State-Dependent Estimation of Delay Distributions in Fork-Join Networks

Nitzan Carmeli, Galit B. Yom-Tov
Technion—Israel Institute of Technology,
nitzany@campus.technion.ac.il, gality@technion.ac.il,

Onno J. Boxma
TU/e—Eindhoven University of Technology,
o.j.boxma@tue.nl,

Problem definition: Developing estimators for delay distributions in a queueing network with a Fork-Join structure.
Academic / Practical Relevance: Delay announcements have become an essential tool in service system operations; they influence customer behavior and network efficiency. Most current delay announcement methods are designed for relatively simple call-center environments. Here we focus on a complex Fork-Join network of an Emergency Department. Such systems are mostly in a transient state and although queues for each station are fairly short, delays are long. These delays are the result of both resource scarcity and synchronization delays.
Methodology: We develop an exact analysis of the system. The analysis is based on recursively constructing the Laplace-Stieltjes transform of the joint distributions, conditioning on customers' movements in the network. We then examine the accuracy of the proposed approach on data of an Emergency Department.
Results: Estimations are provided throughout the customer's stay in the network under Markovian assumptions; we then discuss (and provide) relaxations of both service time distribution and structural model assumptions. We provide evidence that the methodology developed is better than Last-to-Enter-Service (LES) estimators (which are based on snapshot principle arguments), reducing Root-Mean-Squared-Error (RMSE) by 25–30%. Our use case demonstrates the importance of incorporating speedup effects of service rates into delay estimators, which are naturally captured by our model, improving accuracy by 10%.
Managerial Implications: The results of our study allow management to implement delay announcements in complex Fork-Join networks. We also highlight the power of using exact approaches instead of relying on approximations, though the latter might be necessary for larger systems.

1. Introduction

Delay announcements have become an essential tool in service system operations. It has been long known that delay announcements serve as a means to reduce customer anxiety and ambiguity while waiting, thus shortening the perceived waiting time and increasing customer satisfaction (Maister 1984, Munichor and Rafaeli 2007). It was further shown, both theoretically (Armony et al. 2009, Jouini et al. 2011) and empirically (Dong et al. 2018, Yu et al. 2016) that providing delay estimations influences customer behavior, which in turn impacts system performance.

In order to provide delay announcements one needs accurate delay estimations. Methodologies for delay estimations have been mostly developed for single station queues (Ibrahim and Whitt

2009a,0,0,0), or for a predetermined route of queues in tandem (Gal et al. 2017). Those are able to capture delays that stem from a lack of resources. However, in service networks, on which the current paper concentrates, there are also synchronization delays that stem from coordination needs in the network. The prediction of the combined effect of those two types of delay is the essence of this paper.

Our primary motivation comes from healthcare systems, such as Emergency Departments (EDs); these are complex queueing networks that involve multiple resource types (physicians, nurses, etc.), multiple patient types (differentiated by severity and medical specialty), long and highly variable processes, and a time-varying environment (Armony et al. 2015). Figure 1 presents a data-based snapshot of real patient flow throughout an Israeli ED. One can clearly see that modeling the ED process as a queueing network imposes a complex structure. Moreover, one of the basic features of this network is the Fork-Join (FJ) structure. A FJ structure models two (or more) activities that can be done in parallel, with different resources, both (or all) of which need to be completed in order to continue the process. Consider, for example, a patient that needs to have an X-ray imaging test, as well as a blood sample tested at the lab before a physician can decide on the rest of the treatment. Those two processes, X-ray and blood test, can be (mostly) done in parallel; hence, the processing times may overlap. FJ networks are prevalent in healthcare systems, but also in other service systems such as the legal system (e.g. when a judge assigns several specialists to produce their opinion before ruling). The goal of this paper is to develop delay estimations for such FJ networks.

Patients in healthcare systems often experience long delays, which are accompanied by confusion, stress and a sense of helplessness (Sullivan et al. 2012, Taylor 1979), as a result of the ambiguity of their service processes and waiting durations. This stress inflicts high pressure on personnel as well. It has often been claimed that providing real-time information for patients on process development and durations may improve patient satisfaction and reduce patients' stress and its negative effects (Altman et al. 2016, Moriah et al. 2011, Rafaeli et al. 2014).

Within healthcare, estimating delays for EDs is especially challenging, first and foremost since the system is small and mostly in transient state. As an example, we present in Figure 2 the average queue length during the day in three stations within an Israeli ED: nurse admission, doctor admission, and lab tests. It is clear that the queue size varies throughout the day, which suggests that a steady-state analysis will not be adequate, and a transient analysis would in fact be necessary. We also observe that queue lengths are fairly small; on average, up to 10 patients are waiting for

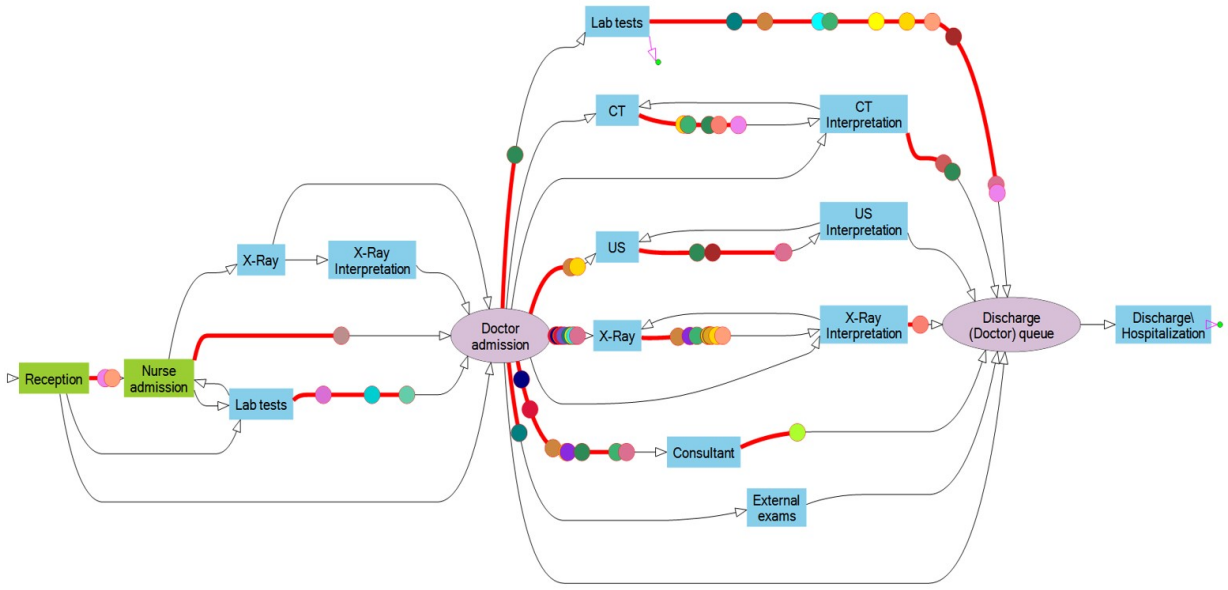


Figure 1 **A snapshot of patient flow throughout an Israeli ED. Nodes represent ED processes, and each colored disc moving along an edge represents a patient who completed the process at the edge origin, and is waiting for, or currently undergoing, the process at the edge destination.**
An animation of the data is available on: `https://youtu.be/HP_au996Ffw`

doctor admission, and no more than 6–7 patients are waiting for nurse admission, or for their lab test results. This suggests that approximation methods may not be accurate either, which led us to develop exact delay estimators for transient-state queueing networks.

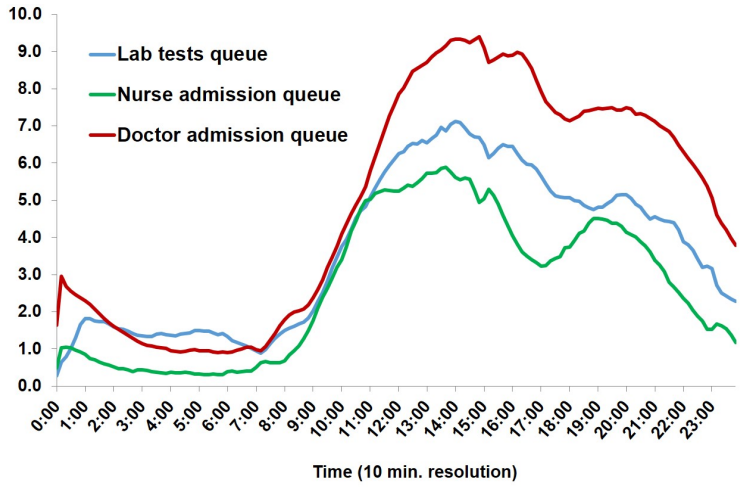
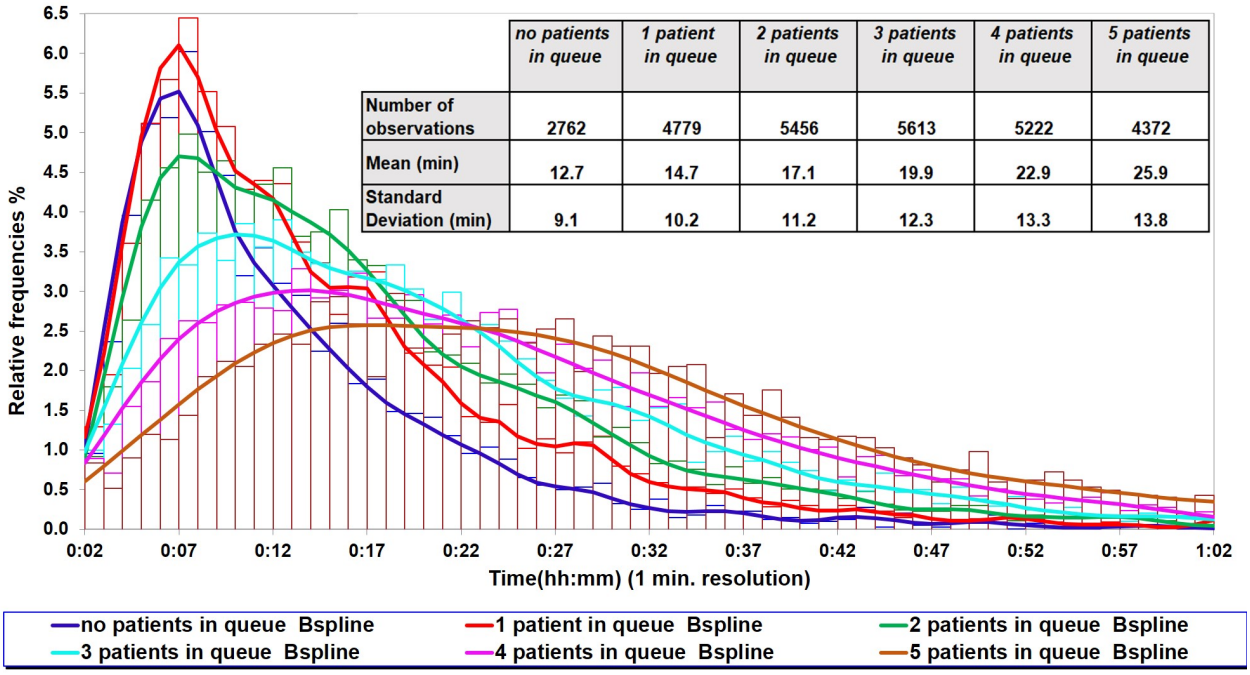


Figure 2 **Average number of patients in queue (April, 2014 - August, 2016, Weekdays)**

The fact that queues for each stage in the treatment process seems to be short, does not mean that waits and sojourn times are short too. Average lengths of stay in EDs are typically long (Armony

et al. 2015), and our use case ED is no different, suffering 7 hours of average delay. Figure 3 presents the sojourn time (waiting time + service time) distribution at the nurse admission station, given the number of patients already in the station (either waiting or undergoing service) at the time of arrival. This provides yet another incentive to consider real-time estimators, as it clearly shows that the sojourn time distribution, as well as its mean and standard deviation, changes significantly with the different queue lengths seen upon patient arrival.



	no patients in queue	1 patient in queue	2 patients in queue	3 patients in queue	4 patients in queue	5 patients in queue
Number of observations	2762	4779	5456	5613	5222	4372
Mean (min)	12.7	14.7	17.1	19.9	22.9	25.9
Standard Deviation (min)	9.1	10.2	11.2	12.3	13.3	13.8

Figure 3 **Sojourn time at nurse admission, as a function of the number of patients already in the station (April, 2014 - August, 2016, Weekdays.) (bars represent histograms in 1-minute resolution, and bold lines show smoothed Bspline functions of relative frequencies)**

.

