Queueing Systems (2021) 97:39–80
https://doi.org/10.1007/s11134-021-09688-w

A survey of parameter and state estimation in queues

Azam Asanjarani[1] · Yoni Nazarathy[2] · Peter Taylor[3]

Received: 1 November 2020 / Revised: 17 December 2020 / Accepted: 21 January 2021 /
Published online: 17 February 2021

Abstract
We present a broad literature survey of parameter and state estimation for queueing systems. Our approach is based on various inference activities, queueing models, observations schemes, and statistical methods. We categorize these into branches of research that we call estimation paradigms. These include: the classical sampling approach, inverse problems, inference for non-interacting systems, inference with discrete sampling, inference with queueing fundamentals, queue inference engine problems, Bayesian approaches, online prediction, implicit models, and control, design, and uncertainty quantification. For each of these estimation paradigms, we outline the principles and ideas, while surveying key references. We also present various simple numerical experiments. In addition to some key references mentioned here, a periodically updated comprehensive list of references dealing with parameter and state estimation of queues will be kept in an accompanying annotated bibliography.

Keywords Queueing inference · Queueing parameter estimation · Inverse problems · Queue inference engine · Queueing prediction

Mathematics Subject Classification 60K25 · 68M20 · 62M20

1 Introduction

Queues occur in retail, health-care, telecommunications, manufacturing, road traffic, social justice systems, call centres, and many other environments. To aid in understanding such systems, mathematical queueing models have been studied and employed for over a century. Such models allow researchers and practitioners to predict congestion

✉ Azam Asanjarani
azam.asanjarani@auckland.ac.nz

[1] The University of Auckland, Auckland, New Zealand

[2] The University of Queensland, Brisbane, Australia

[3] The University of Melbourne, Melbourne, Australia

and delay behaviour based on assumptions about the underlying stochastic processes. The field has grown together with the field of applied probability and constitutes a significant part of the world of stochastic operations research. Indeed, queueing phenomena are both fascinating and important to understand from a practical perspective.

The basic building block of most queueing analysis research involves a *queueing model*. As an example, consider a model called the M/D/1 queue. In such a model, customers arrive to a server according to a Poisson process. The server processes customers one at a time, taking a fixed deterministic time, m, for each customer and idles when there are none left. When a customer arrives to see the server busy, the customer queues. In this model, there are only two parameters of interest: λ the arrival rate (customers per time unit), and m the service time. If $\lambda m > 1$, then the average number of customers that arrive during a service is greater than one and the queue will build up over time without bound. However, if $\lambda m < 1$, then the system will settle down to a stochastic equilibrium.

Each customer experiences a waiting time (which may be 0 if arriving to an empty queue) and a sojourn time which is the customer's total time in the system (waiting time + service time). In the $\lambda m < 1$ regime, it makes sense to analyse equilibrium mean waiting times and other performance measures. There are multiple service disciplines that the server may use such as first come first served (FCFS), random order of service, or other disciplines. The service discipline generally affects the distribution of the waiting time, but does not affect the mean waiting time as long as the server does not idle when customers are present. Queueing theory thrives on results for models such as M/D/1. For example, by setting $\rho = \lambda m < 1$, the mean waiting time of an arbitrary customer for this model is

$$m\frac{\rho}{2(1-\rho)}. \tag{1}$$

Formulas such as this immediately lead to elementary system insights. First observe that as $\rho \to 1$ the mean waiting time grows without bound. For example, with $\rho = 2/3$ the mean time a customer waits in the system is equal to the time it takes the customer to be served. However, for higher values of ρ, the mean waiting time of a customer is longer than the service time.

Results such as this are the cornerstone of queueing theory and analysis. However, how can queueing theory be employed? An immediate answer is to use queueing analysis for arriving at general insights about real world systems. For example, an insight gained from the result presented above is that mean waiting times are of the order $\frac{\rho}{1-\rho}$, as $\rho \to 1$.

Such insights have helped with the design of computing systems, telecommunication systems, manufacturing systems, health-care operations, and more. With an abundance of queueing models, one may wish to go further than just providing insight. Indeed, there is the opportunity to use these models to predict, manage, and design explicitly. This requires using queueing models that are realistic in the context of actual systems and the associated data collection processes. As an example, say that we know that a telecommunications switch takes exactly 1 ms to handle a packet ($m = 0.001$ s). Now, under the Poisson assumption for packet arrivals, for $\lambda \in (0, 1000)$ we know

that the mean waiting time can be computed using (1). For example, if $\lambda = 700$ then with an M/D/1 model, the mean waiting time is about 1.166 ms. Such conclusions require data collection for estimation of m, λ, and verification of the suitability of the model.

Such an application of queueing results generates a variety of questions associated with statistical analysis. In an actual system, how would we measure λ? Or what about the Poisson arrival assumption, is it sensible and supported by data? Further, what if different aspects of the queueing system were observed, not necessarily giving us a full indication of all the underlying processes. How would we then fit a queueing model to the system?

It turns out that while there are thousands of papers dealing with queueing theory and analysis, there are far fewer papers dealing with these types of estimation problems. In fact, this state of affairs was identified as early as 1965 by David R. Cox in [41], where he stated,

> There are a very large number of papers on particular probabilistic models for queues and, by comparison, extremely few papers on the corresponding problems of statistical analysis. When a simple mathematical model is investigated primarily to get qualitative insight into the behavior of queueing systems, the statistical problems are not very relevant. When, however, there is the possibility of quantitative application, or when a practical congestion problem is tackled by rather empirical methods, non-trivial statistical problems arise.

Our purpose in this survey is to present results that are available and work that has already been done. Such problems were considered quite early with the seminal work of Clarke [37], the survey [22] by Bhat, Miller, and Rao, and an updated survey with the same title [23] by Bhat and Rao. Since these were published, there has been a significant body of additional work. We survey both the classical queueing estimation work and more recent results in the current paper.

Structure of this survey

This survey is structured as follows: We start in Sect. 2 by describing the general framework. This includes outlining in Sect. 2.1 a variety of problems which we call **inference activities**. We then present a brief illustration of queueing models in Sect. 2.2. This section is geared towards those that have not been exposed to queueing theory. We then go on to present what we define as **observation schemes** in Sect. 2.3. In Sect. 2.4, we illustrate some of the complexities involved with parameter estimation in queues. The survey continues in Sect. 3 where we lay out the various **estimation paradigms** and outline some of the key contributions in the literature. See Table 2 for an overview. We conclude in Sect. 4 where we outline a few broad areas that have received less attention in the literature. The computational examples that we present in Sect. 2.4 are also accessible via the GitHub repository [11]. A (periodically updated) annotated bibliography aiming to contain an exhaustive list of references in the area is in [9].

2 Framework and background

Research that deals with inference in a queueing model usually has a number of characteristics that describe the type of inference and the type of modelling that is involved. We can classify a paper in this area according to four general attributes:

(i) **The inference activity** that is performed, for instance, parameter estimation, state estimation, hypothesis testing, or sample size planning.
(ii) **The models** employed such as an M/M/1 model, an open queueing network model, or an M/G/∞ model.
(iii) **The observation scheme** used, such as whether a continuous record of data is available or just data observed at certain time points.
(iv) **The statistical methods and principles** used, for instance, likelihood based methods, moment matching, Bayesian inference, or nonparametric inference.

We discuss (i)–(iv) in the subsections below.

2.1 Inference activities

Performing inference on queues can have different objectives. Here are some common activities and their objectives:

1. *To find the parameters of a model* In this case, we believe or assume that a real queueing system for which we have data behaves according to a specific model. The task is to estimate parameters of such a model. It could, for example, be to estimate the arrival rate λ for the M/D/1 queue discussed in the introduction. The majority of the work that we survey in this paper deals with this type of activity.
2. *To select a suitable model based on data* The act of choosing a model for a scenario is often performed without reference to data. However, in certain cases, we may want to incorporate data into the model selection process. For example, we may want to test if interarrival times to a queueing system are independent and exponentially distributed to decide whether a Poisson arrival process assumption is suitable. There has not been much work on this type of activity (in the context of queues) to date. We survey the few papers that we found.
3. *To plan observation schemes and experiments* In classical statistical contexts, elementary considerations in design of experiments involve determining the number of samples to take, the various treatment classes, and stratifying and randomising subjects. In a queueing context, there is an additional complication due to the fact that a queueing system is a dynamic process evolving over time. Most studies are necessarily observational studies. Many such schemes involve indirect methods for observing the quantities of interest. Important considerations involve efficiency of information retrieval and understanding such things as sampling bias. Only a few of the papers that we survey deal with such an activity and we believe that there is room for further research on this area.
4. *To carry out state prediction or filtering* In (1) above, we discussed estimating parameters of models. However, in many practical situations, a parameterized model is already present and the question is about the state. Given past partial

observations, it is of interest to either estimate the full state in the past or predict future states. Such problems may often be tackled via black box methods such as neural networks and/or hidden Markov models. However, in our context, the methods are based on actual queueing models.

