Computational Management Science
https://doi.org/10.1007/s10287-019-00355-3

ORIGINAL PAPER

Robustness analysis of generalized Jackson network

Joost Berkhout[1] · Bernd Heidergott[2] · Jennifer Sommer[3] · Hans Daduna[4]

Received: 16 April 2018 / Accepted: 12 July 2019

Abstract
Queuing networks are a well-established approach to modeling and analysis of complex systems. This paper develops an approach to risk-analysis of queuing network models, where "risk" is understood as the possible impact of ignoring parameter insecurity. Our approach allows to compute the value at risk of performance characteristics of queuing networks under parameter insecurity.

Keywords Robustness analysis · Queuing network · Parameter insecurity

1 Introduction

Jackson networks (henceforth JN), to be formally introduced later on, are a well established class of models in, e.g., production, telecommunication, computer systems; for surveys see Kelly (1979) and Chen and Yao (2001). JN's have the desirable property that the distribution of the stationary queue length vector is of product-form, which allows for quick numerical evaluation of performance measures, such as the the mean

✉ Bernd Heidergott
b.f.heidergott@vu.nl

Joost Berkhout
j.berkhout@cwi.nl

Jennifer Sommer
j.sommer@hpc-hamburg.de

Hans Daduna
daduna@math.uni-hamburg.de

[1] Centrum Wiskunde & Informatica (CWI), Science Park 123, 1098 XG Amsterdam,
The Netherlands

[2] Department Econometrics and Operations Research, School of Economics and Business,
De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

[3] HPC Hamburg Port Consulting GmbH, Hamburg, Germany

[4] Department of Mathematics, University of Hamburg, Hamburg, Germany

Published online: 23 July 2019

queue length, mean sojourn times and the throughput at nodes. In this paper we consider JN's with the additional features:

- *simultaneous breakdown and repair of groups of servers* (i.e., repair can be grouped). This allows in particular (i) to model simultaneous breakdown of groups of servers, and (ii) model group repair strategies. For example, repairing several servers simultaneously may lead to more efficient repair actions and thus may reduce the repair time.
- *infinite supply*, where infinite supply has the aim to utilize the capacity of a server to the fullest. For example, in service center models it is typically assumed that an agent, when not answering a call, switches to low priority works such as answering email and administrative duties.

Like for classical JN's, one can obtain for these extended JN's the steady-state distribution of the queue-length vector at stable nodes in product-from for different type of failure-regimes (a precise definition of "stable node" will be provided later in the text), see Sommer et al. (2017). In addition, closed-form solutions for the long-run throughput of subnetworks and of the complete network are provided.

Design and analysis of stochastic networks are often challenged by the fact that the exact specifications of the network are either not known. This is even more so true for models including breakdowns as there is typically only limited information on breakdowns available. Indeed, during usual operation breakdowns are to be avoided and typically only censored observations are available, which is in contrast to, for example, repairs (indeed, repair times are observable and can often be influenced by a decision maker).

Elaborating on the product-form results in Sommer et al. (2017), we will in this paper investigate the impact of the distribution of the time between breakdowns of the individual servers on the throughput of the network. More specifically, we will model the breakdown behavior through a parameterized distribution, and provide a robustness analysis of the system throughput with respect to the uncertainty parameters. Our analysis shows how for different breakdown and repair regimes, the corresponding risk profiles for system-oriented and customer-oriented performance metrics can be evaluated. It is worth noting that this efficient risk analysis step is only possible due to the simple closed-form solutions obtained for the performance measures. The framework provided in this paper allows to combine robustness analysis and performance modeling in an efficient way.

The research for robustness analysis of stochastic models is a predominant research line in Georg Pflug's work, see, for example the monographs (Ermoliev et al. 2006; Pflug 2000). Next to his impressive work on stochastic optimization the study of risk and the investigation on how to deal with uncertainty in stochastic models.

The paper is organized as follows. Section 2 gives a brief introduction to the class of generalized JN's. Robustness analysis is introduced in Sect. 3. The general approach to robustness analysis is presented in Sect. 4. We conclude the paper with discussion of possible future research directions.

2 Jackson networks with breakdowns and repairs

We present a brief review of the theory of JN's with breakdowns and repairs. For details we refer to Sommer et al. (2017). The network consists of J exponential single server nodes with service discipline "First-Come-First-Served" (FCFS), the node set is denoted by $\tilde{J} = \{1, \ldots, J\}$. At node i a Poisson stream with rate $\lambda_i \geq 0$ arrives from the exterior node 0, and service times at node i are exponential with rate μ_i. All service times constitute an independent family of variables which are independent of the arrival streams. Standard customers are indistinguishable and follow the same rules. Routing is Markovian

Nodes in $V \subseteq \tilde{J}$ have an *infinite supply* from which customers are put into an idling server. We denote $W := \tilde{J} \backslash V$ and require $V \neq \emptyset$ (unless otherwise specified). Customers from the infinite supply have low priority, and (standard) customers arriving from the outside or from another server have high priority with preemptive-resume regime: Service of a low priority customer is interrupted as soon as a high priority customer arrives. Service of low priority customers is resumed only when the server idles again. When a low priority customer is served and fed into the network, he becomes a high priority customer and follows the rules for standard customers. Service times of the low priority customers are independent from the external arrival streams and the service times of high priority customers.

Let D denote the set of nodes that can breakdown. The breakdown-repair process $Y = (Y(t) : t \geq 0)$ is Markov on state space $\mathcal{P}(D)$, where $\mathcal{P}(D)$ denotes the power set of D. $Y(t) = I$, for $\emptyset \subseteq I \subseteq D$, indicates that (exactly) the nodes in I are broken down. The transition rates of Y out of $I \subseteq D$ are given as

1. if $I \subset H \subseteq D$, the nodes in $H \backslash I$ break down with rate $\alpha(I, H) \geq 0$,
2. if $\emptyset \subseteq K \subset I$, the nodes in $I \backslash K$ are repaired with rate $\beta(I, K) \geq 0$.

