applications, including telephone and communication networks, computer systems, and inventory management and workforce management systems. One of the main performance measures of interest in such applications is the stationary loss probability for each customer class. The Erlang fixed-point approximation is the most popular approach for computing these loss probabilities. However, it is well known that this approximation can provide relatively poor results for various model instances. In particular, we found that the Erlang fixed-point approximation can provide relatively poor loss probability estimates when the network is critically loaded, which is often the natural regime for stochastic loss models of many applications.

Given this motivation, we proposed a general algorithm for estimating the stationary loss probabilities in loss networks based on

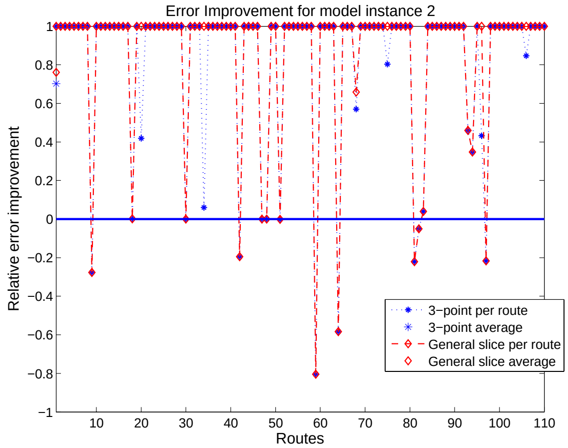

Figure 7: Relative improvement of the approximation errors in loss probabilities for model instance 2.

the properties of "slices" of the exact stationary distribution along the polytope over which it is defined. We established that our algorithm always converges with an exponentially fast rate, where convergence comparisons favor our slice method approach over the Erlang fixed-point approximation. Through a variational characterization of the stationary distribution, we further established that the loss probabilities from our slice method are asymptotically exact. Using this characterization, we also provided an alternative proof of an important result due to Kelly [11], which is simpler and of interest in its own right. Numerical experiments investigate various issues of both theoretical and practical interest. Our general slice method provides an effective approach for computing accurate estimates of the stationary loss probabilities in loss networks.

Stochastic loss networks are often the underlying model in a wide variety of resource allocation and capacity planning problems. One of the difficulties of such optimization problems concern their large feasible regions and the lack of known structural properties on the relationships among resource capacities and loss probabilities for searching through the feasible region. Although bounds on the resource capacity required to achieve a loss probability constraint have been established in the single-resource, single-class case, the corresponding bounds for the multidimensional version represent a significantly more difficult problem. To address this problem, we determined structural properties for the region of resource capacities that ensures offered traffic will be served within a given set of loss probabilities. In addition to the theoretical characterization of relationships between the link capacity and loss probability vectors, our results can be exploited to efficiently search the feasible region of many optimization problems involving stochastic loss networks.

Acknowledgments

The authors thank Sem Borst for helpful comments on Section 6.

8. REFERENCES

[1] S. Berezner, A. Krzesinski, P. Taylor. On the inverse of Erlang's formula. *J. Appl. Prob.*, 35:246–252, 1998.

[2] S. Bhadra, Y. Lu, M. Squillante. Optimal capacity planning in stochastic loss networks with time-varying workloads. In *Proc. ACM SIGMETRICS*, 227–238, 2007.

[3] T. Bonald. The Erlang model with non-Poisson call arrivals. In *Proc. SIGMETRICS-Performance*, 276–286, 2006.

[4] D. Burman, J. Lehoczky, Y. Lim. Insensitivity of blocking probabilities in a circuit-switching network. *J. Appl. Prob.*, 21:850–859, 1984.

[5] A. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. In E. Brockmeyer, H. Halstrom, A. Jensen, eds., *The Life and Works of A.K. Erlang*, Denmark, 1948.

[6] P. Hunt, F. Kelly. On critically loaded loss networks. *Adv. Appl. Prob.*, 21:831–841, 1989.

[7] P. Jelenkovic, P. Momcilovic, M. Squillante. Scalability of wireless networks. *IEEE/ACM Trans. Networking*, 15:295–308, 2007.

[8] K. Jung, Y. Lu, D. Shah, M. Sharma, M. Squillante. Revisiting stochastic loss networks: Structures and algorithms. Research Report, IBM Research, Nov. 2007.

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

[10] F. Kelly. Stochastic models of computer communication systems. *J. Royal Stat. Soc., B*, 47:379–395, 1985.

[11] F. Kelly. Blocking probabilities in large circuit-switched networks. *Adv. Appl. Prob.*, 18:473–505, 1986.

[12] F. Kelly. Routing in circuit-switched networks: Optimization, shadow prices and decentralization. *Adv. Appl. Prob.*, 20:112–144, 1988.

[13] F. Kelly. Loss networks. *Ann. Appl. Prob.*, 1:319–378, 1991.

[14] J. Lasserre. Global optimization with polynomials and the problems of moments. *SIAM J. Opt.*, 11:796–817, 2001.

[15] J. Little. A proof of the queuing formula $L = \lambda W$. *Oper. Res.*, 9:383–387, 1961.

[16] G. Louth, M. Mitzenmacher, F. Kelly. Computational complexity of loss networks. *Theor. Comp. Sci.*, 125:45–59, 1994.

[17] Y. Lu, A. Radovanović, M. Squillante. Optimal capacity planning in stochastic loss networks. *Perf. Eval. Rev.*, 35, 2007.

[18] Z. Luo, P. Tseng. On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM J. Cont. Opt.*, 30:408–425, 1992.

[19] D. Mitra, J. Morrison, K. Ramakrishnan. ATM network design and optimization: A multirate loss network framework. *IEEE/ACM Trans. Networking*, 4:531–543, 1996.

[20] D. Mitra, P. Weinberger. Probabilistic models of database locking: Solutions, computational algorithms and asymptotics. *J. ACM*, 31:855–878, 1984.

[21] K. Ross. *Multiservice Loss Models for Broadband Telecommunication Networks*. Springer-Verlag, 1995.

[22] A. Saleh, J. Simmons. Evolution toward next-generation core optical network. *J. Lightwave Tech.*, 24:3303–3321, 2006.

[23] B. Sevastyanov. An ergodic theorem for Markov processes and its application to telephone systems with refusals. *Theor. Prob. Appl.*, 2:104–112, 1957.

[24] L. Valiant. The complexity of computing the permanent. *Theor. Comp. Sci.*, 8:189–201, 1979.

[25] W. Whitt. Blocking when service is required from several facilities simultaneously. *AT&T Bell Lab. Tech. J.*, 64:1807–1856, 1985.

[26] S. Xu, J. Song, B. Liu. Order fulfillment performance measures in an assemble-to-order system with stochastic leadtime. *Oper. Res.*, 47:131–149, 1999.

[27] I. Ziedins, F. Kelly. Limit theorems for loss networks with diverse routing. *Adv. Appl. Prob.*, 21:804–830, 1989.