5. *Adaptive control* The process of estimating states and parameters and the process of controlling the system are often decoupled. However, in certain cases, one may make decisions online, while parameter and state estimation are ongoing. This is the case of adaptive control. In general, methods from reinforcement learning where the parameters are unknown and partially observable Markov decision processes where the state is unknown present a variety of techniques for dealing with such problems. However, a few selected papers deal with such problems utilising the queueing structure. A related concept is robust control which is not exhaustively covered in this survey.

2.2 Queueing processes and models

We now briefly illustrate key aspects of queueing theory. Our purpose is to present the reader with a taste of key phenomena, models, and quantities involved. Hence, this section is not about inference but is rather about the mathematical (stochastic) models of queues. Specialists in queueing theory may wish to skip this section as it contains elementary material.

As illustrated in the introduction, a model like M/D/1 may be used to predict expected waiting times, mean queue lengths, and other measures in a system subject to Poisson arrivals. We now generalize this model to a broader class called the GI/G/1 queue. Other special cases are the M/G/1 and the M/M/1 models. All of these models are single-server queueing models (hence the "1" in the name). The difference between them lies in the probabilistic assumptions imposed on the arrival process. In all these models, the interarrival times are assumed i.i.d. (independent and identically distributed). In the "GI" case, they follow an arbitrary distribution, while in the "M" case, they follow an exponential distribution and in the "D" case, take a deterministic value. Similar comments apply to the service time process. More general queueing models allow dependence between interarrival times of the arrival process, and rarely between service times, which is less natural from a modelling point of view.

Exogenous processes As model inputs, consider the following two basic *exogenous processes*: the arrival process and the service time process. The *arrival process* may be described by $A(t)$ where t is a continuous time variable. Here, $A(t)$ counts the number of arrivals during the time interval $[0, t]$. This process is exogenous because in the basic suite of models, it is not considered to be affected by the internal dynamics of the queueing system. An alternative representation is via a sequence $T_1, T_2, \ldots$ that marks the arrival times of customers.

Like the arrival process, the *service time process* is usually assumed not to be influenced by the internal dynamics of the queue; hence, it is an exogenous process. It is naturally described by the sequence, $S_1, S_2, \ldots$, where S_n is the service time of the nth customer arriving to the queue. However, we can also define

$$S(t) = \sup\left\{n \ : \ \sum_{i=1}^{n} S_i \leq t\right\}, \tag{2}$$

as the "service time analog" of $A(t)$. We need to keep in mind that as time progresses, there are periods where the system is empty. Hence, $S(t)$ is not necessarily the number of customers served during $[0, t]$. We discuss this in more detail below.

Endogenous processes Given a realisation of $\{A(t),\ t \geq 0\}$ and $\{S_n\}_{n=1}^{\infty}$ together with some initial conditions, the essence of queueing modelling is the description and analysis of *endogenous processes* that evolve. These include:

The *system size process*, $Q(t)$. This process specifies the number of items in the system at time t. Note that it is often called the *queue length process* even though it includes the customers being served (if any), as well as any customers waiting in the queue.

The *waiting time sequence*, $\{W_n\}_{n=1}^{\infty}$. Here, W_n is the waiting time before entering service for the nth customer.

The *workload process*, $V(t)$. This process determines the volume of work in the system at any time t. It is sometimes called the *virtual waiting time process* as it indicates how long a customer arriving at time t will need to wait (under the FCFS regime).

The *departure process*, $D(t)$. This process counts the number of departures (service completions) from the system during $[0, t]$.

The *busy period sequence*, $\{B_n\}_{n=1}^{\infty}$. A *busy period* is a duration of time during which the server is busy. It starts at time τ when $Q(\tau^-) = 0$ and $Q(\tau) = 1$, with $A(\tau) - A(\tau^-) = 1$ due to a customer arrival. It then ends in the first time $\tilde{\tau} > \tau$ such that $Q(\tilde{\tau}) = 0$.

There are various ways to define the functional relationship mapping the exogenous processes and initial conditions to the above endogenous processes. One such simple example is based on the customer conservation equation,

$$Q(t) = Q(0) + A(t) - D(t).$$

This equation is useful if we describe the endogenous departure process, $D(t)$, in a different manner. To do so, we observe that $S(t)$ in (2) determines how many customers could potentially be served during $[0, t]$ if there was always a customer present. Now, also define the *idle-time process*,

$$I(t) = \int_0^t \mathbf{1}\{Q(u) = 0\}\mathrm{d}u,$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. As a consequence, we can represent the departure process via the composition

$$D(t) = S\big(t - I(t)\big) = S\Big(\int_0^t \mathbf{1}\{Q(u) > 0\}\mathrm{d}u\Big). \tag{3}$$

Given initial conditions, there is a unique endogenous process $Q(t)$ satisfying

$$Q(t) = Q(0) + A(t) - S\Big(\int_0^t \mathbf{1}\{Q(u) > 0\}\mathrm{d}u\Big), \tag{4}$$

see [35] for details. A solution of $Q(\cdot)$ for (4) also describes $D(\cdot)$ via (3). In our context, Eqs. (3) and (4) serve the purpose of illustrating that the exogenous processes, $A(\cdot)$ and $S(\cdot)$, can be used to construct the endogenous processes $D(\cdot)$ and $Q(\cdot)$. Similarly, the workload process $V(\cdot)$ can be expressed in terms of the exogenous processes via

$$V(t) = V(0) + \sum_{i=1}^{A(t)} S_i - \int_0^t \mathbf{1}\{V(u) > 0\}\mathrm{d}u. \tag{5}$$

Note that the integrals in (4) and (5) are identical because $Q(t) > 0$ if and only if $V(t) > 0$.

Assume $Q(0) = 0$ and define the sequences u_n for $n = 1, 2, \ldots$ and v_n for $n = 0, 1, 2, \ldots$, with $v_0 = 0$, such that

$$u_n = \inf\{t > v_{n-1} \,:\, Q(t) > 0\}, \qquad v_n = \inf\{t > u_n \,:\, Q(t) = 0\}. \tag{6}$$

Then, $B_n = v_n - u_n$ is the duration of the nth busy period. An analogous definition can be constructed when $Q(0) > 0$. This shows how the endogenous process $\{B_n\}$ can be constructed based on the exogenous processes.

In a similar spirit, the classic Lindley recursion,

$$W_{n+1} = \max\{W_n + S_n - (T_{n+1} - T_n), 0\},$$

determines W_n based on the primitive sequences $\{T_n\}$ and $\{S_n\}$. Here, in agreement with $Q(0) = 0$, we initialize the recursion with $W_1 = 0$. See, for example, [29] for a modern treatment.

Stability and traffic intensity The notion of stability and a parameter known as traffic intensity appears in almost all queueing models. A canonical example is a GI/G/1 queueing system with arrival rate λ and mean service time m. Such a system may behave differently depending on the value of $\rho = \lambda m$. If $\rho < 1$, then queues are stochastically stable, meaning that (under regularity conditions on the interarrival and service time distributions) as $t \to \infty$, the queue length process $Q(t)$ and waiting time processes $\{W_n\}$ converge to limiting distributions. Similarly, if $\rho > 1$, then there is not enough capacity in the system to serve the incoming traffic and as t grows, queues grow without bound almost surely. Further, at the critical value $\rho = 1$, there is not a limiting distribution; however, the queue "grows at a slower rate" as attested by the fact that the system empties infinitely often (yet with heavy tailed gaps between such instances). See, for example, [55]. In all these cases, it is clear that the *offered load*, ρ, also known as the *traffic intensity*, is a key quantity.

Probabilistic performance measures Queueing theory assumes a stochastic description of the exogenous processes and endeavours to determine stochastic descriptions of

the endogenous processes. The types of stochastic processes involved include Markov processes, diffusion processes, Lévy processes, and the analysis often involves associated limiting results for these types of processes. Frequently, it is not possible to obtain a full description of the probability law of the endogenous processes, and we settle for summary measures such as the stationary queue length distribution or the stationary mean waiting time.