Rates $\alpha(\cdot, \cdot)$ and $\beta(\cdot, \cdot)$ are constructed from any pair of functions $A, B : \mathcal{P}(D) \to [0, \infty)$, subject to (i) $A(\emptyset) = B(\emptyset) = 1$, (ii) $\forall\, I \subset H \subseteq D :\ A(H)/A(I) < \infty$, and (iii) $\forall\, \emptyset \subseteq K \subset I :\ B(I)/B(K) < \infty$ (where we set $0/0 = 0$).

With these functions we set for all subsets of down nodes $I \subseteq D$

$$\alpha(I, H) = \frac{A(H)}{A(I)},\ \ I \subset H \subseteq D,\ \text{ and } \beta(I, K) = \frac{B(I)}{B(K)},\ \ \emptyset \subseteq K \subset I. \qquad (1)$$

Remark 1 With suitable functions A and B we can model, e.g., that nodes may break down isolated or in groups, and repair may happen similarly. It is not required that nodes which are broken down are repaired simultaneously. A statistical procedure to check whether this form is justified, is to determine in a first step all possible values $A(I) = \alpha(\emptyset, I\}$ and $B(I) = \beta(I, \emptyset), \forall I \subseteq D$, and then to check (1) stepwise.

The availability process Y is an ergodic Markov process with stationary distribution

$$\pi(I) = \left(\sum_{K \subseteq D} \frac{A(K)}{B(K)} \right)^{-1} \cdot \frac{A(I)}{B(I)}, \quad \forall I \subseteq D. \qquad (2)$$

From this the stationary (time) point availability (PA) of a Jackson network with infinite supply and unreliable nodes (or subnetworks thereof) may be computed similar to Sauer and Daduna (2003), p.185) as $\mathrm{PA}(H) := \sum_{K \subseteq D \setminus H} \pi(K)$, for $H \subseteq D$ and $t \geq 0$, where $\pi(I)$ is the probability that exactly the nodes in $I \subseteq D$ are under repair as given by (2).

The following regime is set in force whenever a node breaks down:

- service at this node is interrupted, customers (of high as well as of low priority) are frozen there to wait for restart of the service, which is resumed at the point where it was paused,
- no new customers are admitted to enter that node,
- customers who select a broken down node to visit are rerouted according one of the classical rules: stalling, skipping or blocking rs-rd, which will be defined below,
- all these rules, if applicable, are valid for high and low priority customers.

Rerouting is a functional of Y and applies only to high priority customers, because on departure from a node with infinite supply low priority customers are transformed immediately to high priority, and only thereafter are rerouted. We distinguish the following rerouting schemes:

- **Stalling:** Whenever a node breaks down the service system is frozen: All arrival processes are interrupted and service everywhere in the network is stopped until all broken down nodes are repaired again. Stalling is applied, e.g., in the automotive industry to decrease variability of the flow of materials. Indeed, stalling prevents servers to send parts to a server that is broken down and thereby prevents piling up inventory.
- **Skipping:** If as next destination of a customer a down node is selected, the customer jumps to this node, spends no time there, and immediately performs the next jump according to routing regime R until he arrives at a node in up status or leaves the network. Skipping is applied, e.g., in production networks where skipping a production step yields a product of lower but sufficient quality.
- **Blocking rs-rd:** Broken down stations are blocked. A customer whose next destination is down stays at his present node to obtain immediately another service there. After the repeated service (rs) the customer chooses his next destination anew according to R (random destination (rd)). Blocking rs-rd is applied, e.g., in communication networks where packages are rerouted in case a link is not available.

Throughout this article, it is assumed that all nodes in W are stable, i.e., the traffic rate η_i, following from the general traffic equations for Jackson networks with infinite supply but no breakdowns and repairs, is smaller than its service rate μ_i for every node i in W, i.e., $\eta_i < \mu_i$. Without breakdowns, i.e., $D = \emptyset$, the traffic equations of Jackson networks with infinite supply is

$$\eta_i = \lambda_i + \sum_{j \in W} \eta_j r(j,i) + \sum_{j \in V} \mu_j r(j,i), \quad i \in \tilde{J}. \tag{3}$$

By assumption, under **blocking rs-rd** the following reversibility constraints hold:

$$\eta_i r(i,j) = \eta_j r(j,i) \quad \forall i,j \in W, \tag{4}$$

$$\eta_i r(i,j) = \mu_j r(j,i) \quad \forall i \in W, j \in V, \tag{5}$$

and in case of **skipping** the following rate stability constraints hold:

$$\eta_i = \mu_i, \quad \forall i \in V \cap D. \tag{6}$$

These constraints ensure that solution η_i, $i \in \tilde{J}$, from (3) is also the solution of the traffic equations for unreliable Jackson networks under any breakdown scenario. For more details see Sommer et al. (2017). Under these assumptions, we provide an overview on the studied performance characteristics in the following. For an overview of other performance characteristics see Sommer et al. (2017).

Under **stalling** the stationary throughput at nodes $i \in W$ (no infinite supply) is

$$\eta_i \cdot \pi(\emptyset).$$

Under **blocking rs-rd** and **skipping** the stationary throughput at a node $i \in W$ is

$$\eta_i \cdot \sum_{I \subseteq D, i \notin I} \pi(I).$$

To simplify the presentation, we will in the following only consider individual breakdowns and repairs, where for a server $i \in D$ the breakdown rate will be denoted by τ_i and the repair rate by ρ_i. We conclude this section with a short discussion on the robustness of JN's as modeling class for stochastic networks.

Remark 2 Suppose that the physical layout of the network, i.e., the number of nodes, and the topology are known, as well as mean service times, mean inter-arrival times of customers at the network and routing decisions. Provided that this rough information constitute the only available data in advance, arguments exploiting entropy properties lead to use so-called product-form models as conservative first order models. Indeed, if mean service times and mean inter-arrival times are given, the exponential distribution is known to maximize the entropy over all distributions with support $\mathbb{R}_+ := [0, \infty)$ and these means,[1] see, for example, Park and Bera (2009), Lisman and van Zuylen (1972). Furthermore, for given arrival and service rates at the stations, a product-form solution maximizes the entropy of the stationary queue length distribution (Ferdinand 1970; Walstra 1985). Therefore, product-form solutions are conservative and robust models with respect to model insecurity in the service time and inter-arrival time distributions. Hence, working with a model with exponentially distributed service times and inter-arrival times that does have a product-form solution for the stationary queue length distribution, provides a robust model for performance analysis.