Some attention also must be given to the amount of information provided to customers. Yom-Tov et al. (2017) showed that delay announcements which are given as a range (e.g. "your wait will be between 45 and 60 minutes") influence customer patience differently than announcing expected wait or a lower bound; and it contributes in reducing customer abandonment. In addition, in a preliminary study of ED announcements, Efrat and Parush (2017) found that customers perceive such range-based delay announcements as more understandable and credible than delay announcements which refer to a single point ("your wait will be approximately 45 minutes" or "your wait will be at least 45 minutes"). Sun et al. (2012) developed a quantile regression model to provide delay

estimations in ED as a range—between the median and the 95^{th} percentile. It was claimed that due to the great variability in patients' waiting time (in this case, between triage and first encounter with a physician) it is extremely challenging to provide a single point estimation (e.g. the mean waiting time), and patients are more satisfied when they receive overestimations rather than underestimations. Jouini et al. (2015) also discussed the importance of estimating a different quantile of the delay distribution, as a means for utility maximization. Therefore, our goal is to provide an estimate for the entire delay distribution, and not just for a specific function of it.

1.1. Contributions

The main contributions of this paper are the following:

1. Developing delay, or sojourn time, estimators for service *networks* which include *FJ* structures, focusing on *individual, real-time, state-dependent* estimation, as opposed to the average, steady-state waiting time of all customers. In this work, we first focus on a network of one single server service station followed by an FJ structure, assuming exponential service times and First-Come-First-Served (FCFS) priority policy. The model is specified in §2. We then relax some of those assumptions in §5.

2. Providing estimations of the waiting (sojourn) time *distribution*, rather than merely means or medians. Such distribution estimation enables announcing delays as a range rather than as a single point. In §3, we use an exact analysis approach to compute the Laplace-Stieltjes Transform (LST) of the *sojourn time in each of the stations*, and the LST of the *total sojourn time*, given the real-time system state at the time of arrival to the first station. The corresponding distributions can then be derived by inverting the LST (for example, by numerical inversion as in Abate and Whitt (1995), Witkovskỳ (2016)).

3. Providing waiting (sojourn) time estimations for *all* stations in a service network, *at the time of arrival to the first station in the network*. This approach can provide answers to questions such as: "when will a customer arrive at the next station? What will be her waiting time once getting there?" Predicting delays at some future point in time is important, and could be used to overcome synchronization problems that stem from providing lagged information (see Dong et al. (2018) for a discussion on the impact of time-lagged delay estimations in networks). This approach has further implications in other domains, too; for example, in project planning and management. One can use our approach to estimate, at the project starting point, when the next stages of the project will start and end, given the current state of all other assignments.

4. *Empirical validation*. We validate our approach in an empirical case study of an Emergency Department, using a unique transaction-level database. This database was established and developed as one of the first steps in our research. It allowed us to test our method's performance in an actual healthcare system. The results are specified in §4. It was found that, although most of the model microscopic assumptions *do not* hold in reality—that is, service durations are not exponential at all stations, priority policy is not FCFS, and even the model structure itself may not accurately represent the actual processes—still in many cases, and specifically for the FJ part tested, our model approximates the actual sojourn time distribution well and better than current approaches. Moreover, this case study revealed some interesting features that influence both system performance and accuracy of estimators. For example, we found that it is essential to incorporate the 'speedup' phenomenon in which service rate grows with the queue length (Chan et al. 2014, Kc and Terwiesch 2009).

5. *Modeling strength.* We expand our basic model to include far more general systems. For example, we show that the 'speedup' phenomenon can be easily incorporated into the model, and by doing so the estimator accuracy is improved by almost 10% (§4.2). We also expand the model to include general service times in the first station (§5.1 and EC A). We further show that one of the FJ stations could be an infinite-server service station, having general service times (see §5.2 and EC B). This means that several activities may be replaced with one single delay node, hence substantially reducing the network complexity. The idea was implemented successfully for ED staffing problems by Yom-Tov and Mandelbaum (2014). We claim it is appropriate here too, and provides a way to address the "curse of dimensionality" common with this type of exact analysis. Furthermore, this approach may be adopted where there is some ambiguity regarding the role of one of the stations in the model—considering it as a delay node may provide the required generalization in order to account for such ambiguity. For example, the lab in our case study provides services to the whole hospital, not just the ED, hence may be considered as a delay node instead of multiple-server station with Exponential service durations.

1.2. Literature review

There is ample literature on waiting time and sojourn time estimators in service systems, mostly concentrating on steady-state behavior. An extensive survey of such results is given by Boxma and Daduna (1990). Our goal, on the other hand, is to provide real-time (state-dependent) estimators for delay distributions, that capture the transient state of the system. Specifically, as explained, we analyze FJ networks. An additional survey on sojourn times in FJ networks is provided by

Boxma et al. (1994). Ko and Serfozo (2004) provided a closed-form approximation for the sojourn time distribution in an FJ network consisting of m $M/M/s$ nodes, in steady state. The sojourn time distribution is approximated as a mixture of the sojourn time distributions in each of the FJ stations. In Ko and Serfozo (2008) the model was extended to an FJ network of m $M/G/1$ nodes. Qiu et al. (2015) suggested a Phase-Type representation of the waiting time distribution and the sojourn time distribution in an FJ network, relying on steady-state assumptions and assuming the difference between the longest queue and the shortest queue is bounded. Rizk et al. (2015) developed computable bounds on the steady-state cumulative distribution functions of the waiting and sojourn time, having both renewal processes and non-renewal processes as an input. All of the above used steady-state analysis to derive expressions for the required distributions. However, healthcare systems, which are the motivation for our work, are rarely in steady state, as is also evident through the data of our ED case study (see for example Figure 2).

Reiman (1982) studied conventional heavy-traffic diffusion approximations for sojourn times in Jackson networks with K single server stations. One of his main results is the snapshot principle: "It is as if the customer takes a snapshot of the network when he enters and all queues remain at the same value during the customer's sojourn through the network." Snapshot principle-based estimators follow the assumption that changes in queue lengths are negligible during a customer visit within the system; this gave rise to the Last-to-Enter-Service (LES) and Head-of-the-Line (HOL) estimators which basically approximate the waiting time of an arriving customer by the waiting time of the last customer to enter service, or the elapsed waiting time of the customer who is currently at the head of the queue. Huang et al. (2015), for example, used the snapshot principle to approximate queue length, virtual waiting time and sojourn times in a single server multi-class queue, specifically, an ED physician treating both newly arrived patients and in-process patients. Nguyen (1993) studied heavy-traffic limits of FJ networks, and showed that under the snapshot principle assumptions, the sojourn time of an arriving customer, in each of the FJ stations, is the average service time in that station times the minimal number of customers in all the station queues. The total sojourn time of an arriving customer can be then regarded as a function of the longest path. We compare our methodology to snapshot-based estimators, though it can only be compared to the average delay, not the whole distribution (§4.2). Note however that the snapshot principle-based estimators are appropriate in critically loaded systems ($\rho \approx 1$) and when the priority policy within each class is FCFS. These criteria may not hold in healthcare systems, which often alternate between overloaded and underloaded periods, and in which the priority policy is rarely FCFS.

Ibrahim and Whitt (2009a,0,0,0) proposed a series of *real-time* waiting time estimators for a single queue with abandonments, time-varying arrival rate and capacity, general service durations, and general inter-arrival and patience distributions. Their predictors could be categorized as two types: queue-length-based (QL) predictors, and delay-history-based predictors (or snapshot principle predictors). Thiongane et al. (2016) expand the delay-history-based predictors to a weighted average form that can cope with multi-class queues. Although these real-time predictors cover a wide range of service systems, they only account for single stations and provide solely an estimation for the average waiting time. Nakibly (2002) focused on estimating waiting time distributions, given the system state at the time of arrival, but this was again for single station service systems, albeit considering skills-based-routing (the motivation came from call centers).

Sun et al. (2012) and Ang et al. (2016) focused on predicting the *"time-to-treatment"* in an ED. This is the time elapsed between the patient registration or triage and the first encounter with a physician or a nurse practitioner. In both papers it was shown that incorporating information about the ED state (the number of patients in queue, the processing rate or the number of providers), as well as estimators such as given in Ibrahim and Whitt (2009a,0), with statistical learning methods (e.g. regularized linear regression) significantly improves the accuracy of the time-to-treatment estimator. In Gal et al. (2017) process-mining techniques were used to predict bus traveling times given a scheduled bus journey. It was also shown that enriching statistical learning methods with queueing perspective variables, such as snapshot-based predictors, results in better prediction models. Similarly to our approach, in Sun et al. (2012) an estimation for a waiting time range ($50^{th}-90^{th}$ percentile) was given, however only for a single station. In Gal et al. (2017) the focus was on delay prediction in networks; yet they do not provide an estimation for the delay distribution, but rather for the mean delay.

From a performance analysis perspective, the most closely related to our work is the literature on closed service networks of two tandem queues, using exact analysis. Stadje (1996) derived a closed-form solution for the total sojourn time within a closed network of two single server queues in tandem with Exponential service durations, *given the number of customers in each of the queues at the time of arrival*. Boxma (1983), Daduna (1986), and Boxma and Daduna (2014) yielded expressions for the joint LST of the sojourn time in the two tandem queues. In those papers, the analysis starts with the assumption that there is a total of N customers in the system, and there are k customers in the first queue, right after a customer departure from the second queue (hence

there are $N-k$ customers in the second queue). The results are then multiplied by the steady-state probabilities of having k customers in the first queue after a departure from the second one. Boxma (1983) considered the case of one queue with Exponential service times followed by a queue with general service times. It was also shown that reversing the order of the queues will yield a different expression for the joint LST. Daduna (1986) generalized the service times distribution to an Erlang-mixture, and Boxma and Daduna (2014) derived an expression for the joint LST of the sojourn times in both queues, having first general service times and then Exponential service times. In our analysis we shall follow that approach and extend it to FJ queues, adding concurrency to the system, which is prevalent in healthcare systems, thus providing an additional important layer to the analysis.

2. The fork-join network model

We consider an open queueing network (see Figure 4) with two parts. The first part consists of a single server queue, followed by the second part with two queues in an FJ structure. We refer to these queues as the 'upper FJ' station (UFJ) and the 'lower FJ' station (LFJ). We assume that service times in the first part are Exponentially distributed with rate λ. The FJ stations can be either single server or multiple server queues, having Exponential service times with rates μ_1 and μ_2, respectively. We assume FCFS service policy throughout the network. From here on we shall assume that both stations in the FJ part have a single server, but this is only in order to simplify notations, and is not a constraint in any way.

We use the following notation to describe the system state at the time of arrival of a *target* customer (for which we would like to provide the delay announcement): k is the number of customers in the first single server station (including customers in queue and in service), h_1 is the number of customers in the UFJ station, and h_2 is the number of customers in the LFJ station. Let S_0 be the target customer sojourn time in the first station, and S_1 (S_2) be her sojourn time in the UFJ (LFJ) station. Let S_{FJ} be the target customer total sojourn time in the FJ part. See Figure 4 for illustration. Note that: $S_{FJ} = \max\{S_1, S_2\}$. Since we are assuming FCFS policy, and the first part queue has a single server, customers entering the system after the target customer will not influence her time in the system. It can therefore be assumed that no customers enter the system after the target customer, and her total sojourn time in the system is actually the time it will take the system to drain.