An example of a key result in classical queueing theory is the Pollaczek–Khinchin (P-K) formula for M/G/1 queues. Here, the arrival rate is λ, and the service distribution has Laplace–Stieltjes transform $G(s)$ with mean m such that $\rho = \lambda m < 1$. One version of the P-K formula gives an expression for the probability generating function $K(z)$ of the steady-state queue length, whose random variable we denote by Q. In this case, P-K reads

$$K(z) = \sum_{k=0}^{\infty} z^k \, \mathbb{P}(Q = k) = (1-\rho)\frac{(1-z)G\big(\lambda(1-z)\big)}{G\big(\lambda(1-z)\big) - z}, \qquad \text{for} \qquad |z| \le 1.$$

In the M/D/1 case, $G(s) = e^{-sm}$, and hence

$$K(z) = (1-\rho)\frac{(1-z)e^{-\rho(1-z)}}{e^{-\rho(1-z)} - z}. \tag{7}$$

Now, the stationary mean queue length can be computed by taking the first derivative and evaluating the limit as $z \to 1$. Further, the second factorial moment can be computed by taking the second derivative and evaluating the limit as $z \to 1$. From these, we get the M/D/1 mean and variance:

$$\mathbb{E}[Q] = \frac{2-\rho}{2}\frac{\rho}{1-\rho}, \quad \mathrm{Var}(Q) = \frac{12 - \rho(18 + \rho(\rho - 10))}{12}\frac{\rho}{(1-\rho)^2}. \tag{8}$$

We use these formulas in Sect. 2.4, illustrating statistical methods. When the service time distribution is exponential, the system is called an M/M/1 queue, in which case $G(s) = (1+ms)^{-1}$, and thus

$$K(z) = \frac{1-\rho}{1-\rho z},$$

which is the generating function of a geometric distribution with support $0, 1, 2, \ldots$, mean $\mathbb{E}[Q] = \rho/(1-\rho)$, and variance $\mathrm{Var}(Q) = \rho/(1-\rho)^2$. Comparing with (8), we see that as $\rho \to 1$, the mean queue length is reduced by a factor of almost 2 and the variance by a factor of almost 4.

Little's law Under general conditions, we can show that queueing systems in steady state satisfy Little's Law:

$$\ell = \lambda\tau, \tag{9}$$

where ℓ is the steady-state number of customers in the system, λ is the arrival rate of customers through the system, and τ is the mean steady-state sojourn time of each customer. The interpretation of "system" can change depending on context. For example, Little's law holds for the waiting room of customers, or the total service facility (the waiting room together with customers in service).

To illustrate the use of Little's law, we can use the expectation in (8) to obtain (1). For this, observe that the mean sojourn time τ is the sum of the mean service time m and the mean waiting time, w. Now, solving $\frac{2-\rho}{2}\frac{\rho}{1-\rho} = \lambda(w+m)$ for w, we obtain (1).

Stochastic process limits in queueing theory We often wish to consider models with more complexity than the M/G/1 queue. This can include either a more detailed description of how customers, servers, and allocation policies interact, or more general assumptions on the endogenous processes. In such cases, exact results such as the P-K formula are often not attainable. Nevertheless, much of the effort in queueing theory research over the past few decades has focused on more involved models. A key approach is to use stochastic process limits which give a theoretical basis for approximating the endogenous processes via limiting processes which are easier to handle.

The basic building blocks of such mechanisms involve *fluid limits* and *diffusion limits*. The idea of a fluid limit is to consider only the first-order deterministic approximation of associated processes. For example, the arrival counting process $A(t)$ can be approximated by the function $\bar{A}(t) = \lambda t$, and similarly for $S(t)$. Such a view of queueing systems ignores the randomness but often captures the essence of the system, especially when considering stability or the behaviour at large. For example, a fluid limit approximation of a GI/G/1 queue starting with $Q(0) = q_0$ customers is

$$\bar{Q}(t) = \max\left\{q_0 - (\frac{1}{m} - \lambda)t, 0\right\}.$$

Such an approximation indicates that if $\rho \geq 1$, then the queue will not "drain", whereas if $\rho < 1$, then at approximately $t = m\, q_0/(1-\rho)$ the queue will hit zero. This is a good approximation if the process starts with a large initial state q_0. Fluid limit-based approximations clearly ignore the subtle stochastic variations that play a key role in results such as the P-K formula presented above.

A second-order refinement considers the fact that deviations between exogenous processes such as $A(t)$ and their fluid limit $\bar{A}(t)$ can often be approximated via diffusion processes. Here, the idea is to consider a sequence of systems indexed by $n = 1, 2, \ldots$ and construct processes such as

$$\breve{A}_n(t) = \frac{A(nt) - \bar{A}(nt)}{\sqrt{n}}.$$

It then turns out that, under mild assumptions on $A(\cdot)$, the sequence of processes $\{\breve{A}_n(\cdot)\}$ converges weakly to a drift-less Brownian motion process. Carrying out such approximations on all or some of the exogenous processes allows us to derive approximations for the endogenous processes.

A very fruitful framework occurs when we also scale the parameters of the exogenous processes such that $\rho_n \to 1$ from below. This suite of limiting regimes yields

heavy traffic approximations for the endogenous processes and their performance measures. Other forms of scaling such as the Halfin–Whitt regime, also known as the quality and efficiency driven (QED) regime, are also very popular. See, for example, [127] for an overview.

To get a feel for the strength of such methods consider approximating the waiting time distribution of a GI/G/1 queue as $\rho_n \to 1$; see Corollary 7.5, [12]. Referring to our M/D/1 example, while we can use the P-K formula to compute clean expressions such as (1), computing the actual distribution of the stationary waiting time is not as simple. Nevertheless, a heavy traffic approximation such as that in Corollary 7.5 in [12] implies that with $\rho \approx 1$, the waiting time distribution is approximately exponential with parameters that depend on the mean and variance of the interarrival time T and the service time S. Specifically, for the M/D/1 model

$$\lim_{\rho\to 1} \mathbb{P}(W_\rho > x) = \exp\left\{-\rho^2 \frac{2(1-\rho)}{m\rho} x\right\} = \exp\left\{-\rho^2 \frac{1}{\mathbb{E}[W_\rho]} x\right\}, \qquad \text{for} \quad x \geq 0.$$

This is an exponential distribution with mean $\rho^{-2}\mathbb{E}[W_\rho] \approx \mathbb{E}[W_\rho]$. For more general GI/G/1 queueing systems, we are not able to compute $\mathbb{E}[W_\rho]$ explicitly; nevertheless, a heavy traffic limit approximation such as Corollary 7.5 in [12] is very powerful because all that is needed is the mean and variance of the interarrival and service times.

Additional branches of queueing theory We should also mention several other subfields of queueing theory that have allowed us to obtain results for the endogenous processes. One major branch is *matrix analytic methods*, which involves modelling the queueing processes with structured Markov chains which are amenable to algorithmic computation of certain performance measures; see [89]. Another branch deals with networks of queues, where even though the system is often quite high dimensional and complex, under general assumptions, one may often show that the stationary distribution possesses a *product form* structure. See, for example, the classic book [82], or more modern treatments in [30]. A third branch involves *tail asymptotics*, where results dealing with probabilities such as $\mathbb{P}(W > x)$ are approximated for large x.

2.3 Observation schemes

Having explored elements of queueing theory and related processes, we now present a possible classification of observation schemes for inference. When characterising methods and results associated with queueing inference, a first step is to consider which processes are observed and which are not. For example, we may observe the queue length process, the workload process, the arrival process, the service process, or some combination thereof. A second step is to consider how well these processes are observed, for example, fully or only at discrete time points. Table 1 contains a few major descriptors that we shall use to organize the discussion below.

Note that it is possible that a particular inference activity could potentially be classified under more than one descriptor. Our purpose is not to fully categorize a situation via the classification but rather indicate the general nature of the information

Table 1 Different kinds of observation schemes

Notation	Observation scheme
(F)	Full observation: This is the situation where the relevant random processes are observed continuously, possibly during some bounded time interval
(DI)	Discrete intervals: This is the situation where one or more of the random processes are sampled at some discrete time points. An example can be the case where in a single server queue, we sample every Δ time units. Hence, the data are $\{Q(i\Delta)\}$, for $i = 1, \ldots, n$
(IO)	Input and output process observation: This is the situation where only the arrival process and departure process are sampled. An example is an infinite server queue where we see customer arrival times and departure times but do not know with certainty how to match arrivals and departures
(P)	Probing: This is the situation where customer journeys of only selected (often manually injected) customers are observed. An example can be a communication network where there is a major traffic flow and we are injecting occasional probe customers to measure behaviour
(T)	Transactional observations: This is the situation where only service commencements and completions are observed together with an indication of server businesses. For example, such an observation scheme may occur in an automatic teller machine where queues are unobservable but server activity is being logged
(IP)	Independent primitives: In this scheme, we do not actually observe the queueing process, but rather observe some of the random variables that construct it, often with a pre-specified number of observations

available. In certain cases and when using certain models, without considering initial conditions on the queue length, observation schemes (F) and (IO) are equivalent. In fact, based on the queue length process, we can determine the arrival and departure times uniquely for single-server models where the customer order is deterministic. However, finding the queue length from the arrival and departure epochs requires information about the initial value of the queue length.

As an illustrative example, let us consider an M/M/5 queue. This is a system with five servers, Poisson arrivals, and exponential service time for each customer. That is, if $Q(t) \leq 5$, then all customers in the system are served simultaneously during time t, and further, when $Q(t) > 5$, then $Q(t) - 5$ customers are waiting for service. Figure 1 illustrates a simulated trajectory of such a system with the purpose of highlighting several types of data sequences, relating to the different observation schemes above.