[1] Loosely speaking, entropy can be seen in this context as the amount of uncertainty. Results based on distributions given certain means with maximal entropy should be 'least surprising' in terms of predictions that follow from the model. Therefore, the most conservative probabilistic model for service times (or inter-arrival times) with support $\mathbb{R}_+$ consistent with a given mean value is an exponential distribution.

3 Robustness analysis

As pointed out in the introduction, breakdown rates are hard to estimate via historical data. Therefore, one is typically confronted with uncertainty about the true value of the parameters defining the distributions of the time between breakdowns, in our case the failure rate. This is known as *parameter uncertainty* in the literature, see, for example, Haverkort and Meeuwissen (1992), for a discussion on integration of parameter uncertainty into queueing models and Henderson (2003) for a discussion on parameter insecurity from a broader perspective. In the following, the focus is on robustness analysis of our queuing model with respect to uncertainty about the breakdown rates.

In modeling parameter uncertainty, the choice of the distribution is of importance and one typically chooses a particular distribution based on (possible incomplete) knowledge that is available. For example, if the mean and the variance are known, and if, in addition, we know that the parameter may take values in $\mathbb{R}$, the most general distribution is the normal distribution, where "most conservative" refers to the fact that this distribution maximizes the entropy. On the other hand, when, due to expert knowledge, it is known that the parameter falls into an interval, say, $[a, b]$, then the uniform distribution on $[a, b]$ is the entropy maximizing distribution; see, for example, Kullback (1959). Alternatively, there may be statistical knowledge available on θ based on measurements. Then, the distribution of the statistic used for estimating θ is a natural candidate for the distribution of θ.

Formally, we assume that the breakdown rate τ is a random variable defined on some underlying probability field, and that the probability density function for τ, denoted by f_τ, is known. We let $h(\tau)$ denote some reward function. Think, for example, of h as the stationary throughput at node i. Provided that h is invertible and the inverse is differentiable with respect to the throughput

$$g(y) = f_\tau(h^{-1}(y)) \left| \frac{d}{dy} \left(h^{-1}(y) \right) \right| \tag{7}$$

yields the density of the stationary throughput. Based on the distributional assumptions or statistical information comprised in f_τ, one may, as is common practice in applied probability, take the expected value of τ, denoted by $\mu_\tau = \int y f_\tau(y) dy$, as a noise-free approximation of τ and subsequently $h(\mu_\tau)$ as output for the throughput. Since $h(\mu_\tau)$ is typically not close to $\mathbb{E}[h(\tau)]$, simply using μ_τ instead of τ falls short of bringing the risk incurred by the insecurity on τ to light. To analyze the impact τ has on $h(\tau)$, we consider the value at risk of $h(\tau)$, denoted in short by $\mathrm{VaR}(\alpha)$, where $\mathrm{VaR}(\alpha) = q$ if and only if

$$G^{-1}(\alpha) = q,$$

where $G(\cdot)$ denotes the cumulative distribution function of $h(\tau)$, which is, for ease of presentation, assumed to be continuous and invertible. The potential misspecification at an α probability level is thus $h(\mu_\tau) - \mathrm{VaR}(\alpha)$. Note that for the throughput we want

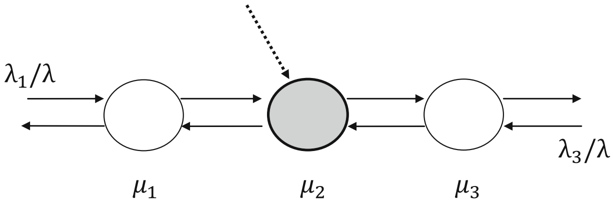

Fig. 1 The two-way-tandem network as analyzed in Sect. 3.1

to hedge against the risk of low values, so we use the α-quantile, whereas for cost functions one would measure the risk through the $(1-\alpha)$-quantile.

In the following, we make the reasonable assumption that the "true" breakdown rate at node i, that is, τ_i, is not revealed to us and we therefore assume that τ_i follows a given distribution F_i. Instances of the throughput can be easily obtained by sampling the τ_i's according to their assumed distribution and evaluating the realization of the stationary throughput. Creating a sufficient number of samples, the density and the cumulative distribution function of the throughput can be estimated and evaluated for further robustness analysis. We would like to point out that a similar robustness analysis can be performed for other performance measures of the queuing network as given (Sommer et al. 2017).

We illustrate the application of the above results to robustness analysis with the help of the following examples.

3.1 A tandem system

Consider the network with $J=\{1,2,3\}$ nodes given in Fig. 1. The network is a two-way tandem of three nodes. The infinite supply is depicted by a dashed arrow pointing to server 2, and the node that is prone to failure is depicted as a grey circle. Note that by incorporating node 0, the linear topology is transformed into a ring. To summarize, $V=\{2\}$, $W=\{1,3\}$, and $D=V$, i.e., the infinite supply node is prone to failure. Routing is given by

$$\begin{aligned} r(1,2)=a, r(1,0)=1-a, r(2,3)=b, r(2,1) \\ =1-b, r(3,0)=c, r(3,2)=1-c, \end{aligned}$$

and

$$r(0,1)=\frac{\lambda_1}{\lambda_1+\lambda_3}, r(0,3)=\frac{\lambda_3}{\lambda_1+\lambda_3},$$

for $0<a,b<1$ and $\lambda_i>0, i=1,2$.