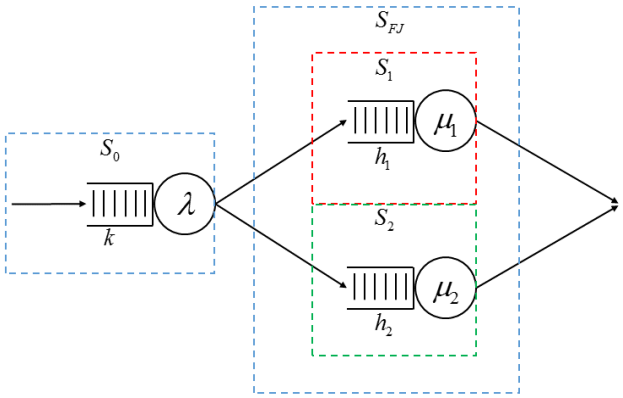


Figure 4 **Two queues in tandem; a single server queue followed by a fork-join queue.**

Let $\Psi_{k,h_1,h_2}(w_0, w_{FJ})$ be the joint Laplace-Stieltjes transform of S_0 and S_{FJ} given k customers in the first station, h_1 customers in the UFJ station and h_2 customers in the LFJ station. If the target customer arrives and finds customers in all three stations, that is, $k, h_1, h_2 > 0$ then,

$$\Psi_{k,h_1,h_2}(w_0, w_{FJ}) \equiv \mathbb{E}_{k,h_1,h_2}\left[e^{-w_0 S_0 - w_{FJ} S_{FJ}}\right] = \int_0^\infty e^{-w_0 t} X_{k,h_1,h_2} \lambda e^{-\lambda t}\, \mathrm{d}t, \tag{1}$$

where X_{k,h_1,h_2} is given by (see the derivation below):

$$\begin{aligned} X_{k,h_1,h_2} = &\sum_{r_1=0}^{h_1-1}\sum_{r_2=0}^{h_2-1} e^{-\mu_1 t}\frac{(\mu_1 t)^{r_1}}{r_1!} e^{-\mu_2 t}\frac{(\mu_2 t)^{r_2}}{r_2!} \Psi_{k-1,h_1-r_1+1,h_2-r_2+1}(w_0, w_{FJ}) \\ &+ \sum_{r_1=0}^{h_1-1}\sum_{r_2=h_2}^{\infty} e^{-\mu_1 t}\frac{(\mu_1 t)^{r_1}}{r_1!} e^{-\mu_2 t}\frac{(\mu_2 t)^{r_2}}{r_2!} \Psi_{k-1,h_1-r_1+1,1}(w_0, w_{FJ}) \\ &+ \sum_{r_1=h_1}^{\infty}\sum_{r_2=0}^{h_2-1} e^{-\mu_1 t}\frac{(\mu_1 t)^{r_1}}{r_1!} e^{-\mu_2 t}\frac{(\mu_2 t)^{r_2}}{r_2!} \Psi_{k-1,1,h_2-r_2+1}(w_0, w_{FJ}) \\ &+ \sum_{r_1=h_1}^{\infty}\sum_{r_2=h_2}^{\infty} e^{-\mu_1 t}\frac{(\mu_1 t)^{r_1}}{r_1!} e^{-\mu_2 t}\frac{(\mu_2 t)^{r_2}}{r_2!} \Psi_{k-1,1,1}(w_0, w_{FJ}). \end{aligned} \tag{2}$$

From the memory-less property of the Exponential distribution, the residual service time of the customer currently in service in the first station is also Exponential with rate λ. (The assumption that service times at the first station are Exponential can be relaxed; an analysis for general service times at the first station is given in §5.1 and EC A).

In order to derive X_{k,h_1,h_2}, we condition the distribution according to changes happening in the FJ part while the first station completed one service. For this, we track the number of customers *potentially* who completed their service by the time that the customer, who is currently in service at the first station, completed service there. We add the term 'potentially' to emphasize that we

account for the number of service completions that could have taken place, if there was an infinite number of customers waiting to be served. There are generally four possible scenarios for the state of the FJ part by the time the customer in service at the first station leaves that station:

1. There are still customers in queue (or in service) in both the UFJ and the LFJ stations. The probability of this scenario is

$$\sum_{r_1=0}^{h_1-1}\sum_{r_2=0}^{h_2-1} e^{-\mu_1 t}\frac{(\mu_1 t)^{r_1}}{r_1!}e^{-\mu_2 t}\frac{(\mu_2 t)^{r_2}}{r_2!} = \left(\sum_{r_1=0}^{h_1-1} e^{-\mu_1 t}\frac{(\mu_1 t)^{r_1}}{r_1!}\right)\left(\sum_{r_2=0}^{h_2-1} e^{-\mu_2 t}\frac{(\mu_2 t)^{r_2}}{r_2!}\right).$$

2. There are still customers in queue (or in service) in the UFJ station, but the server at the LFJ station completed service of all customers that were present at his station when the target customer arrived. Hence, there are no customers in the LFJ station, except for the one who just joined after completing service at the first station. The probability of this scenario is

$$\sum_{r_1=0}^{h_1-1}\sum_{r_2=h_2}^{\infty} e^{-\mu_1 t}\frac{(\mu_1 t)^{r_1}}{r_1!}e^{-\mu_2 t}\frac{(\mu_2 t)^{r_2}}{r_2!} = \left(\sum_{r_1=0}^{h_1-1} e^{-\mu_1 t}\frac{(\mu_1 t)^{r_1}}{r_1!}\right)\left(1-\sum_{r_2=0}^{h_2-1} e^{-\mu_2 t}\frac{(\mu_2 t)^{r_2}}{r_2!}\right).$$

3. There are still customers in queue (or in service) in the LFJ station, but the server at the UFJ station completed service of all customers that were present at that station when the target customer arrived. Hence, there are no customers in the UFJ station, except the one who just joined after completing service at the first station. The probability of this scenario is

$$\sum_{r_1=h_1}^{\infty}\sum_{r_2=0}^{h_2-1} e^{-\mu_1 t}\frac{(\mu_1 t)^{r_1}}{r_1!}e^{-\mu_2 t}\frac{(\mu_2 t)^{r_2}}{r_2!} = \left(1-\sum_{r_1=0}^{h_1-1} e^{-\mu_1 t}\frac{(\mu_1 t)^{r_1}}{r_1!}\right)\left(\sum_{r_2=0}^{h_2-1} e^{-\mu_2 t}\frac{(\mu_2 t)^{r_2}}{r_2!}\right).$$

4. Both the UFJ and the LFJ servers finished serving all customers that were present at their stations when the target customer arrived. Hence, there are no customers in both stations, except for the one who just joined after completing service at the first station. The probability of this scenario is

$$\sum_{r_1=h_1}^{\infty}\sum_{r_2=h_2}^{\infty} e^{-\mu_1 t}\frac{(\mu_1 t)^{r_1}}{r_1!}e^{-\mu_2 t}\frac{(\mu_2 t)^{r_2}}{r_2!} = \left(1-\sum_{r_1=0}^{h_1-1} e^{-\mu_1 t}\frac{(\mu_1 t)^{r_1}}{r_1!}\right)\left(1-\sum_{r_2=0}^{h_2-1} e^{-\mu_2 t}\frac{(\mu_2 t)^{r_2}}{r_2!}\right).$$

Note that the number of potential service completions in the UFJ station and the LFJ station are independent given t—the service duration of the customer in service in the first queue. Thus, we derive X_{k,h_1,h_2} as in (2).

3. Laplace-Stieltjes transform computation

Based on Equations (1) and (2) the joint LST will now be computed recursively. Each iteration corresponds to the number of remaining customers in the first station, until it has no customers, except for the target customer. That is, in the first iteration there are k customers in the first station; in the next iteration there will be $k-1$ customers in the first station, and so on. The joint LST in each iteration is conditioned on the number of potential service completions in the FJ part, by the time the current customer in service in the first station leaves this station. In order to efficiently compute recursively the joint LST, we use the following notation:

$$
\begin{aligned}
a(r_1,r_2) &\equiv \int_0^\infty e^{-(w_0+\mu_1+\mu_2)t}\frac{(\mu_1 t)^{r_1}}{r_1!}\frac{(\mu_2 t)^{r_2}}{r_2!}\lambda e^{-\lambda t}\,\mathrm{d}t,\\
b(r_1,h_2) &\equiv \sum_{r_2=h_2}^{\infty} a(r_1,r_2) = \int_0^\infty e^{-(w_0+\mu_1+\mu_2)t}\sum_{r_2=h_2}^{\infty}\frac{(\mu_1 t)^{r_1}}{r_1!}\frac{(\mu_2 t)^{r_2}}{r_2!}\lambda e^{-\lambda t}\,\mathrm{d}t,\\
c(h_1,r_2) &\equiv \sum_{r_1=h_1}^{\infty} a(r_1,r_2),\\
d(h_1,h_2) &\equiv \sum_{r_1=h_1}^{\infty} b(r_1,h_2).
\end{aligned}
$$

Using the above notation, and a shorthand notation in which we omit the w_0, w_{FJ}, the recursive formula of $\Psi_{k,h_1,h_2}(w_0,w_{FJ}) \equiv \Psi_{k,h_1,h_2}$, given via Equations (1) and (2) (for the case where $k,h_1,h_2>0$), will yield

$$
\begin{aligned}
\Psi_{k,h_1,h_2} = &\sum_{r_1=0}^{h_1-1}\left[\sum_{r_2=0}^{h_2-1}\left(\Psi_{k-1,h_1-r_1+1,h_2-r_2+1}\cdot a(r_1,r_2)\right)+\Psi_{k-1,h_1-r_1+1,1}\cdot b(r_1,h_2)\right]\\
&+\sum_{r_2=0}^{h_2-1}\left(\Psi_{k-1,1,h_2-r_2+1}\cdot c(h_1,r_2)\right)+\Psi_{k-1,1,1}\cdot d(h_1,h_2). \qquad (3)
\end{aligned}
$$

By simple algebraic manipulations we get

$$
\begin{aligned}
a(r_1,r_2) &= \frac{\lambda}{\lambda+w_0}\int_0^\infty(\lambda+w_0)e^{-(w_0+\lambda)t}e^{-\mu_1 t}\frac{(\mu_1 t)^{r_1}}{r_1!}e^{-\mu_2 t}\frac{(\mu_2 t)^{r_2}}{r_2!}\,\mathrm{d}t \qquad (4)\\
&= \binom{r_1+r_2}{r_1}\frac{\lambda}{\lambda+w_0+\mu_1+\mu_2}\left(\frac{\mu_1}{\lambda+w_0+\mu_1+\mu_2}\right)^{r_1}\left(\frac{\mu_2}{\lambda+w_0+\mu_1+\mu_2}\right)^{r_2},\\
b(r_1,h_2) &= \frac{\lambda}{\lambda+w_0}\int_0^\infty(\lambda+w_0)e^{-(w_0+\lambda)t}e^{-\mu_1 t}\frac{(\mu_1 t)^{r_1}}{r_1!}\left(1-\sum_{r_2=0}^{h_2-1}e^{-\mu_2 t}\frac{(\mu_2 t)^{r_2}}{r_2!}\right)\mathrm{d}t,
\end{aligned}
$$

$$c(h_1, r_2) = \frac{\lambda}{\lambda + w_0} \int_0^\infty (\lambda + w_0) e^{-(w_0+\lambda)t} e^{-\mu_2 t} \frac{(\mu_2 t)^{r_2}}{r_2!} \left(1 - \sum_{r_1=0}^{h_1-1} e^{-\mu_1 t} \frac{(\mu_1 t)^{r_1}}{r_1!}\right) \mathrm{d}t,$$

$$d(h_1, h_2) = \frac{\lambda}{\lambda + w_0} \int_0^\infty (\lambda + w_0) e^{-(w_0+\lambda)t} \left(1 - \sum_{r_1=0}^{h_1-1} e^{-\mu_1 t} \frac{(\mu_1 t)^{r_1}}{r_1!}\right) \left(1 - \sum_{r_2=0}^{h_2-1} e^{-\mu_2 t} \frac{(\mu_2 t)^{r_2}}{r_2!}\right) \mathrm{d}t.$$