2.4 Statistical methods

The vast field of statistics provides methods for carrying out a variety of tasks. In this survey, we focus mainly, but not solely, on estimation in which case either parameters of models, state estimates, or nonparametric estimates are produced based on col-

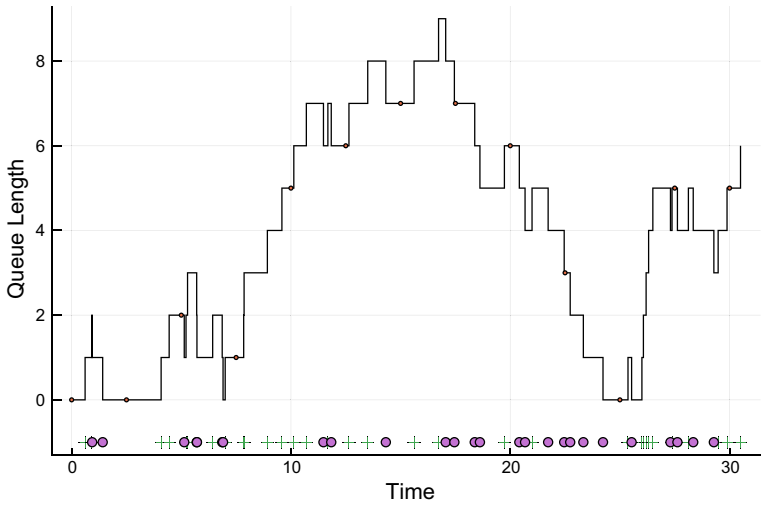

Fig. 1 A simulated sample path of $Q(t)$ for an M/M/5 system over $t \in [0, 30]$. If observed using (F), then the full trajectory of $Q(t)$ is available. If observed using (DI), then discrete samples are collected every $\Delta = 2.5$ time units. If observed using (IO), then the sequence of arrivals to the queue (marked with $+$ symbols) and the sequence of service completions (marked with $\bullet$ symbols) are available as data. The figure does not present samples using (P), (T) or (IP)

lected data. Such estimation can be carried out either in the classic frequentist setting or a Bayesian setting. The reader should keep in mind that many methods of elementary statistics are typically presented in the context of independent and identically distributed (i.i.d.) random variables. Adapting such methods to queueing inference often requires considering the dependencies and dynamics of the underlying queueing models. We now present an example.

We return to the M/D/1 queue and explore statistical inference under the (DI) observation scheme. Here, the queue length process is sampled n times every Δ time units. Taking the first observation at time Δ, the sample can then be represented as,

$$X = \Big(Q(\Delta), Q(2\Delta), Q(3\Delta), \ldots, Q(n\Delta) \Big). \tag{10}$$

We explore two alternative inference activities, both assuming the underlying system is in steady state. First assume that we simply wish to estimate the steady-state mean queue length (an endogenous performance measure) and obtain confidence intervals for our estimate. Later, we consider parameter estimation for the arrival rate λ.

Confidence intervals for $\mathbb{E}[Q]$ Using classic statistical formulas a naive approach would be to estimate the mean queue length via the sample mean,

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} Q(i\,\Delta),$$

and then to continue to obtain a 95% confidence interval

$$\left(\overline{X} - 1.96\frac{S}{\sqrt{n}}, \overline{X} + 1.96\frac{S}{\sqrt{n}}\right) \tag{11}$$

for $\mathbb{E}[Q]$, with

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Q(i\Delta) - \overline{X})^2}. \tag{12}$$

If X is an i.i.d. vector of observations that are normally distributed with mean θ, then $\sqrt{n}(\overline{X} - \theta)/S$ follows a t-distribution with $n - 1$ degrees of freedom which, for large or even reasonably sized n, is approximately a standard normal distribution. Then, 1.96 is approximately the 0.975th quantile of a standard normal distribution and this yields the confidence interval formula (11). Even if the observations are not normally distributed, the central limit theorem ensures that $\overline{X}$ has an approximate normal distribution if n is large.

While we may get away with assuming stationarity, queuing processes generally exhibit strong dependence over time. There are versions of the central limit theorem that apply to dependent sequences (see, for example, [12], p. 30). However, there is still the problem of estimating the variance of the limiting normal distribution. In particular, if Δ is not large, the covariances strongly influence the variance of $\overline{X}$ via

$$\mathrm{Var}(\overline{X}) = \frac{1}{n}\mathrm{Var}(Q) + \frac{1}{n^2}\sum_{i\neq j}\mathrm{Cov}(Q(i\Delta), Q(j\Delta)),$$

where Q represents a generic random variable of the queue size in steady state and $\mathrm{Var}(Q)$ is as in (8). Also, similar calculations yield

$$\mathbb{E}[S^2] = \mathrm{Var}(Q) + \frac{1}{(n-1)n}\sum_{i\neq j}\mathrm{Cov}(Q(i\Delta), Q(j\Delta)).$$

Hence, due to the covariance terms, the estimation of the standard deviation via (12) may be heavily biased. This jeopardizes the validity of the confidence interval (11). We demonstrate this effect via a numerical experiment.

Assume a ground truth with mean service time $m = 1$ and $\lambda = 0.9$, and hence, $\rho = 0.9$. This implies the steady-state unknown mean queue length is 4.95 as per (8). To estimate it, we could take $n = 100$ samples and consider different scenarios when Δ is in the range $10, 20, \ldots, 300$. For each value of Δ, we simulate $M = 10^4$ Monte Carlo simulations of the queue, letting it "warm up" for 10^3 time units each time. (This sets it close to "steady state".) Each simulation run samples the queue as per (10). We then estimate the coverage probability of the resulting confidence interval via

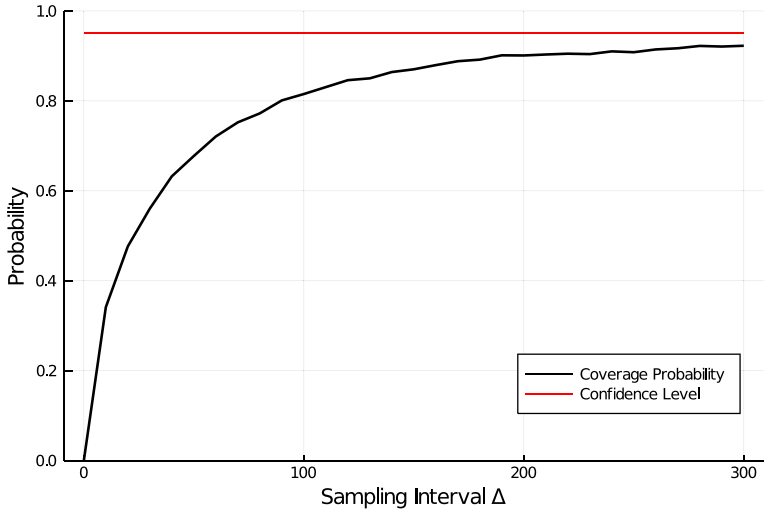

Fig. 2 Coverage probability obtained by using the confidence interval (11) with increasing Δ and $n = 100$

$$\mathcal{C}_\Delta = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}\Big\{4.95 \in \big(\overline{X} - 1.96\frac{S}{\sqrt{n}}, \overline{X} + 1.96\frac{S}{\sqrt{n}}\big)\Big\}.$$

The estimates are plotted in Fig. 2. As expected, as Δ grows, the coverage probability agrees with the i.i.d. case. However, for small Δ we see a big discrepancy between the actual coverage probability and the desired 95%. Hence, for small Δ we see that the classic confidence interval formula (11) breaks down.

Clearly, in our construction we used a naive confidence interval that assumes no covariance between $Q(i\Delta)$ and $Q(j\Delta)$ for $i \neq j$ and this is the cause of the error. We should mention that there has been extensive work on such estimation for time series where the samples are not i.i.d., see, for example, [32]. Still, in the context of queueing, one may often try to use the explicit model structure, as opposed to assuming arbitrary covariance structures as is common in the time-series literature. We survey examples of this in the sequel.

Estimating λ Say now that under the same observation scheme, we know that $m = 1$ and we wish to estimate λ. For this, we can develop an estimator based on $\mathbb{E}[Q]$ from (8). If we set $\overline{X}$ on the left-hand side of $\mathbb{E}[Q]$ in (8) and solve for λ, we obtain $\lambda = 1 + \overline{X} \pm \sqrt{1 + \overline{X}^2}$. We can take the negative sign ensuring that $\rho \in (0, 1)$. This is not hard to check for any positive $\overline{X}$. Hence, our estimator is

$$\hat{\lambda} = 1 + \overline{X} - \sqrt{1 + \overline{X}^2}. \tag{13}$$

Let us now evaluate the quality of this estimator using the mean squared error criterion,

$$\text{MSE} = \mathbb{E}[(\hat{\lambda} - \lambda)^2],$$

and determine how Δ affects the MSE. We can also reason about the limiting MSE as Δ becomes large and $n \to \infty$. For large Δ, it is reasonable to assume that $Q(i\Delta)$ and $Q(j\Delta)$ for $i \neq j$ are independent observations of the stationary queue length random variable Q. Then, from the central limit theorem, $\overline{X}$ is approximately normally distributed with mean $\mathbb{E}[Q]$ and variance $\text{Var}(Q)/n$ and the limiting mean square error is

$$\widetilde{\text{MSE}} = \int_{-\infty}^{\infty} (1 + z - \sqrt{1+z^2} - \lambda)^2 \frac{1}{\sqrt{\text{Var}(Q)/n}} \phi\Big(\frac{z - \mathbb{E}[Q]}{\sqrt{\text{Var}(Q)/n}}\Big) \text{d}z, \quad (14)$$

where $\phi(\cdot)$ is the standard normal density. Note that we are not able to evaluate the right-hand side of (14) analytically, but rather use numerical integration.