For ease of analysis, we parameterize the model and set $\lambda_1=(1-a)t$, for $t>0$, and $\lambda_3=at$, and $b=1-c$. Regarding the service rates it holds that

$$\mu_1>t,\ \mu_2=t\frac{a}{c} \quad \text{and} \quad \mu_3>t\frac{a}{c}.$$

In this model, we assume for a breakdown scenario that the rate with which a breakdown of server $i = 2$ occurs is given by τ_2, and we denote the corresponding repair rate by ρ_2. Then,

$$\pi(\emptyset) = \frac{\rho_2}{\rho_2 + \tau_2}$$

and the throughput at node 3 under stalling as a function of a certain τ_2 is given by

$$h(\tau_2) = \frac{\eta_3 \rho_2}{\rho_2 + \tau_2},$$

and by computation

$$h^{-1}(y) = \frac{\eta_3 \rho_2}{y} - \rho_2 \quad \text{and} \quad \frac{d}{dy} h^{-1}(y) = -\frac{\eta_3 \rho_2}{y^2}.$$

Let h denote the stationary throughput at node $i = 3$ and model τ_2 as being random. In the following, two distributions for τ_2 are elaborated, the uniform and exponential distribution, respectively:

- Assume that τ_2 is uniformly distributed on $[a, b]$, with $0 < a < b < \infty$. Then,

  $$g(y) = \frac{1}{b-a} \frac{\eta_3 \rho_2}{y^2}, \quad \text{for } \frac{\eta_3 \rho_2}{\rho_2 + b} \leq y \leq \frac{\eta_3 \rho_2}{\rho_2 + a},$$

  and zero otherwise. It holds that $\mu_{\tau_2} = (a + b)/2$. The value at risk, i.e., the α-quantile of the stationary throughput, is also easily computable to be

  $$\mathrm{VaR}(\alpha) = \frac{\eta_3 \rho_2}{\rho_2 + b - (b - a)\alpha},$$

  for $\alpha \in [0, 1]$. In words, for $\alpha \cdot 100\%$ of the possible breakdown rates the actual throughput of the system will fall below $\mathrm{VaR}(\alpha)$. Observe, that for $\alpha = 1$ we have $\mathrm{VaR}(\alpha) = \eta_3 \rho_2 / (\rho_2 + a)$, which is the right bound of the support of τ_2, and for $\alpha = 0$ we have $\mathrm{Var}(\alpha) = \eta_3 \rho_2 / (\rho_2 + b)$, which is the left bound of the support of $h(\tau_2)$.

  Suppose $b = k \cdot a$ with $k > 1$, then τ_2 is uniformly distributed on $[a, ka]$, and τ_2 becomes rather uncertain for large values of k. Then the above analysis uncovers the exposed risk by expecting the throughput to be of order $h(\mu_{\tau_2})$ without taking the stochasticity into account. In particular, with chance α the realized throughput $h(\tau_2)$ is at least

  $$\left(1 - \frac{\frac{2\rho_2}{a} + k + 1}{\frac{2\rho_2}{a} + 2(1-\alpha)k + 2\alpha}\right) \cdot 100\% \tag{8}$$

  smaller than $h(\mu_{\tau_2})$. For example, let $\alpha = 0.1$, then with probability 0.1 the actual throughput $h(\tau_2)$ is at least approximately 44.4% smaller than $h(\mu_{\tau_2})$ for relatively large k.

- Let τ_2 be exponentially-λ-distributed so that $\mu_{\tau_2} = 1/\lambda$. Then the density for the throughput via (7) equals

$$g(y) = \frac{\lambda \eta_3 \rho_2}{y^2} \exp\left(-\lambda \rho_2 \left(\frac{\eta_3}{y} - 1\right)\right),$$

for $y \in (0, \eta_3]$. It can be shown that the cumulative distribution function of $g(y)$ is given by

$$G(y) = \exp\left(-\frac{\lambda \eta_3 \rho_2}{y} + \lambda \rho_2\right), \quad \text{for } y \in (0, \eta_3], \tag{9}$$

which leads to

$$\mathrm{VaR}(\alpha) = \frac{\lambda \rho_2 \eta_3}{\lambda \rho_2 - \ln(\alpha)}, \tag{10}$$

for $\alpha \in (0, 1)$. Note that the "naive" throughput is given by

$$h(1/\lambda) = \frac{\lambda \rho_2 \eta_3}{\lambda \rho_2 + 1}$$

and the difference between $h(1/\lambda)$ and $\mathrm{VaR}(\alpha)$ expresses the model risk at probability α.

Remark 3 Consider the uniform model for τ_2 and consider the reasonable case that the breakdown rate is small, i.e., assume that τ_2 is close to zero. More specifically, let $\tau_2 \sim U(0,\ 2 \cdot \epsilon)$, and assume that $\rho_2 = c \cdot \epsilon$, for $c > 1$. It then holds for the relative error that

$$\frac{h(\mathbb{E}[\tau_2]) - \mathrm{VaR}(\alpha)}{h(\mathbb{E}[\tau_2])} = 1 - \frac{1 + c}{2(1 - \alpha) + c},$$

which implies that $\mathrm{VaR}(\alpha)$ is $(1 - \frac{1+c}{2(1-\alpha)+c}) \cdot 100$ percent smaller than $h(\mathbb{E}[\tau_2])$ for $\alpha \leq 1/2$, and $\mathrm{VaR}(\alpha)$ is $(\frac{1+c}{2(1-\alpha)+c} - 1) \cdot 100$ percent lager than $h(\mathbb{E}[\tau_2])$ for $\alpha \geq 1/2$. This reasoning allows for a quick assessment of the impact of the postulated model on the breakdown rate.

3.2 A star-like system

Consider a network with $J = \{1, 2, \ldots, 6\}$, $V = \{2, 3, 4\}$, $W = \{1, 5, 6\}$, and $D = V$, i.e., all infinite supply nodes are prone to failure, depicted by grey circles in Fig. 2. Jobs arrive from outside with rate λ at the central node 1. From node 1 they go with probability $r/5$ to any of the nodes 2 to 6, for $r \in (0, 1)$. After finishing service at node $i = 2, \ldots, 6$, jobs are sent back to the central node 1. Being served there, they either leave the system with probability $1 - r$, or are sent back to one of the servers