These integrals can also be evaluated in the same way as the integral expression for $a(r_1, r_2)$. Furthermore, note that if we define X to be a Poisson process with rate μ_1, Y as a Poisson process with rate μ_2, and $\Omega \sim exp(\lambda + w_0)$, then,

$$\begin{aligned}
a(r_1, r_2) &= \frac{\lambda}{\lambda + w_0} \mathbb{P}\left(X(\Omega) = r_1,\, Y(\Omega) = r_2\right), \\
b(r_1, h_2) &= \frac{\lambda}{\lambda + w_0} \mathbb{P}\left(X(\Omega) = r_1,\, Y(\Omega) \geq h_2\right), \\
c(h_1, r_2) &= \frac{\lambda}{\lambda + w_0} \mathbb{P}\left(X(\Omega) \geq h_1,\, Y(\Omega) = r_2\right), \\
d(h_1, h_2) &= \frac{\lambda}{\lambda + w_0} \mathbb{P}\left(X(\Omega) \geq h_1,\, Y(\Omega) \geq h_2\right).
\end{aligned}$$

3.1. Stopping conditions

The recursive computation of the joint LST ends when there are no customers in the first station except for the target customer, that is, when $k = 0$. In this case, the target customer sojourn time in the first station will only be her service time, which is Exponential with rate λ. We define $Z_{x,y}(w_{FJ})$ as the LST of the total sojourn time of the target customer in the FJ part given that there are x customers in the UFJ station, and y customers in the LFJ station, as follows:

$$Z_{x,y}(w_{FJ}) := E\left[e^{-w_{FJ} S_{FJ}} \mid h_1 = x, h_2 = y\right] = \int_0^\infty e^{-w_{FJ} t}\, \mathrm{d}F_{S_{FJ}|x,y}(t). \tag{5}$$

The sojourn time in the UFJ (LFJ) station, of a target customer arriving to that station, and finding x (y) customers there, will be the sum of x (y) service durations, each Exponentially distributed with rate μ_1 (μ_2). Therefore, $(S_1|h_1 = x) \sim Gamma\,(x, \mu_1)$ $((S_2|h_2 = y) \sim Gamma\,(y, \mu_2))$, and from the Gamma-Poisson relationship, we get:

$$\begin{aligned}
F_{S_{FJ}|x,y}(t) &= \mathbb{P}\left(S_{FJ} \leq t \mid h_1 = x, h_2 = y\right) = \mathbb{P}\left(S_1, S_2 \leq t \mid h_1 = x, h_2 = y\right) \\
&= \mathbb{P}\left(S_1 \leq t \mid h_1 = x, h_2 = y\right) \mathbb{P}\left(S_2 \leq t \mid h_1 = x, h_2 = y\right) \\
&= \mathbb{P}\left(S_1 \leq t \mid h_1 = x\right) \mathbb{P}\left(S_2 \leq t \mid h_2 = y\right) \\
&= \left(1 - \sum_{m=0}^{x} e^{-\mu_1 t} \frac{(\mu_1 t)^m}{m!}\right) \left(1 - \sum_{n=0}^{y} e^{-\mu_2 t} \frac{(\mu_2 t)^n}{n!}\right).
\end{aligned}$$

It follows that once $k=0$, the recursive formula is done, since it is only left to calculate $Z_{x,y}(w_{FJ})$ for all possible values of x and y, and the probabilities of having $h_1=x$ and $h_2=y$ by the time the target customer ends service in the first station.

We shall now review the four possible cases in which $k=0$, and the resulting recursive formula (again, for notation simplicity we omit w_{FJ}):

1. $k=0$, $h_1,h_2>0$:

$$\Psi_{0,h_1,h_2}=\sum_{r_1=0}^{h_1-1}\left[\sum_{r_2=0}^{h_2-1}\left(Z_{h_1-r_1+1,h_2-r_2+1}\cdot a(r_1,r_2)\right)+Z_{h_1-r_1+1,1}\cdot b(r_1,h_2)\right] + \sum_{r_2=0}^{h_2-1}\left(Z_{1,h_2-r_2+1}\cdot c(h_1,r_2)\right)+Z_{1,1}\cdot d(h_1,h_2). \tag{6}$$

2. $k=0$, $h_1=0$, $h_2>0$:

$$\Psi_{0,0,h_2}=\sum_{r_2=0}^{h_2-1}Z_{1,h_2-r_2+1}\cdot c(0,r_2)+Z_{1,1}\cdot d(0,h_2). \tag{7}$$

3. $k=0$, $h_2=0$, $h_1>0$:

$$\Psi_{0,h_1,0}=\sum_{r_1=0}^{h_1-1}Z_{h_1-r_1+1,1}\cdot b(r_1,0)+Z_{1,1}\cdot d(h_1,0). \tag{8}$$

4. $k=0$, $h_1,h_2=0$:

$$\Psi_{0,0,0}=\frac{\lambda}{\lambda+w_0}Z_{1,1}. \tag{9}$$

Here the LST $Z_{x,y}\equiv Z_{x,y}(w_{FJ})$ is defined in Equation (5), and $a(\cdot,\cdot)$, $b(\cdot,\cdot)$, $c(\cdot,\cdot)$, and $d(\cdot,\cdot)$ are defined in Equation (4). Note that $(S_{FJ}|1,1)$ is the maximum of two Exponential random variables with rates μ_1 and μ_2, ergo,

$$F_{S_{FJ}|1,1}(t)=\mathbb{P}\left(S_1\le t\mid h_1=1\right)\mathbb{P}\left(S_2\le t\mid h_2=1\right)=\left(1-e^{-\mu_1 t}\right)\left(1-e^{-\mu_2 t}\right).$$

Therefore,

$$\begin{aligned}Z_{1,1}&=\int_0^\infty e^{-w_{FJ}t}\left[\mu_1 e^{-\mu_1 t}\left(1-e^{-\mu_2 t}\right)+\mu_2 e^{-\mu_2 t}\left(1-e^{-\mu_1 t}\right)\right]\mathrm{d}t\\&=\frac{\mu_1}{\mu_1+w_{FJ}}-\frac{\mu_1}{\mu_1+\mu_2+w_{FJ}}+\frac{\mu_2}{\mu_2+w_{FJ}}-\frac{\mu_2}{\mu_1+\mu_2+w_{FJ}}.\end{aligned}$$

For completeness, we shall also review the special cases in which $k>0$ and h_1 or h_2 (or both) equal zero:

1. $k > 0$, $h_1 = 0$, $h_2 > 0$:

$$\Psi_{k,0,h_2} = \sum_{r_2=0}^{h_2-1} \Psi_{k-1,1,h_2-r_2+1} \cdot c(0, r_2) + \Psi_{k-1,1,1} \cdot d(0, h_2).$$

2. $k > 0$, $h_1 > 0$, $h_2 = 0$:

$$\Psi_{k,h_1,0} = \sum_{r_1=0}^{h_1-1} \Psi_{k-1,h_1-r_1+1,1} \cdot b(r_1, 0) + \Psi_{k-1,1,1} \cdot d(h_1, 0).$$

3. $k > 0$, $h_1, h_2 = 0$:

$$\Psi_{k,0,0} = \Psi_{k-1,1,1} d(0,0) = \Psi_{k-1,1,1} \cdot \frac{\lambda}{\lambda + w_0}.$$

3.2. Recursive formula calculation

The recursive formula can be easily and efficiently calculated if written in compact matrix form. We define the following two matrices: The first matrix is $\bar{\Psi}_{k-1,h_1+1,h_2+1}$, given by

$$\bar{\Psi}_{k-1,h_1+1,h_2+1} = \begin{pmatrix} \Psi_{k-1,h_1+1,h_2+1} & \Psi_{k-1,h_1+1,h_2} & \dots & \Psi_{k-1,h_1+1,1} \\ \Psi_{k-1,h_1,h_2+1} & \Psi_{k-1,h_1,h_2} & \dots & \Psi_{k-1,h_1,1} \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{k-1,1,h_2+1} & \Psi_{k-1,1,h_2} & \dots & \Psi_{k-1,1,1} \end{pmatrix},$$

and its dimensions are $(h_1 + 1) \times (h_2 + 1)$. The second matrix, Φ_{h_1,h_2}, is defined as

$$\Phi_{h_1,h_2} = \begin{pmatrix} a(0,0) & a(1,0) & \dots & a(h_1-1,0) & c(h_1,0) \\ a(0,1) & a(1,1) & \dots & a(h_1-1,1) & c(h_1,1) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a(0,h_2-1) & a(1,h_2-1) & \dots & a(h_1-1,h_2-1) & c(h_1,h_2-1) \\ b(0,h_2) & b(1,h_2) & \dots & b(h_1-1,h_2) & d(h_1,h_2) \end{pmatrix}.$$

and its dimensions are $(h_2 + 1) \times (h_1 + 1)$. We now observe that

$$\Psi_{k,h_1,h_2} = trace\left(\bar{\Psi}_{k-1,h_1+1,h_2+1} \cdot \Phi_{h_1,h_2}\right).$$

If there are k customers in the first station, h_1 customers in the UFJ station, and h_2 in the LFJ station, by the time the target customer completes service in the first station, there could be up

to h_1+k customers in the UFJ station, and up to h_2+k customers in the LFJ station. Thus, to recursively compute Ψ_{k,h_1,h_2} we start with the $(h_1+k)\times(h_2+k)$ matrix:

$$\bar{\Psi}_{0,h_1+k,h_2+k}=\begin{pmatrix} \Psi_{0,h_1+k,h_2+k} & \Psi_{0,h_1+k,h_2+k-1} & \dots & \Psi_{0,h_1+k,1} \\ \Psi_{0,h_1+k-1,h_2+k} & \Psi_{0,h_1+k-1,h_2+k-1} & \dots & \Psi_{0,h_1+k-1,1} \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{0,1,h_2+k} & \Psi_{0,1,h_2+k-1} & \dots & \Psi_{0,1,1} \end{pmatrix};$$

each of its elements can be easily computed using Formulas (6)–(9). As a basis for all iterations it will be of use to define the following four $(h_1+k-1)\times(h_2+k-1)$ matrices:

$$A_{h_1+k,h_2+k}=\begin{pmatrix} a(0,0) & a(0,1) & \dots & a(0,h_2+k-2) \\ a(1,0) & a(1,1) & \dots & a(1,h_2+k-2) \\ \vdots & \vdots & \ddots & \vdots \\ a(h_1+k-2,0) & a(h_1+k-2,1) & \dots & a(h_1+k-2,h_2+k-2) \end{pmatrix},$$

$$B_{h_1+k,h_2+k}=\begin{pmatrix} b(0,1) & b(0,2) & \dots & b(0,h_2+k-1) \\ b(1,1) & b(1,2) & \dots & b(1,h_2+k-1) \\ \vdots & \vdots & \ddots & \vdots \\ b(h_1+k-2,1) & b(h_1+k-2,2) & \dots & b(h_1+k-2,h_2+k-1) \end{pmatrix},$$

$$C_{h_1+k,h_2+k}=\begin{pmatrix} c(1,0) & c(1,1) & \dots & c(1,h_2+k-2) \\ c(2,0) & c(2,1) & \dots & c(2,h_2+k-2) \\ \vdots & \vdots & \ddots & \vdots \\ c(h_1+k-1,0) & c(h_1+k-1,1) & \dots & c(h_1+k-1,h_2+k-2) \end{pmatrix},$$

$$D_{h_1+k,h_2+k}=\begin{pmatrix} d(1,1) & d(1,2) & \dots & d(1,h_2+k-1) \\ d(2,1) & d(2,2) & \dots & d(2,h_2+k-1) \\ \vdots & \vdots & \ddots & \vdots \\ d(h_1+k-1,1) & d(h_1+k-1,2) & \dots & d(h_1+k-1,h_2+k-1) \end{pmatrix}.$$

The elements of these matrices are computed by Equation (4).