In our case with the ground truth of $\lambda = 0.9$ and $m = 1$, using (8) we have $\mathbb{E}[Q] = 4.95$ and $\text{Var}\big(Q\big) = 23.7825$. For $n = 100$, (14) yields $\widetilde{\text{MSE}} = 0.010074^2$. Also, the central limit theorem typically "kicks in" for moderate values of n. Hence for $n = 100$, assuming independence, normality effectively holds. However, for smaller Δ, the situation is different as we present in this numerical experiment.

As before, we consider different scenarios for Δ in the range $10, 20, \ldots, 300$. For each value of Δ, we simulate $M = 10^5$ Monte Carlo simulations of the queue, letting it "warm up" for 10^3 time units. Over each simulation run, we estimate $\hat{\lambda}$ and then for each value of Δ, we estimate the root MSE via

$$\text{RMSE}_\Delta = \sqrt{\frac{1}{M}\sum_{i=1}^{M} (\hat{\lambda} - 0.9)^2}. \quad (15)$$

The resulting Monte Carlo RMSE estimates are plotted in Fig. 3, and we indeed see that as Δ grows, our limiting approximation $\sqrt{\widetilde{\text{MSE}}}$ holds.

The broader view of statistics The two examples above illustrate that classical statistical methods can break down when carrying out inference for queues if dependencies are ignored. Nevertheless, when designing queue inference procedures, it is important to be aware of the vast set of tools developed in classic and contemporary statistics.

For example, the two most common ways of finding the estimated parameters are the method of moments and maximum likelihood estimation (MLE). The most important benefit of the method of moments is that it is usually fast and often non-iterative. However, like MLE, the method of moments often yields non-unique estimators and both methods are difficult to apply when the number of parameters is large. Further, in many cases, method of moment estimators are less efficient than MLE.

Likelihood-based approaches view the observed data under a certain model. MLE provides estimates for model parameters which yield the largest likelihood of the observed data. The widespread use of maximum likelihood is due to the asymptotic

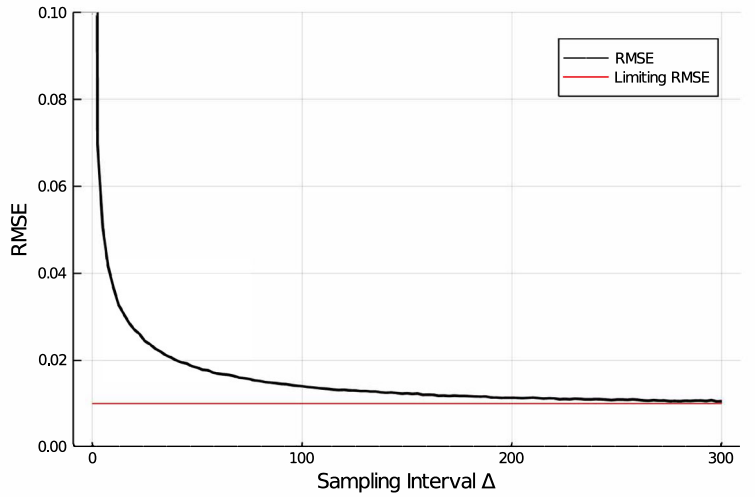

Fig. 3 The root MSE estimate (15) with increasing Δ, $n = 100$

properties of the MLE. For instance, model estimates coming from the MLE are optimal in terms of asymptotic variance. Another advantage of using MLE is that many model selection procedures such as AIC (Akaike's information criterion) and BIC (the Bayesian information criterion) are based on MLE. The computational effort required for the MLE is its major drawback. Note that even with i.i.d. data, MLEs are often biased estimators, however, under general conditions, MLEs are asymptotically unbiased. Nevertheless, with dependent data, such as waiting times, one often ends up with biased parameter estimates; see [18] for more details.

In queueing inference, we are often faced with incomplete data. One generic useful tool for this scenario is the *EM algorithm*. It is a way of maximizing the likelihood that is quite effective for estimating parameters of models with some hidden or incomplete data. The name "EM algorithm" stems from the alternating application of an *expectation* step (E-step) and a *maximization* step (M-step) that yields a successively higher likelihood of the estimated parameters. See, for example, [13], in the context of *phase-type* (PH) distributions, often closely related to queueing models. The EM algorithm is also broadly applicable for *hidden Markov models* (HMMs) which can be used to represent certain queueing inference scenarios.

An alternative is the Bayesian approach where we consider parameters to be random variables. We assign a prior probability distribution to each unknown parameter. Applying Bayes' rule, we update the parameter belief distribution from the prior to the posterior.

Another important avenue is nonparametric analysis. For stochastic models, such as queues, ideas were developed by Grübel [65]. This method is based on considering the queueing model as a mapping from the exogenous processes to the endogenous processes. Then, under certain conditions such as continuity of the mapping, applying differentials can provide information about changes to the endogenous processes based on changes of the exogenous processes. Further local properties of the functional can

give valuable information about the robustness, consistency, and asymptotic properties of the estimators; see [66].

3 Various estimation paradigms

We now move to the heart of the survey. We present and summarize a variety of results and methods. In considering queueing inference problems, there are several dimensions at play. These include:

- The physical/real-world problem being investigated and the goals associated with the application.
- The queueing model (or class of models) being used. See Sect. 2.2.
- The observation scheme. See Sect. 2.3.
- The type of statistical methodology being used. See Sect. 2.4.

The interplay of these aspects are woven into a specific domain that we call an **estimation paradigm**.

Having studied the broad literature, we decided to partition the scope of this field into ten paradigms. Each paradigm shares a specific sub-field of research. One major characteristic of each paradigm is the typical observation scheme. We summarize the paradigms and their typical observation schemes in Table 2 and survey them in the subsequent subsections. Note that for each estimation paradigm we refer only to a few selected key references. The reader can find a more complete list of references dealing with queueing inference in the annotated bibliography [9].

3.1 The classical sampling approach

When considering parameter estimation for queueing systems, a natural first step is to consider observation scheme (F), see Table 1, where all the data are available. In that case, one may think that there are not any challenges because we can simply employ the state of the art parameter estimation methods for the queueing primitives (interarrival and service times). This is not far from the truth when we have large samples available, however, for small samples there are some technical complications. These complications have driven most of the classic research on parameter estimation of queues. Much of this work is summarized in the last queueing estimation survey by Bhat et al., [22]. See also the earlier survey, [23] by Bhat and Rao.

As an example, consider a single server queue where we observe the sequence of interarrival times $\{A_1, \ldots, A_{n_a}\}$ and the sequence of service times $\{S_1, \ldots, S_{n_s}\}$, where these sequences are i.i.d., independent of each other, and the sample sizes n_a and n_s are fixed. We can then naturally estimate λ and μ via

$$\hat{\lambda} = \frac{n_a}{\sum_{i=1}^{n_a} A_i}, \qquad \text{and} \qquad \hat{\mu} = \frac{n_s}{\sum_{i=1}^{n_s} S_i}. \tag{16}$$

However, when observing a queue, n_a and n_s are often not fixed and can be dependent on the sequences $\{A_i\}$ and $\{S_i\}$. Complications may arise, not only due

Table 2 Key references for different estimation paradigms and their most relevant observation schemes

Estimation paradigm	Typical obs scheme	Key references
The Classical Sampling Approach: This is the situation where either endogenous or exogenous processes are sampled, or both. The actual number of samples might be fixed or might be a random variable determined endogenously	(F), (IP)	[19,20,22,23,37,81,83,108,110,122,124,131]
Inverse Problem Estimation: This is a situation where certain attributes of the system are observed and these observations are used to infer model parameters	(P)	[1–3,16,25,36,39,69,70,72,73,80,86,96,98,103,104,117]
Inference for Non-Interacting Systems: This paradigm deals with models where customers do not interact such as the M/G/∞ queue and generalisations	(IO), (DI)	[24,27,31,33,53,62–64,67,68,102,107,111,112]
Inference with Discrete Sampling: This paradigm focuses on cases where systems are sampled discretely over time	(DI)	[26,46,51,92,95,106,109]
Inference with Queueing Fundamentals: This paradigm describes situations where queueing theory fundamentals aid parameter and state estimation. The most prominent example is the use of Little's law	(F)	[34,52,57–61,84,85,91,99]
Queue Inference Engine Problems: This paradigm deals with a branch of problems where transactional observations are recorded and the trajectory of the queue within a given cycle is inferred	(T)	[21,42–45,47,54,56,71,77–79,88,90,101]
Bayesian Approaches: In most of the Bayesian work to date the parameter estimation utilizes known queueing performance analysis formulas, considering their posterior distributions given a sensible choice of priors	(IP)	[4–7,15,28,40,76,93,94,97,105,119,121,125],
Online Prediction: In this paradigm, we observe the states up to a given time and make prediction about future states	(F), (DI)	[38,75,100,118,120,126,129]
Implicit Models: This paradigm deals with new developments combining data science and queueing theory where queue-like models are introduced without explicitly modelling every component of the system	(F)	[14,49,50,113–116,123,128,130]
Control, Design, and Uncertainty Quantification: This paradigm deals with work related to parameter and state estimation where control and design decisions based on inferred values are to be made		[8,10,17,48,74,87]

to censoring, but also due to the dependency structure of the various quantities. Here are some possibilities:

- We may sample the system for a fixed duration $[0, T]$ in which case both the number of arrivals and the number of service completions are dependent random variables.
- We may sample a fixed number of arrivals, n_a, in which case the observation time, T, and the number of service completions are dependent random variables.
- We may sample a fixed number of service completions, n_s, in which case the observation time, T, and the number of arrivals are dependent random variables.
- We may sample using some other similar scheme (such as a fixed number of transitions) which will again imply that other quantities are random variables.

In each of these cases, the estimation procedure exhibits what we refer to as *endogenously determined sample sizes*. That is, the total number of either interarrival times, service times, or both is a random quantity resulting from the model. This aspect drove much of the early research on parameter estimation of queues and is well described in [20]. In fact, the subtle problems that arise in such cases were considered in one of the first parameter estimation papers for queues by Clarke in 1957, [37].

The work in [37] focused on parameter estimation for a stationary M/M/1 system where the parameters are the arrival rate λ and the service rate μ. When sampling an M/M/1 queue for a fixed duration $[0, T]$, it is difficult to obtain a simple likelihood expression for the unknown parameters λ and μ. Hence, a more creative sampling scheme was proposed where a set duration $\tilde{T}$ is determined and sampling takes place for as long as the busy time of the server is less than $\tilde{T}$. In such a case, standard properties of the M/M/1 queue imply that the likelihood can be written as

$$L(\lambda, \mu\ ;\ \text{data}) = \left(1 - \frac{\lambda}{\mu}\right) e^{-\mu\tilde{T} - \lambda T_{n_s}}\ \mu^{n_s - \nu}\ \lambda^{n_a + \nu} K(n_a, n_s, \nu, T_{n_s}),$$

where K does not depend on the unknown parameters λ and μ. This expression is useful because the (full observation) data is summarized via the statistics n_a, n_s, ν, and T_{n_s}. As defined previously, the statistics n_a and n_s are the number of arrivals and number of service completions (only this time endogenously determined via the sample). The statistic ν is the initial queue size. Finally, the statistic T_{n_s} is the time of the last service completion during $[0, \tilde{T}]$. This structure of the likelihood allows one to maximize with respect to λ and μ given measurements of the sufficient statistics. Further, there is also the (minor) extra added benefit that the initial queue length, $\nu = Q(0)$ can yield more information for this observation scheme.

As exemplified by the results of [37], the (F) observation scheme, while simple, still entails some interesting challenges. However, when considering larger sample sizes, the subtle issues associated with the construction of MLEs and similar estimators are not as crucial. Nevertheless, a significant body of literature has handled such queuing inference problems. For instance, in [122] the problem of estimation for tandem queues was discussed as a special case of Jackson networks. Further, in [108], the case of a general G/G/1 retrial queue was considered. Here, the flow of repeated attempts can be non-Markovian and the system is observed until there is a fixed number of departures.

Then, in [19] estimation of the parameters of GI/G/1 queues was studied where only the incomplete information of the differences between service and interarrival times was observed. Also, in [124], MLEs for service demands in closed queueing networks with load-independent and load-dependent stations were proposed.

Going back to the simple estimator (16), even in the situation where n_a and n_s are fixed, there may be anomalies in the inference process. For example, combining $\hat{\lambda}$ and $\hat{\mu}$ from (16) we have an estimator for the offered load,

$$\hat{\rho} = \frac{\hat{\lambda}}{\hat{\mu}}, \tag{17}$$

which may appear straightforward. However, in [110], Schruben and Kulkarni showed that for an M/M/1 queue, if we wish to use $\hat{\rho}$ to compute the steady state mean queue length, some unexpected behaviour may occur. The ratio of the estimated traffic intensity to the true traffic intensity has an F distribution with $2n_s$ and $2n_a$ degrees of freedom. Further, they showed that this estimator has undesirable sampling properties. For example, even when we restrict the estimated workload to be strictly less than one (for instance, by re-sampling for the case that $\hat{\rho} \geq 1$), the expected value of the plug-in estimator $\frac{\hat{\rho}}{1-\hat{\rho}}$ for the average number of customers is infinite. That is,

$$\mathbb{E}\left[\frac{\hat{\rho}}{1-\hat{\rho}}\ \mathbf{1}\{\hat{\rho} < 1\}\right] = \infty. \tag{18}$$

These types of results indicating *anomalies in inference* are useful to keep in mind when a practitioner estimates primitives as inputs into basic queueing models such as M/M/1, but also for more complex discrete event simulation models. That is, in simulation modelling practice one often considers system primitives as inputs into a complex discrete event simulation. Then, a discrete event (say agent-based) simulation model can be used for performance analysis. The take-home message from a simple result such as (18) is that such problems can also occur in much more complex models.

The results from [110] were generalized in [131] where alternative estimators for the limiting expected number of customers in the queue (and several other performance measures) were constructed. These estimators require the analyst to choose a value $\rho_0 < 1$. Under the assumption that $\rho < \rho_0$, the estimator has finite mean and finite mean square error. Further, in [81], similar estimators were considered including the consideration of bootstrap-based confidence intervals as well as other statistical aspects. A third notable paper dealing with this aspect of queueing estimation was Kiessler and Lund [83], where the authors proposed and analysed two alternative estimators for ρ in M/G/1 queues.

The first estimator discussed in [83] uses the sample average of the work arriving into the system during $[0, T]$:

$$\hat{\rho}_{\text{work}} = \frac{\sum_{i=1}^{N(T)} S_i}{T} = \hat{\rho}\,\frac{\sum_{i=1}^{N(T)} A_i}{T}, \tag{19}$$

where $N(T)$ is the total number of customers arriving up to time T, $\{A_i\}$ and $\{S_i\}$ are as above, and $\hat{\rho}$ is similar to the estimator in (17) with n_a and n_s equalling $N(T)$.

The second estimator uses the proportion of time during which the server is busy:

$$\hat{\rho}_{\text{virtual}} = \frac{\int_0^T \mathbf{1}\{V(u) > 0\}\,\mathrm{d}u}{T}, \tag{20}$$

where $V(t)$ is the workload process. The paper showed that (20) is an asymptotically unbiased estimator and further provided analysis of asymptotic means, biases, and variances.

3.2 Inverse problem estimation

In general, an inverse problem arises in a situation where we observe endogenous processes and need to estimate or predict parameters of exogenous processes. As a simplest example, consider a stationary M/D/1 queue and the mean waiting time given in (1). Assume we know the value of the mean service time m and observe the waiting times of n customers. If the observed sample average waiting time is $\overline{W}$, then we can estimate ρ via the equation

$$\overline{W} = m\frac{\rho}{2(1-\rho)}. \tag{21}$$

Solving for ρ, we obtain an estimator

$$\hat{\rho} = \frac{\overline{W}}{\overline{W} + m/2}. \tag{22}$$

Much of the work dealing with inverse problems has focused on the (P) observation scheme; see Table 1. In certain situations, endogenous processes, such as $\{W_n\}$ yielding $\overline{W}$, are not directly observable, and we need to obtain observations by actively probing the system via artificial customers (packets). The prober usually chooses the sizes of probes and time epochs in which to send them. In this case, we say that the probing is *active*. However, sometimes the probe sizes are determined by the selected application and associated network protocol such as transport control protocol (TCP). In this case, we have *passive* probing.

Probing is depicted in Fig. 4. In such a case, the probes slightly affect the system, and via their measurements we aim to solve an inverse problem.