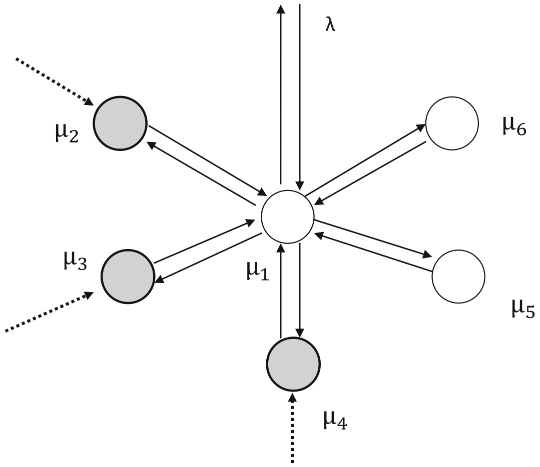

Fig. 2 The star-like network as analyzed in Sect. 3.2

in the set $\{2, \ldots, 6\}$ according to the routing scheme described above. Let h denote the stationary throughput at node $i = 2$ and model τ_2 as being random. Then, the throughput at node 2 as a function of certain τ_2 under skipping is

$$h(\tau_2) = \eta_2 \pi(\emptyset) \left(1 + \frac{\tau_3}{\rho_3} + \frac{\tau_4}{\rho_4} + \frac{\tau_3 + \tau_4}{2 \min(\rho_3, \rho_4)}\right),$$

with $\pi(\emptyset)$ given as

$$\pi(\emptyset) = \left(1 + \sum_{i=2}^{4} \frac{\tau_i}{\rho_i} + \frac{\tau_2 + \tau_3}{2\min(\rho_2, \rho_3)} + \frac{\tau_2 + \tau_4}{2\min(\rho_2, \rho_4)} \right.$$
$$\left. + \frac{\tau_3 + \tau_4}{2\min(\rho_3, \rho_4)} + \frac{\tau_2 + \tau_3 + \tau_4}{3\min(\rho_3, \rho_3, \rho_4)}\right)^{-1}.$$

Letting

$$a_1 = 1 + \frac{\tau_3}{\rho_3} + \frac{\tau_4}{\rho_4} + \frac{\tau_3 + \tau_4}{2\min(\rho_3, \rho_4)} + \frac{\tau_4}{2\min(\rho_2, \rho_4)}$$
$$+ \frac{\tau_3}{2\min(\rho_2, \rho_3)} + \frac{\tau_3 + \tau_4}{3\min(\rho_2, \rho_3, \rho_4)},$$
$$a_2 = \frac{1}{\rho_2} + \frac{1}{2\min(\rho_2, \rho_4)} + \frac{1}{2\min(\rho_2, \rho_3)} + \frac{1}{3\min(\rho_2, \rho_3, \rho_4)},$$

and

$$a_3 = \eta_2 \left(1 + \frac{\tau_3}{\rho_3} + \frac{\tau_4}{\rho_4} + \frac{\tau_3 + \tau_4}{2\min(\rho_3, \rho_4)}\right),$$

we may write for a constant τ_2

$$h(\tau_2) = \frac{a_3}{a_1 + a_2\tau_2}.$$

Hence,

$$h^{-1}(y) = \frac{a_3}{a_2 y} - \frac{a_1}{a_2} \quad \text{and} \quad \frac{d}{dy}h^{-1}(y) = -\frac{a_3}{a_2 y^2}.$$

In the following we show that for uniformly and exponentially distributed τ_2, closed form expressions for the value at risk can be obtained:

- Assume that τ_2 is uniformly distributed on $[a, b]$, with $0 < a < b < \infty$. Then,

$$g(y) = \frac{1}{b-a}\frac{a_3}{a_2 y^2}, \quad \text{for } \frac{a_3}{a_1 + a_2 b} \le y \le \frac{a_3}{a_1 + a_2 a}$$

  and zero otherwise. The value at risk, i.e., the α quantile of the stationary throughput, is also easily computable to be

$$\mathrm{VaR}(\alpha) = \frac{a_3}{a_1 + a_2 b - \alpha(b-a)a_2},$$

  for $\alpha \in [0, 1]$.
- For τ_2 exponentially distributed with parameter λ it holds that

$$g(y) = \lambda \exp\left(-\lambda\left(\frac{a_3}{a_2 y} - \frac{a_1}{a_2}\right)\right)\frac{a_3}{a_2 y^2}, \tag{11}$$

  for $y \in (0, \frac{a_3}{a_1})$, so that

$$G(y) = \exp\left(-\lambda\left(\frac{a_3}{a_2 y} - \frac{a_1}{a_2}\right)\right), \tag{12}$$

  and thus

$$\mathrm{VaR}(\alpha) = \frac{\lambda a_3}{\lambda a_1 - a_2 \ln(\alpha)},$$

  for $\alpha \in (0, 1)$.

As we have shown in this section, for uniform and exponential distributions, VaR(α) can be explicitly solved, which is due to the simplicity of both distributions. In the following we study the more challenging problem when the distribution of τ_2 is of general form.

4 The general approach

In this section we provide a general approach to approximately computing $\mathrm{VaR}(\alpha)$. Revisit the two-way network from example in Sect. 3.1. Let τ_2 be normally distributed with mean μ and standard deviation σ but conditioned on interval $[\gamma_l, \gamma_r]$ where $0 \leq \gamma_l < \gamma_r$. The rationale behind the conditioning is that negative values as well as non-realistically large values for τ_2 are avoided. Then, following (7), the throughput at node 3 under stalling has density g for

$$\frac{\eta_3 \rho_2}{\rho_2 + \gamma_r} \leq y \leq \frac{\eta_3 \rho_2}{\rho_2 + \gamma_l} \tag{13}$$

given by

$$g(y) = \frac{\eta_3 \rho_2}{\Delta_\Phi \sigma \sqrt{2\pi} y^2} \exp\left(-\frac{\left(\frac{\eta_3 \rho_2}{y} - \rho_2 - \mu\right)^2}{2\sigma^2} \right), \tag{14}$$

where

$$\Delta_\Phi = \Phi\left(\frac{\gamma_r - \mu}{\sigma}\right) - \Phi\left(\frac{\gamma_l - \mu}{\sigma}\right)$$

and $\Phi(\cdot)$ is the standard normal cumulative distribution function. To obtain the VaR we need to find the inverse of the cumulative distribution of the throughput given by

$$G(y) = \int_{\frac{\eta_3 \rho_2}{\rho_2 + \gamma_r}}^{y} g(t) dt.$$

Computing the inverse of a general function can usually only be performed numerically. However, in case the function of interest is analytical and can thus be written as a power series, a power series representation of the inverse can be obtained. This result is well-known in analysis, see, e.g., Dettman (2012). However, computing the actual elements of the power series is a challenging task. A first result can be found in Whittaker's pioneering paper (Whittaker 1951). In particular, Whittaker provided an explicit expression for the elements of the power series of the inverse in terms of the elements of the power series of the original function. Unfortunately, the computation of the elements is rather demanding. An alternative approach, that suffers from the same computational burden is Lagrange's inversion formula (Abramowitz and Stegun 1992). Dominici (2003) introduced a method for numerical inversion that is very well suited to computing the VaR of a transformation of an exponentially distributed random variable. In the following we will present this approach.