We shall also use the following notation: let X be a matrix, then $X[l:m,n:p]$ is the resulting matrix by taking all the elements of X in rows l up to m, and columns n up to p.

In the first iteration we compute the matrix $\bar{\Psi}_{1,h_1+k-1,h_2+k-1}$. Each of its elements,

$$\Psi_{1,m,n},\ \forall 1\le m\le h_1+k-1,\ 1\le n\le h_2+k-1$$

will be calculated as follows:

$$\Psi_{1,m,n} = trace\left(\bar{\Psi}_{0,m+1,n+1} \cdot \Phi_{m,n}\right),$$

where

$$\bar{\Psi}_{0,m+1,n+1} = \bar{\Psi}_{0,h_1+k,h_2+k}\left[h_1+k-m : h_1+k\ ,\ h_2+k-n : h_2+k\right],$$

and $\Phi_{m,n}$ is composed of the matrices A, B, C, and D, in the following manner:

$$\Phi_{m,n} = \begin{bmatrix} A[1:m,1:n]^T & C[m,1:n]^T \\ B[1:m,n]^T & D[m,n] \end{bmatrix}.$$

In the second iteration we calculate $\bar{\Psi}_{2,h_1+k-2,h_2+k-2}$. Each of its elements,

$$\Psi_{2,m,n},\ \forall 1 \le m \le h_1+k-2,\ 1 \le n \le h_2+k-2,$$

can be calculated as follows:

$$\Psi_{2,m,n} = trace\left(\bar{\Psi}_{1,m+1,n+1} \cdot \Phi_{m,n}\right).$$

We continue this process, until we reach the k_{th} iteration, in which the desired LST will be:

$$\Psi_{k,h_1,h_2}(w_0, w_{FJ}) \equiv \Psi_{k,h_1,h_2} = trace\left(\bar{\Psi}_{k-1,h_1+1,h_2+1} \cdot \Phi_{h_1,h_2}\right).$$

Assigning $w_0 = 0$ ($w_{FJ} = 0$) will yield the LST of the sojourn time at the first (FJ) part, assigning $w_{FJ} = w_0$ will yield the LST of the target customer total sojourn time in the system. From the LST we obtain the Fourier-Stieltjes transform, and then use the numerical inversion of Witkovskỳ (2016) to retrieve the sojourn time distribution.

4. 'All models are wrong, but some are useful' (George Box)

As this known quote states any model, ours included, is constrained by its assumptions on the reality; assumptions that are mostly wrong. In this section, we will demonstrate via a case study, that despite this fact the proposed method is robust enough to be practical in real, complex, challenging service networks, such as healthcare systems. To do so, we compare estimations based on the model to over two years of data (from April 2014 to August 2016) taken from an Israeli ED. We shall first describe the ED environment focusing on the gaps between the model assumptions and the data. Then, we shall compare and evaluate our estimators accuracy, and discuss the reasons for such accuracy. The structure of the ED process was first established by depth interviews with

various ED personnel—the ED manager, physicians, nurses, and administrators—and then verified by obtaining transaction-level data on all treatments provided throughout the patient's stay.

On average, 150 patients arrive to the ED per day. The ED classify them into two groups non-severe independent *walking* patients (47%), and *acute* patient groups in which each patient is assigned a bed (53%). The average length of stay in the ED for walking patients is approximately 5 hours, for acute patients, the average length of stay is approximately 9.5 hours. As previously mentioned, Figure 1 presents a snapshot of real-time patient flow through the entire ED process; the following case study will focus on part of that network, including the following three stations: nurse admission, doctor admission and lab tests. The first assumption we use is the model structure—based on interviews with nurses and physicians we assumed that these three stations can be modeled by an FJ network. During the nurse admission stage, for most patients, the nurse will also draw blood to be tested. In some cases, the test results will arrive before the doctor admission phase, and in others afterwards. Therefore, the blood tests and the doctor admission are done in parallel, and can be modeled as an FJ network (see Figure 5 for illustration). This description was also verified via the data we collected. However, we also note that unlike the FJ network assumptions, sometimes the doctor will wait for the lab tests to arrive *before* examining a patient, even though he is available. In those cases, the right model shall be two queues in tandem. Therefore, the assumption that the "true" model has an FJ structure is only approximately accurate in this particular ED application.

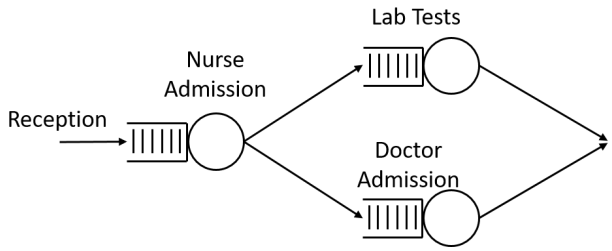


Figure 5 Modeling the beginning of an ED process as a single server queue followed by a fork-join queue.

Other model assumptions are also violated in reality. For example, service duration is not Exponential, but rather fits quite well a Log-normal distribution, as can be seen in Figure 6. This figure presents the sojourn time distribution at the nurse admission station, given that there were

no patients in the station upon a patient arrival. In that case, the patient sojourn time equals service time. To overcome this gap between assumptions and reality, we expand our results to non-exponential service times in the first station in §5.1, where we also discuss the influence of this assumption on all parts of the network via simulation.

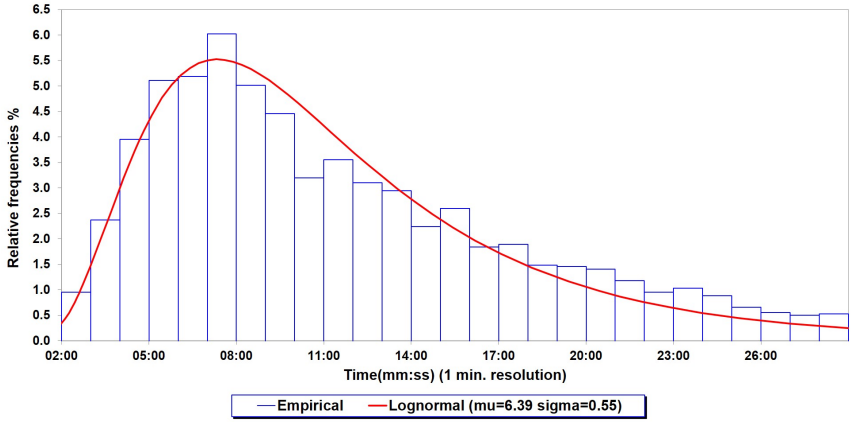


Figure 6 **Fitting Log-normal distribution to nurse admission service time distribution (April, 2014 - August, 2016, Weekdays)**

One of our model assumptions is that each station operates according to a FCFS priority policy. However, in healthcare systems, particularly in EDs, patients are assigned different priorities according to their medical condition. Our data analysis revealed that in each of the three stations considered, less than 40% of the patients maintained their place in queue between arrival to queue and departure from service. This means that the FCFS policy does not hold. Note that the patients order may be violated while waiting in queue or during service, if there is more than one service provider.

Another concern is that the number of people in queue is uncertain. This stems from several reasons. First, the lab provides service not just for ED patients but also to the entire hospital. The number of open orders to the lab is unknown. Second, physicians are providing treatment both to newly arrived patients (who are waiting for doctor admission), as well as patients at later stages of their ED process, i.e., both activities are done by the same resource. Hence, by isolating this sub-network, we underestimate the number of patients in queue, and ignore effects of other parts (both within the ED and outside the ED) on it. Still, assuming the lab commits the same amount of resources to ED patients over time, and since physicians give priority to patients waiting for admission over patients in later stages, we still can provide accurate enough results.

Current wait time estimation models assume either constant service rates or time-varying service rates. Yet, it was shown that service rates may also be a function of the load in the system (Berry Jaeker and Tucker 2016, Kc and Terwiesch 2009), which is expressed via the length of the queues; also known is that such dependency between load and service rates can impact the network equilibrium and stability (Chan et al. 2014, Jing Dong and Yom-Tov 2015). This is indeed the case in our data. Incorporating such phenomena into our model is actually straightforward, simply by changing service rates to be state-dependent. We demonstrate in this case study that such an adjustment is essential to obtain accurate delay predictions.

4.1. Service time estimation

In order to check the influence of queue length on service rates, we first need to evaluate service durations. Our data include completion times for each patient's activity. We do not have data of activity starting times; hence, an approximation was needed in order to determine the duration of each activity. We explain our service duration estimation methodology for nurse admission. Note that all patients' first activity in the ED is administrative reception and the second is nurse admission. Let Patient x be our target customer, and let Patient y be the last patient to go through nurse admission prior to Patient x. We distinguish between two cases; see Figure 7 for illustration. Case 1 in which Patient x reception time was before Patient y nurse admission time. In that case when Patient x arrived to the station the nurse was busy, hence, Patient x service duration (in the nurse admission stage) will be 'Patient x nurse admission time' minus 'Patient y nurse admission time'. Otherwise Case 2 applies, in which the nurse is assumed to be idle. In that case Patient x service duration (in the nurse admission stage) will be 'Patient x nurse admission time' minus 'Patient x reception time'.

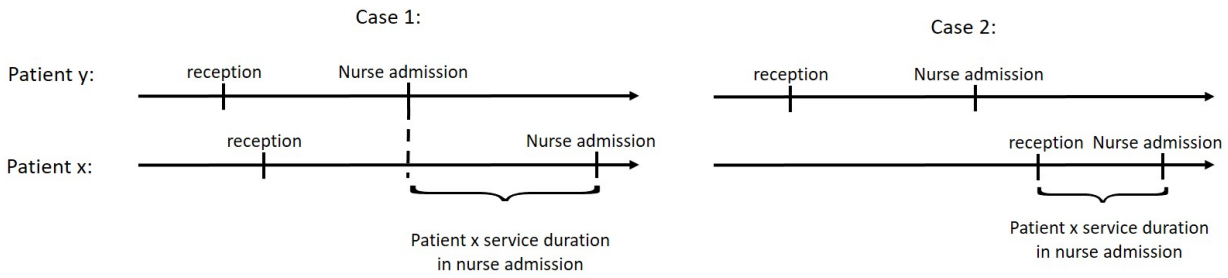


Figure 7 **Estimating service duration, an example.**

Figure 8 presents the service rate in the three stations we examined as a function of the queue length. It is clear that in all three stations, there is a speedup phenomenon; as queue length

increases, the service rate increases. This may be due to service providers accelerating their work in order to mitigate the load in the system, or simply due to extra staffing in loaded times. In any case, one must acknowledge this phenomenon when using real-time estimation methods. In the next section we explore the importance of using such service rates, rather than simply considering the average service rates over all cases.