Continuing with the M/D/1 example, assume we wish to estimate ρ. To do this, the prober sends n probes into the queue at the time points of a truncated Poisson stream with rate γ. Here, γ should typically be quite small so as not to disturb the system. For each probing customer, the prober measures the sojourn times denoted via $\tau_1, \ldots, \tau_n$ with $\overline{\tau}_p$ denoting their average. Then, we can estimate $\overline{W}$ from the waiting times experienced by the probes via $\overline{W} = \overline{\tau}_p - m$. Then, using Eq. (21) again with $\rho = (\lambda + \gamma)m$, we can solve for λm to obtain the estimator

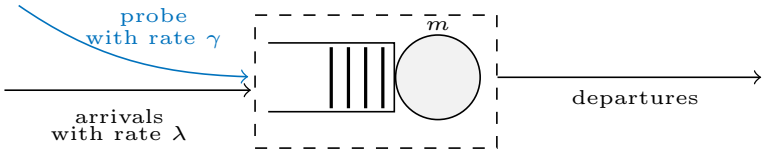

Fig. 4 Probing a single server queue. Regular arrivals arrive at Poisson rate λ. The probe stream is injected with rate γ. The service duration for both probes and regular packets has a mean of m

$$\hat{\rho}_{\text{M/D/1 probes}} = \frac{\overline{\tau}_p - m}{\overline{\tau}_p - m/2} - \gamma\, m. \tag{23}$$

As a comparison, consider the case where the underlying system is a stationary M/M/1 queue with a known mean service time m, and arrival rate λ which is unknown. In this case, since the M/M/1 mean sojourn time is $(m^{-1} - \lambda - \gamma)^{-1}$, a probing-based estimator for the offered load is

$$\hat{\rho}_{\text{M/M/1 probes}} = \frac{\overline{\tau}_p - m}{\overline{\tau}_p} - \gamma\, m. \tag{24}$$

Note that, in practice, we do not always know whether the underlying system is better modelled as an M/D/1 queue, an M/M/1 queue, or some other model. Still, for a given observations yielding $\overline{\tau}_p$, the estimators in (23) and (24) show that treating the system as an M/D/1 queue will yield a higher offered load estimate than treating it as an M/M/1 queue.

One of the first probing papers [36] by Chen et. al. considered a FCFS M/D/1 system. The prober knows the arrival times, waiting times, and departure times of probes. The authors derived a tractable expression for the likelihood function of λ. This allows us to carry out maximum likelihood estimation.

In the general case of G/G/1 queues, assume that λ and μ_1 are the average arrival rate and service rate of local traffic, and γ and μ_2 are the average arrival rate and service rate of the probes. Then, the traffic intensity is

$$\rho = \frac{\lambda}{\mu_1} + \frac{\gamma}{\mu_2}, \tag{25}$$

which, in the case that $\rho < 1$, is the stationary fraction of time during which the server is busy. This can form the basis of an estimator. If γ, μ_1 and μ_2 from (25) are known, then we can estimate λ by using (25) together with an estimator like (20).

However, in reality the $Q(t)$ (or $V(t)$) is often not observed. So, in [117] Sharma and Mazumdar introduced a method based on measuring the delay experienced with active probing. They solved the problem for the cases where the true arrival process is either Poisson or arbitrary and the probing process is Poisson, denoted by M+M/G/1 and M+G/G/1, respectively. Further, they extended the problem to cases where the service times are unknown as well as queueing networks. This was the first paper that proposed an analytic approach to probing of queueing networks.

In [1], Alouf et al. studied the case where the system has limited unknown capacity c. They considered Poisson arrivals and service times that are either exponential or deterministic and denoted the systems via M+M/M/1/c and M+M/D/1/c. Their estimators for c can be used to estimate system capacity.

When $\rho \geq 1$, a different approach can be used. In [72], Hei et al. observed that the ratio $\mathcal{R}$ between the mean interarrival and mean interdeparture time is given by

$$\mathcal{R} = \begin{cases} 1 & \rho < 1, \\ \frac{\lambda}{\mu_1} + \frac{\gamma}{\mu_2} & \rho \geq 1. \end{cases}$$

In this case, an estimate of $\mathcal{R}$ based on averages of observed interarrival and interdeparture times can be used to estimate λ.

When $\rho < 1$, the ratio $\mathcal{R} = 1$. This yields no information about the arrival rate, and we need to consider the second-order characteristics of the departure process. The authors suggested using an approximation of the squared coefficient of variation (SCV) of the interdeparture times to estimate the arrival rate of a D+M/G_i/1 queue where the notation G_i indicates that the service distribution may differ between the customers and the probes. In follow-up work [73], the authors extended the method to the case of M+M/G_i/1 queues.

Analysis for special cases of probing and development of new estimation methods was a very active area of research around the turn of the first decade of the current century. In [96], Nam et al. considered probing for parameter estimation of an M/G/1 queue where both the service rate $\mu = 1/m$ and the input traffic load λ are unknown. Their method estimates the available bandwidth (the residual processing capacity) based on what they call a minimal-backlogging method. Comert and Cetin in [39] considered the application of probing for real-time estimation of the number of vehicles (customers in a queue) in a signalized traffic intersection. This is the case that only the position of the last probing customer in the queue is known. In [98], the authors studied the convergence rate of an M/D/1 queue to its steady state as a function of the load. They attempted to use this performance measure for finding an adequate probe separation threshold.

In a significant paper [16], Baccelli, Kauffmann, and Veitch described how to apply probing methods for queueing networks. Following the initial work in [117], the work in [16] presented a comprehensive survey of probing methods to estimate parameters and design queueing networks. See also [103] where the authors dealt with network tomography and further exploited the EM algorithms for multicast trees.

In [70], Heckmüller and Wolfinger studied estimation of the arrival rate of a G/D/1 queue with probing where only the departure times are observed. Their method was constructed in a discrete time setting, approximating the numbers of customers arriving in time intervals by Gaussian random variables. They also investigated sequences of queues with possibly varying bottleneck capacity. In [80], Kauffmann suggested a new approach with zero probing overhead based on the theory of inverse problems for bandwidth sharing networks. In [2], Antunes et al. considered the problem of estimating the arrival rate and the service rate of an M/G/1 queue with probing. They also studied the time-varying M_t/G_t/1 queues in [3]. In [86], Kim et al. applied a data

driven probing approach to provide a high-fidelity simulation model for an arrival process to a clinic.

Nonparametric problems While the (P) observation scheme, see Table 1, has attracted most of the attention in the realm of inverse problems, other types of settings are also relevant in practice. For example, nonparametric estimation for the M/G/1 queue was considered in [25] based on busy periods and in [69] using the P-K formula. Based on workload observations, the authors of [69] constructed estimators both for the offered load ρ and the (nonparametric) service time distribution. In earlier work, [104], Pitts studied inference for GI/G/1 queueing models and laid foundations for nonparametric inference. Further nonparametric work appears in the context of infinite server queues, a topic which we cover next.

3.3 Inference for non-interacting systems

We now discuss inference associated with the M/G/∞ queue and similar models. We call such systems *non-interacting* because customers do not affect each other in the queue. Almost any probabilistic analysis of an M/G/∞ model is based on a transformation of the Poisson arrival process and this is why many M/G/∞ results (and generalizations) are tractable. While there is not real "queueing" taking place in such a model, infinite server systems naturally appear in applications as they describe a situation where incoming customers experience random delay. One example is a pedestrian crossing tunnel where pedestrians do not really interact, and the delay between the entry time and the exit time of each pedestrian is i.i.d. random variables.

Just as an illustration, we can compare the formulas for the auto-covariance function of the stationary queue length for an M/M/1 queue and an M/M/∞ queue, both with a mean service time of 1. From [107], the auto-covariance is,

$$\mathrm{Cov}\big(Q(0), Q(t)\big) = \begin{cases} \dfrac{2\lambda(1-\lambda)}{\pi} \displaystyle\int_0^\pi \frac{(\sin\theta)^2 e^{-t(1+\lambda-2\sqrt{\lambda}\cos\theta)}}{(1+\lambda-2\sqrt{\lambda}\cos\theta)^3}\mathrm{d}\theta & \text{for M/M/1,} \\ \lambda e^{-t} & \text{for M/M/}\infty. \end{cases}$$

As is evident, having non-interacting customers (M/M/∞) yields a much simpler formula.

The analytic tractability of M/M/∞ queues (and M/G/∞ queues for that matter) goes beyond the auto-covariance function. Many performance measures have closed-form expressions involving the arrival rate and the service time distribution $G(x)$. This has motivated several authors to consider inverse problems in various settings. Specifically, observation schemes of types (DI), (IO), and (T) given in Table 1 have been studied.

The (IO) observation scheme A neat initial M/G/∞ result from Brown in [33] deals with the transformation of the service time distribution

$$H(x) = 1 - (1 - G(x))e^{-\lambda x}. \tag{26}$$

As is easy to see (Lemma 2 of [33]), $H(\cdot)$ happens to be the distribution function of the time since the last arrival when observing the process at a departure point. The key to seeing this is to note that the last arrival does not necessarily correspond to the observed departure.

Now, with (26) in hand, there is an immediate scheme for conducting nonparametric inference of $G(\cdot)$ (and λ) under the (IO) observation scheme. By observing the sequence of arrivals and departures, we can construct the empirical distribution function that estimates $H(\cdot)$ and also obtain an estimate for λ based on the arrival rate. Then, (26) can be used to find an estimate of $G(\cdot)$.