For an infinitely often differentiable mapping f define the nested derivative $\mathcal{D}^n[f](x)$ by the recursion

$$\mathcal{D}^0[f](x) = 1$$

and

$$\mathcal{D}^n[f](x) = \frac{d}{dx}\Big(f(x)\mathcal{D}^{n-1}[f](x)\Big),$$

for $n \geq 1$. Let

$$h(x) = \int_a^x \frac{1}{f(t)} dt \tag{15}$$

with $f(a) \neq 0, \infty$. Then according to Theorem 4.1 in Dominici (2003) the inverse of $h(x)$ is given by

$$h^{-1}(y) = a + f(a) \sum_{n \geq 1} \mathcal{D}^{n-1}[f](a) \frac{y^n}{n!}, \tag{16}$$

where $|y| < \epsilon$ for some $\epsilon > 0$. The elements of the series can be easily evaluated by means of standard computer algebra tools. We refer to Dominici (2003) for details.

Example 1 Consider the exponential mapping e^x and apply the method of nested derivatives. Note that

$$h(x) := e^x - 1 = \int_0^x e^t dt.$$

Let $f(x) = e^{-x}$, then

$$D^n[f](x) = (-1)^n n! e^{-nx}$$

so that $D^n[f](0) = (-1)^n n!$. Since f is analytical we obtain the inverse of $h(x)$ as

$$h^{-1}(y) = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{y^n}{n},$$

which is easily recognizable as the series expansion of $\ln(y + 1)$ around 0.

As illustrated in the above example, the method of nested derivatives allows for a direct analysis of the function under the integral. This is particularly useful in VaR computations as the analysis can directly be applied to the density and computation of the cumulative distribution function can thus be avoided.

In the following we present the main result on nested derivatives, where we write $\bar{g}(t) = 1/g(t)$.

Theorem 1 *Let $G(y)$ be a cumulative distribution function on $B = [b_l, b_r]$, where $B = \mathbb{R}$ is not excluded. Suppose that there exits $g(t)$, for $t \in B$, such that*

(i) *for $y \in B$ it holds that*

$$G(y) = \int_{b_l}^{y} g(t)dt,$$

(ii) *G is analytical on the interior of B as a mapping in y,*
(iii) *there is an $a \in B$ such that $g(a) \neq 0$.*

Let

$$c_a = \int_{b_l}^{a} g(t)dt = G(a),$$

then

$$\mathrm{VaR}(\alpha) = a + \bar{g}(a) \sum_{n\geq 1} \mathcal{D}^{n-1}[\bar{g}](a) \frac{(\alpha - c_a)^n}{n!},$$

for α sufficiently close to c_a.

Proof For $x \geq a$, write

$$G(x) = c_a + \int_{a}^{x} g(t)\, dt$$

and let

$$G_{c_a}(x) = G(x) - c_a = \int_{a}^{x} \frac{1}{\bar{g}(t)} dt.$$

We now apply the nested derivatives method to $\bar{g}(t)$. From (16) [see also Dominici (2003)] it then follows that

$$G_{c_a}^{-1}(y) = a + \bar{g}(a) \sum_{n\geq 1} \mathcal{D}^{n-1}[\bar{g}](a) \frac{y^n}{n!},$$

for $|y|$ sufficiently small. Noting that $\mathrm{VaR}(\alpha) = G_{c_a}^{-1}(\alpha - c_a)$ concludes the proof. □

Note that the advantage of Theorem 1 lies in the fact that the elements of the series expansion have to be computed once for a, yielding a polynomial approximation of VaR on an entire interval. For ease of reference define for $N \in \mathbb{N}$

$$\mathrm{VaR}(N, \alpha) = a + \bar{g}(a) \sum_{n=1}^{N} \mathcal{D}^{n-1}[\bar{g}](a) \frac{(\alpha - c_a)^n}{n!},$$

so that $\lim_{N\to\infty} \mathrm{VaR}(N, \alpha) = \mathrm{VaR}(\alpha)$.

Example 2 Reconsider the tandem network from Sect. 3.1. Let τ_2 be normally distributed with mean μ and standard deviation σ but truncated on interval $[\gamma_l, \gamma_r]$ where $0 \le \gamma_l < \gamma_r$. See (14) for the density $g(y)$ of the throughput. In order to approximate the VaR by Theorem 1, where a is chosen to be $\frac{\eta_3\rho_2}{\rho_2+\gamma_r}$ so that $c_a = 0$ (note that this is allowed since $g(\frac{\eta_3\rho_2}{\rho_2+\gamma_r}) \neq 0$), we will compute the series using the computer algebra algorithm provided in Dominici (2003). Using notation $\bar{g}(y) = 1/g(y)$ it holds that

$$\bar{g}(y) = \frac{\Delta_\Phi \sigma \sqrt{2\pi}\, y^2}{\eta_3\rho_2} \exp\left(\frac{\left(\frac{\eta_3\rho_2}{y} - \rho_2 - \mu\right)^2}{2\sigma^2}\right),$$

for $\frac{\eta_3\rho_2}{\rho_2+\gamma_r} \le y \le \frac{\eta_3\rho_2}{\rho_2+\gamma_l}$.