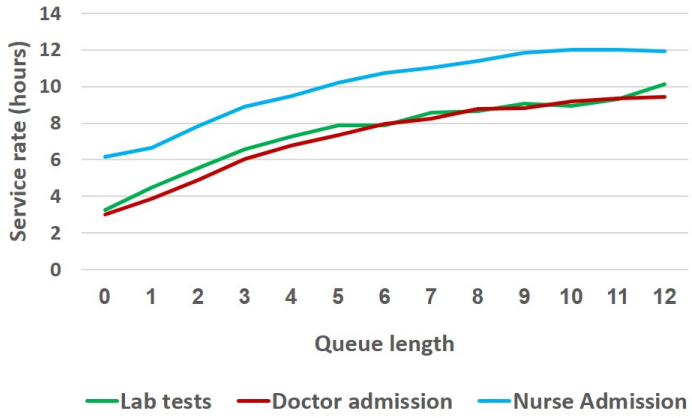


Figure 8 Service rates as a function of queue lengths

4.2. Case study results

In this section, we compare our model results with actual patient sojourn times in the ED data. A patient sojourn time at the first station (S_0) is considered as the time between administrative reception and nurse admission. A patient sojourn time at the doctor admission phase starts at nurse admission and ends at the doctor admission. For most of the blood test (70%) the hospital tracks the time of blood test order, which is considered as the entering time to that station. Hence, the sojourn time at the blood test phase is defined as the time between the blood test order and the arrival of the last blood test result. In cases where there was no record of a blood test order (about 30% of the cases), we assumed that a blood test was ordered at the time of nurse admission. This is based on our interviews and process observation. Consequently, the sojourn time at the FJ part (S_{FJ}) is defined as the maximum of the doctor admission sojourn time and the blood test sojourn time. The total sojourn time was then calculated as the summation of the sojourn time at the first station and the sojourn time at the FJ part. See Figure 9 for illustration.

As explained, the ED is divided into two sections: walking vs. urgent patient areas. We focus our analysis on walking patients; these patients may benefit the most from delay information, as

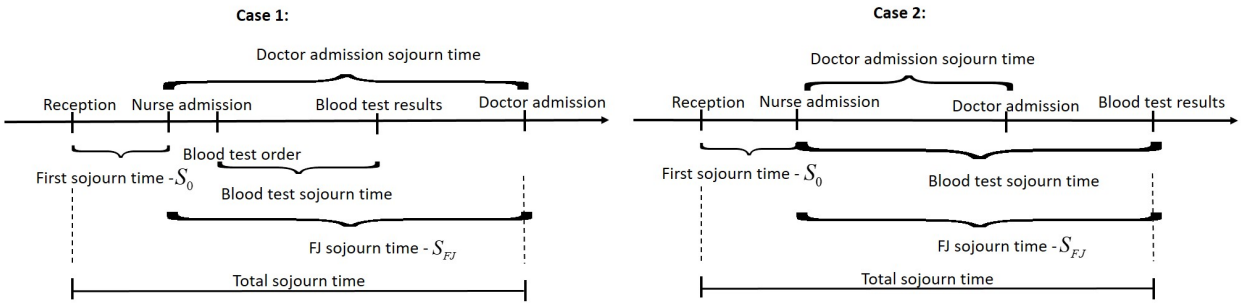


Figure 9 Sojourn times at the different stations - a scheme.

they can form decisions based on this information (for example, decide to wait at the cafeteria if they are expecting a long waiting time). Moreover, data records of acute patients are less reliable, as some patients are treated first, and records may be updated at a later stage; this is mainly due to their urgent medical condition at the beginning of their ED stay.

Recall that the inputs for our model are the three queue sizes observed at the patient arrival time (denoted by k, h_1, and h_2), and the service rates at the three stations (denoted by λ, μ_1, and μ_2). From now on, we shall refer to each combination of k, h_1, and h_2 as a "scenario". We have examined a total of 105 scenarios with sample sizes larger than 50 observations.We removed 5% of the observations in which the nurse admission sojourn time was equal to 0, and a negligible number of observations (less than 1%) in which one of the sojourn times was larger than 5 hours, as we assume this resulted from data corruption and does not reflect reality. Thus, including in our examination a total of 10,000 observations.

As a benchmark for our method we use the Last-to-Enter-Service (LES) estimator. This estimator is easy to implement and fairly robust, hence, widely used (Gal et al. 2017, Huang et al. 2015). The LES method is based on the snapshot principle, which was proven to hold also for FJ networks (Nguyen 1993). Consider a target patient entering the administrative reception. The LES estimator states that this target patient sojourn time in any of the three following stations (nurse admission, doctor admission and lab tests) can be estimated by the sojourn time of the last patient to enter service in that station, prior to the target customer reception time. To estimate the sojourn time of every patient in the FJ part, using an LES estimator, we first compute that patient's LES estimators for the doctor admission and the lab test stations separately, and then take the maximal value of the two.

We examined our model in two versions (a) using constant service rates and (b) using state-dependent service rates. We shall denote the model with the constant service rates by MFJC

Table 1 Comparing performance of LES, MFJC, and MFJV, in estimating sojourn times in the first station, the FJ station and the total sojourn time.

	First station	FJ station	Total
Average (min)	20.45	52.61	73.06
Standard Deviation (min)	14.15	31.47	34.49
LES - RMSE (min)	18.04	45.82	49.06
MFJC - RMSE (min)	14.00	34.39	37.93
MFJV - RMSE (min)	13.49	31.98	34.03

(Markovian, FJ, Constant rates), and the latter by MFJV (Markovian, FJ, Varied rates). The input service rates for the MFJC model were the average service rates among all observations. The input state-dependent service rates, *per each scenario* in the MFJV model, were the average service rates among all observations with the same queue length combinations.

A summary of results is provided in Table 1. The columns correspond to the first station (nurse admission), the FJ part (doctor admission and lab tests), and the total of the two. The top two rows present statistics of the average sojourn time and its standard deviation. We then present the Root-Mean-Squared-Error (RMSE) of the LES estimator and the RMSE of our two models (MFJC and MFJV). Note that our models provide an estimation for the entire distribution, while the LES only provides a single point estimation; hence, for 'fair' comparison, we present here RMSE with respect to expected sojourn times.

Clearly both the MFJC model and the MFJV model outperform the LES estimator in all three cases, reducing RMSE by 25–30%. This gap is probably due to the fact that the ED is in transient-state most of the day, which violates the snapshot principle assumptions upon which the LES estimator is based. Furthermore, the RMSEs of both models are very close to (and even lower than) the actual standard deviation of our data, which is also affected by the natural variability in the personal and medical condition of different patients. As expected, the MFJV model provides more accurate results than the MFJC model, as its input service rates fit better the real-time service rates. To be exact, its RMSE is 10% lower than the RMSE of the MFJC model, for the total sojourn time.

One of the unique features of our proposed method is the fact that it can provide estimations not just for the current station, but also for subsequent stations. Hence, it is more appropriate for networks than LES methods. Still, note that significant differences were found even in the first station.

Next we examine our method's performance in estimating sojourn time *distributions* for several scenarios. Here we consider both walking and acute patients, as it provides a much larger sample size. Figure 10 presents the sojourn time distribution in the FJ part according to the MFJV method (in red) compared with a Kernel smoothing of the empirical sojourn time distribution (in blue). The 12 graphs differ in their initial scenario (k, h_1, h_2), where $k \in \{0,1,2\}$, $h_1 \in \{0,1,2,3\}$ and $h_1 = h_2$. We chose these scenarios as they had enough observations required in order to compare distributions (as opposed to comparing averages which requires less data).

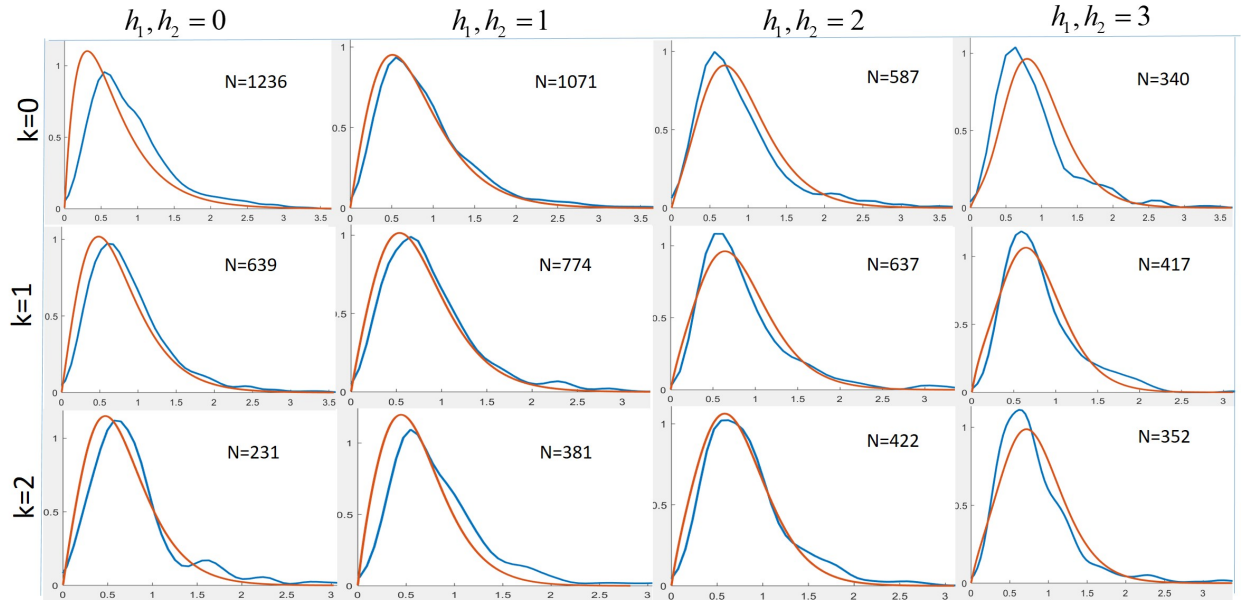


Figure 10 Comparing MFJV (red) distribution estimation with data (blue) of the FJ station sojourn times

We observe that in general there is an excellent fit of the theoretical distribution to the empirical one. The scenario where all queues are empty seems to have the worst fit—our method clearly underestimated sojourn times in that scenario. It seem that in the scenarios with higher queues in the FJ part (e.g. $h_1 = h_2 = 3$) our method slightly overestimated the sojourn time. There could be several reasons for such deviations, as stated before, many of the basic assumptions do not hold. In addition, it could be that some bias resulted from our method of estimating service rates, or from approximating a multi-server multi-queue network using a single-server multi-station network. Either way, considering all limitations above, we find the results more than satisfactory for practical use.

5. Model extensions

In this section we provide some important extensions and robustness checks of our method; we either relax or check the limitation resulting from some of the model's assumptions.

5.1. General service durations

One of our main assumptions was that service durations, in all three stations, are Exponentially distributed. In EC A, we partially relax this assumption by allowing a general service duration distribution in the first station. The case of general service durations in the FJ part is not a trivial extension of our model. However, we performed simulation experiments to validate that the assumption of Exponential service durations does not reduce accuracy significantly. To this end, we have compared estimated delays (mean and variance of the sojourn time at the first station, the FJ part and the total sojourn time) with simulation in which the service time distribution in all stations is Log-normal, as in our data (for example see Figure 6). Simulation results are presented in Figure 11 for 2x6 scenarios. The parameters of the top two graphs are $\lambda = 1$, $\mu_1 = 1$, $\mu_2 = 1$, $h_1 = 2$, $h_2 = 2$, while the queue in the first station, k, varies from 0 to 5. The parameters of the bottom two graphs are $\lambda = 1$, $\mu_1 = 1$, $\mu_2 = 1$, $k = 2$, $h_2 = 2$, while the queue in the UFJ station, h_1, varies from 0 to 5. We ran $50,000$ iterations per each scenario. The graphs on the left (right) present the mean sojourn time (sojourn time variance) at the first station in red, at the FJ part in blue, and in yellow the total sojourn time. Solid lines represent the theoretical model results, having Exponentially distributed service times, while the dotted lines represent the simulation results having Log-normal service times. The simulation reveals that the estimated sojourn time expectation is accurate even when the actual service durations are not Exponential but rather Log-normal; additionally there were only small deviations in the sojourn time variance.

5.2. Fork-Join network with a delay node

In the next extension, we relaxed some of the *structural* assumptions of our model by replacing one of the service stations with a general delay node.