It follows from Maple calculations that $\mathrm{VaR}(1, \alpha)$, i.e., a VaR series approximation based on 1 term, equals

$$\begin{aligned}\mathrm{VaR}(1, \alpha) &= a + \bar{g}(a)\alpha \\ &= \frac{\rho_2\eta_3\left(\Delta_\Phi \sigma\sqrt{2\pi}\alpha \exp\left(\frac{(\gamma_r-\mu)^2}{2\sigma^2}\right) + \rho_2 + \gamma_r\right)}{(\rho_2+\gamma_r)^2}.\end{aligned}$$

For the VaR approximation of order 2 we have to add the term

$$\begin{aligned}&\bar{g}(a)\mathcal{D}^1[\bar{g}](a)\frac{\alpha^2}{2} \\ &= \frac{-\eta_3\rho_2(\gamma_r^2 + (\rho_2-\mu)\gamma_r - \rho_2\mu - 2\sigma^2)\alpha^2\Delta_\Phi^2 2\pi \exp\left(\frac{(\gamma_r-\mu)^2}{\sigma^2}\right)}{2(\rho_2+\gamma_r)^3}.\end{aligned}$$

In general, for the n-th term it holds

$$\bar{g}(a)\mathcal{D}^{n-1}[\bar{g}](a)\frac{\alpha^n}{n!} = \frac{(-1)^{n+1}\sigma^2\eta_3\rho_2 P(n)}{n!(\rho_2+\gamma_r)}\left(\frac{\alpha\Delta_\Phi\sqrt{2\pi}\exp\left(\frac{(\gamma_r-\mu)^2}{2\sigma^2}\right)}{\sigma(\rho_2+\gamma_r)}\right)^n,$$

where $P(n)$ is a homogeneous polynomial of degree $2(n-1)$ in variables γ_r, σ, ρ_2 and μ. In particular for $P(n)$ with $n = 1, 2, 3, 4$ it holds

$$\begin{aligned}P(1) &= 1 \\ P(2) &= \gamma_r^2 + (\rho_2-\mu)\gamma_r - 2\sigma^2 - \rho_2\mu \\ P(3) &= 2\gamma_r^4 + (4\rho_2 - 4\mu)\gamma_r^3 + (-5\sigma^2 + 2\rho_2^2 - 8\rho_2\mu + 2\mu^2)\gamma_r^2 + (-4\sigma^2\rho_2 \\ &\quad + 6\sigma^2\mu - 4\rho_2^2\mu + 4\rho_2\mu^2)\gamma_r + 6\sigma^4 + \sigma^2\rho_2^2 \\ &\quad + 6\sigma^2\rho_2\mu + 2\rho_2^2\mu^2\end{aligned}$$

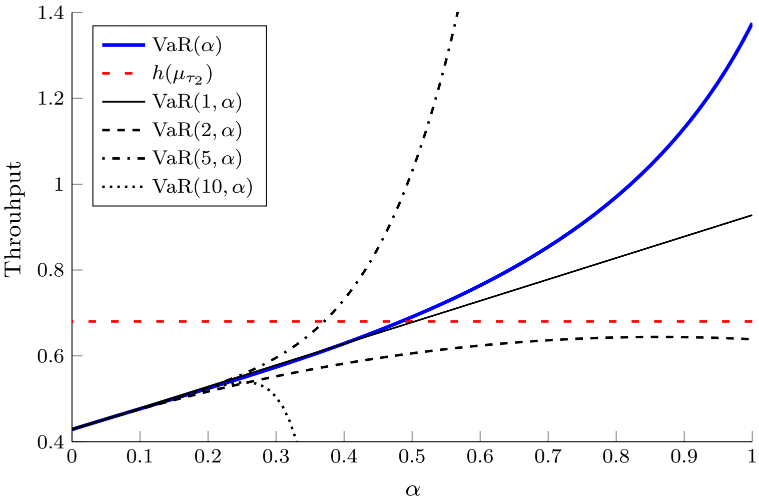

Fig. 3 Plot of VaR(α), VaR(N, α), with $N = 1, 2, 5, 10$, and $c_a = 0$. Numerical values: $\rho_2 = 1.1$, $\eta_3 = 1.5$, $\mu = 1$, $\sigma = 1.5$ and $[\gamma_l, \gamma_r] = [0.1, 2.75]$ so that $E[\tau_2] = 1.3259$ and VaR$(\tau_2) = 0.52134$

$$\begin{aligned}P(4) = {} & 6\gamma_r^6 + (18\rho_2 - 18\mu)\gamma_r^5 + (-15\sigma^2 + 18\rho_2^2 - 54\rho_2\mu + 18\mu^2)\gamma_r^4 \\ & + (-23\sigma^2\rho_2 + 37\sigma^2\mu + 6\rho_2^3 - 54\rho_2^2\mu + 54\rho_2\mu^2 - 6\mu^3)\gamma_r^3 \\ & + (28\sigma^4 - \sigma^2\rho_2^2 + 67\sigma^2\rho_2\mu - 22\sigma^2\mu^2 - 18\rho_2^3\mu + 54\rho_2^2\mu^2 - 18\rho_2\mu^3)\gamma_r^2 \\ & + (20\sigma^4\rho_2 - 36\sigma^4\mu + 7\sigma^2\rho_2^3 + 23\sigma^2\rho_2^2\mu - 44\sigma^2\rho_2\mu^2 \\ & + 18\rho_2^3\mu^2 - 18\rho_2^2\mu^3)\gamma_r - 24\sigma^6 - 8\sigma^4\rho_2^2 - 36\sigma^4\rho_2\mu - 7\sigma^2\rho_2^3\mu \\ & - 22\sigma^2\rho_2^2\mu^2 - 6\rho_2^3\mu^3.\end{aligned}$$

Figure 3 provides a numerical example which illustrates that the VaR series with a few terms already yields an accurate approximation for VaR(α) for α between 0 and 0.1. For the tandem network parameters we take $a = 3/5$, $b = 1/2$, $c = 1/2$, $t = 15/8$, $\mu_1 = \mu_3 = 2$, so that $\lambda_1 = 9/8$, $\lambda_3 = 3/4$ and $\mu_2 = 3/2$. From the traffic equations it follows that $\eta_3 = 1.5$ for this this parameter setting. Regarding the parameter uncertainty parameters, we let $\rho_2 = 1.1$, $\mu = 1$, $\sigma = 1.5$ and $[\gamma_l, \gamma_r] = [0.1, 2.75]$ so that $\mu_{\tau_2} = E[\tau_2] = 1.3259$ and VaR$(\tau_2) = 0.52134$. Furthermore, the example illustrates that significant risk is ignored when taking $h(\mu_{\tau_2})$ as measure for the throughput. Specifically, $h(\mu_{\tau_2}) \approx 0.68$ whereas with probability 0.2 the actual throughput is approximately smaller than 0.52 (a difference of at least 23.5%) and with probability 0.1 the actual throughput is approximately smaller than 0.48 (a difference of at least 29.4%).