The FJ part of our network consists of two stations. It is possible to expand the network to account for more stations; it is quite straightforward. However, it may become computationally cumbersome. Alternatively, we propose to use a delay node within the FJ part to capture the joint delay resulting from multiple stations. This approach was inspired by Yom-Tov and Mandelbaum (2014) and van Leeuwaarden et al. (2017) who used it for making staffing decisions in the complex ED environment. Those papers neglected the FJ structure inherited in ED services. Nevertheless,

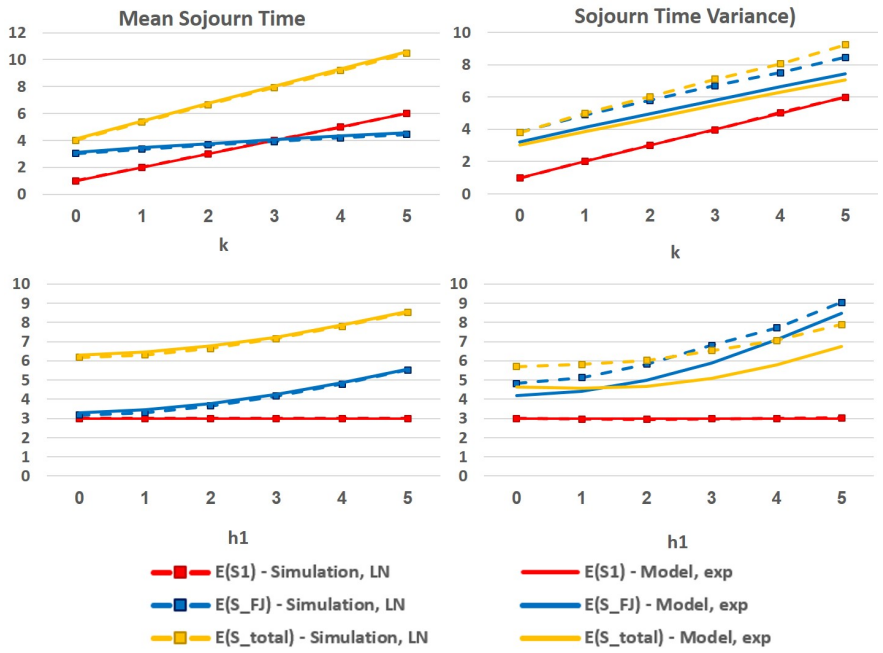


Figure 11 **Comparison of simulation results with Log-normal service times vs. model results with Exponential service times.**

the delay-station approach can bridge the dynamic structure so common in ED services. Patients in EDs usually undergo several activities which are necessary for all patients while other activities are only relevant for a portion of them. For example, all ED patients need to be seen by a doctor. However, some of them will need to go through an X-Ray exam, CT scan, or Ultrasound exam, and possibly more than one exam. Furthermore, it is not known in advance which of the exams, if any, the patients will need, resulting in dynamic processes. Another uncertainty arises from the fact that some of the resources provide service to patients which are not included in the specified system. For example, the lab provides service to the entire hospital, not just for ED patients. We propose to tackle such uncertainties by combining all those activities into a stochastic delay node, positioned in parallel to a focal known activity (e.g. blood tests). A delay node is an infinite server station with a generally distributed service time. The resulting model will then have one single server queue (with either Exponential or general service time distribution), followed by an FJ part consisting of one single (or multiple) server queue with Exponentially distributed service durations and one infinite server queue with some general service durations.

This model can be reduced to a very similar model to the one presented in Boxma and Daduna (2014) of a tandem network consisting of one queue with general service times followed by a

Markovian queue. With the delay node, in each of the recursion iterations, the joint LST is conditioned only on the number of potential service completions in one station with Exponential service durations (and not on two such stations), in accordance with the recursion in Boxma and Daduna (2014). However, the time each customer spends in the last station (FJ with one delay node and one Markovian node), given that when he enters the station there are x customers in the Markovian node, does not follow an Erlang distribution. Instead, it is the maximum of two independent random variables, one with general distribution (corresponding to the sojourn time in the delay node) and one with Erlang distribution. Formulas for this model are presented in EC B.

6. Discussion

Our research goal was to develop *real-time, state-dependent* estimations for delay *distributions* in service *networks*. The motivation came from healthcare systems which are proved to be complex queueing networks. We have developed a recursive formula to estimate sojourn times in a network with one single server Markovian queue followed by an FJ part with two stations, each with Exponentially distributed service duration. We derived an estimation for the sojourn time in the first part, the sojourn time in the FJ part, and the total sojourn time—all are given at the time of a customer's arrival to the system, and are based on the state of the system at that arrival epoch.

We examined our methodology in a case study of an Israeli ED, based on unique transaction-level data. This use case revealed that although essentially none of our model assumptions holds, our methodology provides fairly accurate estimations for sojourn time distributions, and outperforms other commonly used estimators, such as the LES.

Then we relaxed some of the distributional assumptions, providing a natural extension for our model with general service times in the first station. Furthermore, we have shown via simulation that even when services in all stations are not Exponential, our methodology still provides accurate results.

The main limitation of our methodology is the "curse of dimensionality". For example, if we wish to add explicitly more stations to the FJ part, computations may become cumbersome. We approach this problem by suggesting another extension to our model in which one of the FJ stations becomes a delay node. This delay node can represent, in an aggregative manner, all other stations in the system. Computational problems may also occur when queue size grows large. In this case, approximation methods may be more appropriate. We leave this for future research. One of the basic assumptions of our model is the FCFS policy. However, as is evident through data, this may

not hold, especially in healthcare systems in which priority is determined by medical condition. This may be one of the reasons for deviations between the sojourn time estimation provided by the theoretical model and patients' actual sojourn time. Incorporating priorities into such models will also be an interesting future research direction. Another direction of future research could be in combining the estimators developed here with machine-learning techniques, i.e., adding them as features in statistical models. This was shown to be very promising in Senderovich et al. (2014) and Ang et al. (2016), and may help in increasing accuracy for FJ networks too.

References

Abate J, Whitt W (1995) Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal on Computing* 7(1):36–43.

Altman D, Efrat-Treister D, Eisenman A, Lev-Arey D, Rafaeli A, Rosenfeld Y, Shapira C, Solomon S (2016) Aggression toward ED medical staff: Can information help?, International Association of Conflict Management. New-York, NY.

Ang E, Kwasnick S, Bayati M, Plambeck EL, Aratow M (2016) Accurate emergency department wait time prediction. *Manufacturing & Service Operations Management* 18(1):141–156.

Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB (2015) On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* 5(1):146–194.

Armony M, Shimkin N, Whitt W (2009) The impact of delay announcements in many-server queues with abandonment. *Operations Research* 57(1):66–81.

Berry Jaeker JA, Tucker AL (2016) Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science* 63(4):1042–1062.

Boxma O (1983) The cyclic queue with one general and one exponential server. *Advances in Applied Probability* 15(4):857–873.

Boxma O, Daduna H (2014) The cyclic queue and the tandem queue. *Queueing Systems* 77(3):275–295.

Boxma OJ, Daduna H (1990) Sojourn times in queueing networks. Takagi H, ed., *Stochastic Analysis of Computer and Communication Systems*, 401–450 (North-Holland Publishing Company, Amsterdam).

Boxma OJ, Koole G, Liu Z (1994) Queueing-theoretic solution methods for models of parallel and distributed systems. *In: Performance Evaluation of Parallel and Distributed Systems—Solution Methods*, Proceedings of the third QMIPS workshop, Part 1 (CWI Tract 105, Amsterdam).

Chan CW, Yom-Tov GB, Escobar G (2014) When to use speedup: An examination of service systems with returns. *Operations Research* 62(2):462–482.

Daduna H (1986) Two-stage cyclic queues with nonexponential servers: steady-state and cyclic time. *Operations Research* 34(3):455–459.

Dong J, Yom-Tov E, Yom-Tov GB (2018) The impact of delay announcements on hospital network coordination and waiting times. *Management Science* .

Efrat M, Parush A (2017) Personalized information to patients in the emergency department: User centered design and testing, Production and Operations Management Society (POMS) Conference. Tel-Aviv, Israel.

Gal A, Mandelbaum A, Schnitzler F, Senderovich A, Weidlich M (2017) Traveling time prediction in scheduled transportation with journey segments. *Information Systems* 64:266–280.

Huang J, Carmeli B, Mandelbaum A (2015) Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research* 63(4):892–908.

Ibrahim R, Whitt W (2009a) Real-time delay estimation based on delay history. *Manufacturing & Service Operations Management* 11(3):397–415.

Ibrahim R, Whitt W (2009b) Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Science* 55(10):1729–1742.

Ibrahim R, Whitt W (2011a) Real-time delay estimation based on delay history in many-server service systems with time-varying arrivals. *Production and Operations Management* 20(5):654–667.

Ibrahim R, Whitt W (2011b) Wait-time predictors for customer service systems with time-varying demand and capacity. *Operations Research* 59(5):1106–1118.

Jing Dong PF, Yom-Tov GB (2015) Service system with slowdowns: Potential failures and proposed solutions. *Operations Research* 63(2):305–324.

Jouini O, Akşin OZ, Karaesmen F, Aguir MS, Dallery Y (2015) Call center delay announcement using a newsvendor-like performance criterion. *Production and Operations Management* 24(4):587–604.

Jouini O, Aksin Z, Dallery Y (2011) Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management* 13(4):534–548.

Kc DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* 55(9):1486–1498.

Kerner Y (2008) The conditional distribution of the residual service time in the Mn/G/1 queue. *Stochastic Models* 24(3):364–375.

Ko SS, Serfozo RF (2004) Response times in M/M/s fork-join networks. *Advances in Applied Probability* 36(3):854–871.

Ko SS, Serfozo RF (2008) Sojourn times in G/M/1 fork-join networks. *Naval Research Logistics (NRL)* 55(5):432–443.

Maister DH (1984) The psychology of waiting lines. Czepiel JA, Solomon MR, Surprenant CF, eds., *The Service Encounter*, 113–123 (Lexington Books, Lexington, MA).

Mandelbaum A, Yechiali U (1983) Optimal entering rules for a customer with wait option at an M/G/1 queue. *Management Science* 29(2):174–187.

Moriah H, Efrat-Treister D, Rafaeli A, Cheshin A, Agasi S (2011) Situational antecedents of customer conflict and aggression toward healthcare professionals in the hospital setting. *IACM 24TH Annual Conference Paper*.

Munichor N, Rafaeli A (2007) Numbers or apologies? customer reactions to telephone waiting time fillers. *Journal of Applied Psychology* 92(2):511–518.

Nakibly E (2002) *Predicting waiting times in telephone service systems*. Master's thesis, Technion—Israel Institute of Technology.

Nguyen V (1993) Processing networks with parallel and sequential tasks: Heavy traffic analysis and Brownian limits. *The Annals of Applied Probability* 3(1):28–55.

Qiu Z, Pérez JF, Harrison PG (2015) Beyond the mean in fork-join queues: Efficient approximation for response-time tails. *Performance Evaluation* 91:99–116.

Rafaeli A, Lev-Arey D, Efrat-Treister D, Moriah H, Rosenfeld Y (2014) Curtailing anger in emergency departments: Providing information as a way of reducing aggression, Academy of Management Annual Meeting, Philadelphia, Pennsylvania.

Reiman MI (1982) The heavy traffic diffusion approximation for sojourn times in Jackson networks. *Applied Probability Computer Science: The Interface*, 409–421 (Springer).

Rizk A, Poloczek F, Ciucu F (2015) Computable bounds in fork-join queueing systems. *ACM SIGMETRICS Performance Evaluation Review*, 43(1):335–346 (ACM).

Senderovich A, Weidlich M, Gal A, Mandelbaum A (2014) Queue mining–predicting delays in service processes. *International Conference on Advanced Information Systems Engineering*, 42–57 (Springer).

Stadje W (1996) Non-stationary waiting times in a closed exponential tandem queue. *Queueing Systems* 22(1-2):65–77.

Sullivan DR, Liu X, Corwin DS, Verceles AC, McCurdy MT, Pate DA, Davis JM, Netzer G (2012) Learned helplessness among families and surrogate decision-makers of patients admitted to medical, surgical, and trauma ICUs. *Chest Journal* 142(6):1440–1446.