In case one is interested in VaR(α) for α around 0.3, Fig. 3 shows that poor approximations are obtained via the series, even when using 10 terms for the series. The approximation for VaR(α) with α around 0.3 can be improved by choosing a in condition (iii) of Theorem 1 greater than $b_l = \frac{\eta_3\rho_2}{\rho_2+\gamma_r}$ such that c_a lies near 0.3. The downside is that this approach requires numerical evaluation of c_a and the search for an appropri-

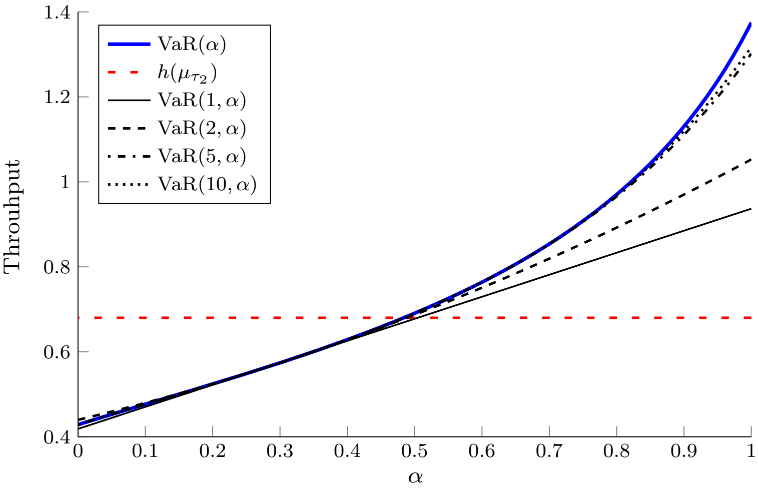

Fig. 4 Plot of $\mathrm{VaR}(\alpha)$, $\mathrm{VaR}(N,\alpha)$, with $N = 1, 2, 5, 10$, and $c_a \approx 0.3$. Numerical values: $\rho_2 = 1.1$, $\eta_3 = 1.5$, $\mu = 1$, $\sigma = 1.5$ and $[\gamma_l, \gamma_r] = [0.1, 2.75]$ so that $E[\tau_2] = 1.3259$ and $\mathrm{VaR}(\tau_2) = 0.52134$

ate a. But after this numerical burden, the series for $\mathrm{VaR}(\alpha)$ from Theorem 1 provides an accurate and efficient approximation for $\mathrm{VaR}(\alpha)$ with α in a relative large interval around c_a. As example, Fig. 4 shows for the same instance as in Fig. 3 that choosing a such that $c_a \approx 0.3$ leads to accurate approximations for $\mathrm{VaR}(\alpha)$ with $\alpha \in (0.15, 0.45)$ even for a small number of series terms.

5 Conclusion

In this paper we have argued the importance of robustness analysis in case of parameter uncertainty in queuing models. For generalized Jackson networks we have provided a framework for evaluating numerically the value at risk incurred by parameter uncertainty. Future research includes uncertainty analysis of multiple parameters and further development of our risk analysis framework. In additional topic of further research is to provide a numerically efficient bound for the remainder of the series approximation of the value at risk.

References

Abramowitz M, Stegun IA (1992) Handbook of mathematical functions with formulas, graphs and mathematical tables. Dower Publications, New York

Chen H, Yao DD (2001) Fundamentals of queueing networks. Springer, Berlin

Dettman JE (2012) Applied complex variables. Dover Publications, New York

Dominici D (2003) Nested derivatives: a simple method for computing series expansions of inverse functions. Int J Math Math Sci 58:3699–3715

Ermoliev Y, Makowski M, Marti K, Pflug GCh (eds) (2006) Coping with uncertainty. Lecture notes in economics and mathematical systems. Springer, Berlin

Ferdinand A (1970) A statistical mechanical approach to systems analysis. IBM J Res Dev 14(5):539–547

Haverkort B, Meeuwissen A (1992) Sensitivity and uncertainty analysis in performance modelling. In: Proceedings 11th symposium on reliable distributed systems. IEEE Computer Society Press, pp 93–102

Henderson S (2003) Input model uncertainty: why do we care and what should we do about it? In: Chick S, Sanchez PJ, Ferrin D, Morrice DJ (eds) Proceedings of the 2003 winter simulation conference, pp 90–100

Kelly FP (1979) Reversibility and stochastic networks. Wiley, Chichester

Kullback S (1959) Information theory and statistics. Wiley, New York

Lisman J, van Zuylen M (1972) Note on the generation of most probable frequency distributions. Stat Neerl 26(1):19–23

Park S, Bera A (2009) Maximum entropy autoregressive conditional heteroskedasticity model. J Econom 150:219–230

Pflug G (2000) Some remarks on the value-at-risk and the conditional value-at-risk. In: Uryasev S (ed) Probabilistic constrained optimization—methodology and applications. Kluwer, Dordrecht, pp 272–281

Sauer C, Daduna H (2003) Availability formulas and performance measures for separable degradable networks. Econ Qual Control 18:165–194

Sommer J, Berkhout J, Daduna H, Heidergott Bernd (2017) Analysis of Jackson networks with infinite supply and unreliable nodes. QUESTA 87:181–207

Walstra R (1985) Nonexponential networks of queues: a maximum entropy analysis. In: SIGMETRICS 85: proceedings of the 1985 ACM SIGMETRICS conference on measurement and modeling of computer systems, New York, p 2737

Whittaker E (1951) On the reversion of series. Gaz Mat Lisboa 12(50):1–12