Sun Y, Teow KL, Heng BH, Ooi CK, Tay SY (2012) Real-time prediction of waiting time in the emergency department, using quantile regression. *Annals of Emergency Medicine* 60(3):299–308.

Taylor SE (1979) Hospital patient behavior: Reactance, helplessness, or control? *Journal of Social Issues* 35(1):156–184.

Thiongane M, Chan W, L'Ecuyer P (2016) New history-based delay predictors for service systems. *Winter Simulation Conference*.

van Leeuwaarden JS, Mathijsen BW, Sloothaak F, Yom-Tov GB (2017) The restricted Erlang-R queue: Finite-size effects in service systems with returning customers, working paper.

Witkovskỳ V (2016) Numerical inversion of a characteristic function: An alternative tool to form the probability distribution of output quantity in linear measurement models. *Acta IMEKO* 5(3):32–44.

Yom-Tov GB, Mandelbaum A (2014) Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management* 16(2):283–299.

Yom-Tov GB, Rafaeli A, Westphal M (2017) An empirical study of customer patience and abandonment in online customer service, working paper, Technion.

Yu Q, Allon G, Bassamboo A (2016) How do delay announcements shape customer behavior? an empirical study. *Management Science* 63(1):1–20.

Technical notes and model extensions

Appendix A: General service times at the first station

In this appendix we provide a formulation for estimating sojourn times in the FJ network when service time in the first station has a general distribution.

Assume that a target customer arrives to the first station and finds k customers there. Let R_k be the remaining service duration of the customer being served in the first station, at the time of that target customer's arrival (i.e., given that there were k customers in the first queue). Expressions for the expected value of R_k and for its LST are given in Mandelbaum and Yechiali (1983), and Kerner (2008).

Let $B(\cdot)$ be the cumulative distribution function of service duration in the first station (i.e, $S_0|k=0 \sim B$). As before, we define $\Psi_{k,h_1,h_2}(w_0, w_{FJ})$ as the joint Laplace-Stieltjes transform of S_0 and S_{FJ} given k customers in the first station, h_1 customers in the UFJ station and h_2 customers in the LFJ station.

$$\Psi_{k,h_1,h_2}(w_0,w_{FJ}) \equiv \mathbb{E}_{k,h_1,h_2}\left(e^{-w_0 S_0 - w_{FJ} S_{FJ}}\right) = \int_0^\infty e^{-w_0 t} X_{k,h_1,h_2}\, \mathrm{d}\mathbb{P}\left(R_k < t\right)$$

$$= \int_0^\infty e^{-w_0 t} \left[\sum_{r_1=0}^{h_1-1} \sum_{r_2=0}^{h_2-1} e^{-\mu_1 t} \frac{(\mu_1 t)^{r_1}}{r_1!} e^{-\mu_2 t} \frac{(\mu_2 t)^{r_2}}{r_2!} \Psi_{k-1,h_1-r_1+1,h_2-r_2+1}(w_0,w_{FJ}) \right.$$

$$+ \sum_{r_1=0}^{h_1-1} \sum_{r_2=h_2}^{\infty} e^{-\mu_1 t} \frac{(\mu_1 t)^{r_1}}{r_1!} e^{-\mu_2 t} \frac{(\mu_2 t)^{r_2}}{r_2!} \Psi_{k-1,h_1-r_1+1,1}(w_0,w_{FJ})$$

$$+ \sum_{r_1=h_1}^{\infty} \sum_{r_2=0}^{h_2-1} e^{-\mu_1 t} \frac{(\mu_1 t)^{r_1}}{r_1!} e^{-\mu_2 t} \frac{(\mu_2 t)^{r_2}}{r_2!} \Psi_{k-1,1,h_2-r_2+1}(w_0,w_{FJ})$$

$$\left. + \sum_{r_1=h_1}^{\infty} \sum_{r_2=h_2}^{\infty} e^{-\mu_1 t} \frac{(\mu_1 t)^{r_1}}{r_1!} e^{-\mu_2 t} \frac{(\mu_2 t)^{r_2}}{r_2!} \Psi_{k-1,1,1}(w_0,w_{FJ}) \right] \mathrm{d}\mathbb{P}\left(R_k < t\right).$$

Here, $\Psi_{l,m,n}\left(w_0, w_{FJ}\right)$ is the LST of the joint distribution of the target customer's sojourn time in the first station and in the FJ part, given that *a new service starts right now in the first station*, and that at this epoch the target customer sees l customers in front of him in the first station, m customers in the UFJ station, and n customers in the LFJ station. We now denote:

$$a(r_1,r_2,w_0) = \int_0^\infty e^{-(w_0+\mu_1+\mu_2)t} \frac{(\mu_1 t)^{r_1}}{r_1!} \frac{(\mu_2 t)^{r_2}}{r_2!}\, \mathrm{d}B(t),$$

$$b(r_1,h_2,w_0) = \int_0^\infty e^{-(w_0+\mu_1+\mu_2)t} \sum_{r_2=h_2}^{\infty} \frac{(\mu_1 t)^{r_1}}{r_1!} \frac{(\mu_2 t)^{r_2}}{r_2!}\, \mathrm{d}B(t),$$

$$c(h_1,r_2,w_0) = \sum_{r_1=h_1}^{\infty} a(r_1,r_2,w_0),$$

$$d(h_1,h_2,w_0) = \sum_{r_1=h_1}^{\infty} b(r_1,h_2,w_0).$$

Let X be a Poisson process with rate μ_1, and let Y be a Poisson process with rate μ_2. The above expressions are reduced to:

$$a(r_1,r_2,w_0) = \int_0^\infty e^{-w_0 t} \mathbb{P}\left(X(t)=r_1,\, Y(t)=r_2\right) \mathrm{d}B(t),$$

$$b(r_1,h_2,w_0)=\int_0^\infty e^{-w_0t}\mathbb{P}\left(X(t)=r_1,\,Y(t)\geq h_2\right)\mathrm{d}B(t),$$

$$c(h_1,r_2,w_0)=\int_0^\infty e^{-w_0t}\mathbb{P}\left(X(t)\geq h_1,\,Y(t)=r_2\right)\mathrm{d}B(t),$$

$$d(h_1,h_2,w_0)=\int_0^\infty e^{-w_0t}\mathbb{P}\left(X(t)\geq h_1,\,Y(t)\geq h_2\right)\mathrm{d}B(t).$$

With these new notations, the recursive formula for calculating $\Psi_{k,h_1,h_2}(w_0,w_{FJ})$ presented in Eq. (3) will not change, nor does the rest of the analysis. There will be only minor differences in the formulas of the stopping conditions. Let $\beta(\cdot)$ be the LST of the specified service duration (which can now be general). Thus, the stopping conditions will now be: $\Psi_{0,0,0}(w_0,w_{FJ})=\beta(w_0)Z_{1,1}(w_{FJ})$. In the same manner, we get

$$\Psi_{k,0,0}(w_0,w_{FJ})=\Psi_{k-1,1,1}(w_0,w_{FJ})\beta(w_0).$$

We shall also define (using a shorthand notation in which we omit w_0):

$$a^*(k,r_1,r_2)=\int_0^\infty e^{-w_0t}\mathbb{P}\left(X(t)=r_1,\,Y(t)=r_2\right)\mathrm{d}\mathbb{P}\left(R_k<t\right),$$

$$b^*(k,r_1,h_2)=\int_0^\infty e^{-w_0t}\mathbb{P}\left(X(t)=r_1,\,Y(t)\geq h_2\right)\mathrm{d}\mathbb{P}\left(R_k<t\right),$$

$$c^*(k,h_1,r_2)=\int_0^\infty e^{-w_0t}\mathbb{P}\left(X(t)\geq h_1,\,Y(t)=r_2\right)\mathrm{d}\mathbb{P}\left(R_k<t\right),$$

$$d^*(k,h_1,h_2)=\int_0^\infty e^{-w_0t}\mathbb{P}\left(X(t)\geq h_1,\,Y(t)\geq h_2\right)\mathrm{d}\mathbb{P}\left(R_k<t\right).$$

Hence, define

$$\Phi^*_{k,m,n}=\begin{pmatrix} a^*(k,0,0) & a^*(k,1,0) & \dots & a^*(k,m-1,0) & c^*(k,m,0)\\ a^*(k,0,1) & a^*(k,1,1) & \dots & a^*(k,m-1,1) & c^*(k,m,1)\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ a^*(k,0,n-1) & a^*(k,1,n-1) & \dots & a^*(k,m-1,n-1) & c^*(k,m,n-1)\\ b^*(k,0,n) & b^*(k,1,n) & \dots & b^*(k,m-1,n) & d^*(k,m,n)\end{pmatrix}.$$

As before, for $1\leq l\leq k-1$:

$$\Psi_{l,m,n}=trace\left(\bar{\Psi}_{l-1,m+1,n+1}\cdot\Phi_{m,n}\right),$$

and

$$\Psi_{k,h_1,h_2}(w_0,w_{FJ})\equiv\mathbb{E}_{k,h_1,h_2}\left(e^{-w_0S_0-w_{FJ}S_{FJ}}\right)=trace\left(\bar{\Psi}_{k-1,h_1+1,h_2+1}\cdot\Phi^*_{k,h_1,h_2}\right).$$

Appendix B: Fork-join with delay node

We consider a model with one single server queue (with either exponential or general service time distribution), followed by an FJ part, which consists of one single (or multiple) server queue with exponentially distributed service durations and one infinite server queue with some general service duration. We denote the CDF of this distribution by $G(\cdot)$. Assuming that the delay node (that is, the infinite server queue), is the UFJ station we get, as in the model of

Boxma and Daduna (2014), that the joint LST of the sojourn times in the first station and in the FJ part is given by the following recursive formula:

$$\Psi_{k,h_2}(w_0,w_{FJ}) = \int_0^\infty e^{-w_0 t}\left\{\sum_{l=0}^{h_2-1} e^{-\mu_2 t}\frac{(\mu_2 t)^l}{l!}\Psi_{k-1,h_2-l+1}(w_0,w_{FJ}) + \sum_{l=h_2}^{\infty} e^{-\mu_2 t}\frac{(\mu_2 t)^l}{l!}\Psi_{k-1,1}(w_0,w_{FJ})\right\}\mathrm{d}B(t),$$

However,

$$\Psi_{0,h_2}(w_0,w_{FJ}) = \int_0^\infty e^{-w_0 t}\left\{\sum_{l=0}^{h_2-1} e^{-\mu_2 t}\frac{(\mu_2 t)^l}{l!}Z_{h_2-l+1}(w_{FJ}) + \sum_{l=h_2}^{\infty} e^{-\mu_2 t}\frac{(\mu_2 t)^l}{l!}Z_1(w_{FJ})\right\}\mathrm{d}B(t),$$

and

$$\Psi_{0,0}(w_0,w_{FJ}) = \beta(w_0)Z_1(w_{FJ}).$$

Here $\beta(\cdot)$ is the LST of the general service times in the first station, and $Z_x(w_{FJ})$ is the LST of the maximum of two independent random variables, one with general distribution (corresponding to the sojourn time in the delay node) and one with Erlang distribution (corresponding to the sojourn time of a customer entering a Markovian multiple/single server queue with x customers already there). Hence

$$Z_x(w_{FJ}) \equiv \mathbb{E}_x\left(e^{-w_{FJ}S_{FJ}}\right) = \int_0^\infty e^{-w_{FJ}t}\,\mathrm{d}F_{S_{FJ}|x}(t),$$

where

$$\begin{aligned} F_{S_{FJ}|x}(t) &= \mathbb{P}\left(S_{FJ}\le t|h_2=x\right) = \mathbb{P}\left(S_1\le t,\, S_2\le t|h_2=x\right)\\ &= \mathbb{P}\left(S_2\le t|h_2=x\right)\mathbb{P}\left(S_1\le t\right) = \left(1-\sum_{m=0}^{x} e^{-\mu_2 t}\frac{(\mu_2 t)^m}{m!}\right)G(t). \end{aligned}$$