This general idea was revived and extended by Blanghaps et al. in a more recent paper, [27], where the distribution of the rth latest arrival, $H^{(r)}(\cdot)$, was considered. The relation between $G(x)$ and $H^{(r)}(\cdot)$ is given by

$$H^{(r)}(x) = 1 - \big(1 - G(x)\big)e^{-\lambda x}\,\frac{(\lambda x)^{r-1}}{(r-1)!} - \sum_{j=0}^{r-2} \frac{e^{-\lambda x}(\lambda x)^j}{j!}.$$

As shown in [27], the improvement in estimating $G(x)$ through $H^{(r)}(\cdot)$ is considerable when ρ is greater than 1.

The paper [67] by Grübel and Wegener treated the same problem, but there the authors seemed to not be aware of the [33] result (and idea). Hence, in Sect. 2 of that paper, they analysed the problem using the concept of *matchmaking* (guessing what departure maps to what arrival). They developed a method for matchmaking for the case in which the distribution of the sojourn times is either exponential, log-convex or log-concave. For the last two cases, they showed that this match is unique. That paper also goes a bit further to provide an hypothesis test for determining whether the service times in M/G/∞ are exponential.

Infinite server queues under (DI) In contrast to many other queueing systems, the M/G/∞ queue has explicit expressions for the joint distribution of $Q(t_1), \ldots, Q(t_n)$. This allows one to carry out effective inference in the discrete intervals (DI) observation scheme. In [62], Goldenshluger constructed a nonparametric estimator based on discrete observations which exploits a relationship between the derivative of the covariance function and the distribution G. The study of nonparametric inference for this queue was then extended in [63], where Goldenshluger considered the variant where the arrival and departure epochs are registered without knowledge of the epoch type. That paper contains further results and comparisons to previous estimators.

Extended models include the time inhomogeneous case studied in [64] by Goldenshluger and Koops. Further, in a discrete time setting, in [53], Edelman and Wichelhaus considered parameter estimation for two-node networks of infinite server queues with geometric arrivals and general service times. A related paper, [111], studied parameter estimation for discrete time G/G/∞ queues. The work was extended in [112] in the context of queueing networks.

Observing busy periods Consider a situation where the sequence of busy periods $\{B_n\}$, as well as idle periods, is observed. This was considered in [68], where a sequence of busy period observations is used to construct empirical approximations to the distribution function of the service time. A related paper is [24].

Inference procedures for M/G/∞ queues based on busy and idle periods may seem attractive from a statistical perspective, but from a practical (and/or queueing) perspective they are less useful. As an example, consider an M/G/∞ queue with $\lambda = 50$ (customers per minute) and $\mu = 1$ (mean service time of a customer is 1 minute). This could model, for example, the underground crossing of a major street mentioned above where the walking time is about 1 minute and there are about 50 people entering the crossing every minute. In such a case, the stationary queue length is known to be Poisson distributed with parameter $\rho = 50$ and the stationary probability of being empty is thus e^{-50}. Now, the time between idle periods, i.e. $\tau = \inf\{t > 0 \mid Q(t) = 0\}$, satisfies

$$\mathbb{E}[\tau \mid Q(0^-) = 0, Q(0) = 1] = \lambda e^{\rho} = e^{50}.$$

Hence, it is not reasonable to expect to actually collect any data in such a scenario. Thus, papers that base their statistical analysis on this observation scheme essentially deal with a situation that is unlikely to occur in practice.

More related work In [102], Pickands and Stine considered a discrete time infinite server system with geometric service times where the queue size is the only observation. They proposed an estimator for the arrival rate and the holding time distribution. Their key contribution was that they model the situation with a HMM where the hidden component was the order of arrivals and departures. They used HMM algorithms and the correlation structure of the process for constructing MLEs. A related line is by Brillinger [31] where he developed a spectral approach for estimator construction of G/G/∞ models and generalizations.

3.4 Inference with discrete sampling

In the previous section, we overviewed work dealing with the (DI) observation scheme for infinite server queues. We now discuss estimation under this observation scheme for other models. In computerized applications, the data often include a full log of queueing information. However, in physical systems, periodic logging is often more sensible. We illustrated such an example in Sect. 2.4, where, for instance, the sample of the queue length over discrete intervals is given in (10).

The difficulty with discrete sampling is that unless we are considering non-interacting systems as in the previous section, the joint distribution of the samples is typically intractable. Hence, research in this area often builds on approximations or modifications of the sampling scheme. As illustrated in Sect. 2.4, if the interval between samples is very large, then we can use a crude approximation where the samples are assumed to be independent. However, one needs to keep in mind that the assumed stationarity of the system is questionable when considering large sampling intervals.

For continuous time Markov chain models with small finite state spaces, sampled at discrete intervals, one can construct maximum likelihood estimation procedures. For instance, see [26], where Bladt and Sørensen established the existence and uniqueness of the MLE and compared the use of the EM-algorithm and alternative Markov chain

Monte Carlo (MCMC)-based procedures. This method can be applied to small finite state Markovian models of queueing systems. However, the method quickly becomes intractable as the state space grows.

In terms of approximations, in [109], Ross et. al. considered M/M/c queueing systems under the (DI) observation scheme. They approximated the process using Ornstein–Uhlenbeck diffusion approximations which work well when the number of servers c is not small. They carried out MLE estimates on the approximated model. A related preprint is [95], where McVinish and Pollett considered the method of "estimating equations", which to the best of our knowledge has not been exploited further in the context of queueing inference.

A modification of the (DI) observation scheme is to use Poisson probing, where samples occur at times dictated by a Poisson process independent of the other system processes. For models such as the M/G/1 queue, as well as more general Lévy-driven storage systems, Poisson probing yields tractable estimators. In [106], Ravner et al. exploited the fact that the dependence structure of the workload process, sampled according to a Poisson process, has closed form. Specifically, given the value of the workload process at a specified time, the Laplace transform of the workload process at an exponential future time was explicitly derived. They exploited this structure to carry out, and analyse, semi-parametric estimation of the Lévy exponent driving such queues. Further, in [92], Mandjes and Ravner considered hypothesis testing for such systems. Related to these papers is [51] by Duffie and Glynn, where the authors introduced a generalization of the method of moments for continuous time Markov chains sampled at random time intervals. Another related paper is [46], where den Boer and Mandjes considered a general estimation problem using Laplace transforms, also with application to the M/G/1 queue.

3.5 Inference with queueing fundamentals

Queueing theory supports many models, each with its own properties and theoretical results. However, there are also basic fundamental properties of queues that are universal to almost any queueing model. These include Little's law, see (9), as well as general properties such as the fact that queue lengths in critically stable queues are often of the order $O\left((1-\rho)^{-1}\right)$, and the fact that tail asymptotics of waiting time and sojourn time distributions often have a known asymptotic form.

Several key papers have exploited such properties for the purpose of inference and estimation. In terms of tail asymptotics, in [58], Glynn and Torres considered how long the arrival process needs to be observed in order to accurately estimate the long-run fraction of time that the workload exceeds a given level. Their conclusion appears to hold regardless of whether the arrival process exhibits complex dependencies or not. In [61], Glynn and Zeevi established logarithmic consistency and studied the efficiency of tail-based estimators. In [52], Duffy and Meyn considered Lindley recursions similar to (6) and studied their estimation properties via a large deviations analysis.

In terms of Little's law, there has also been significant work. In many situations, one may observe either the queue length trajectory, or the sojourn times of customers, or both. Little's law ties the expected value of these two quantities, and hence, whenever

we can use one quantity for estimation, we can also use the other. As an example, refer back to the M/D/1 queue and the waiting time-based estimator, (21). In that case, the sample mean of the waiting time $\overline{W}$ was observed. However, consider a situation where instead we observe the time average of the number of waiting customers,

$$\overline{Q}_r = \frac{1}{T}\int_0^T (Q(u) - 1)\mathbb{1}\{Q(u) \geq 1\}\,\mathrm{d}u.$$

Now, by Little's formula we expect

$$\overline{Q}_r \approx \lambda \overline{W}.$$

This motivates writing down an estimator for ρ via an adaptation of (21),

$$\frac{\overline{Q}_r}{\rho/m} = m\frac{\rho}{2(1-\rho)}.$$

Now, solving the quadratic equation for ρ and choosing the nonnegative solution, we obtain the estimator

$$\hat{\rho} = \sqrt{\overline{Q}_r(2 + \overline{Q}_r)} - \overline{Q}_r. \tag{27}$$

A comparison of the estimator (27) with the estimator (22) or the similar estimator (13) indicates that there are multiple methods to estimate the same quantity. With infinitely many samples, these methods are equivalent; however, in general there is room to investigate the statistical properties of such competing estimation schemes.

Results in this spirit were analysed in depth in [60] by Glynn and Whitt. In that paper, the authors extended variance reduction results by Carson and Law, [34], and investigated trade-offs relating to Little's law-based estimation. They focused on estimation of the means of the endogenous processes, Q_r and W (or, respectively, Q and the sojourn time process). They considered the arrival rate λ as either known or unknown, and they further dichotomized between what they call the "direct estimation" and the "indirect estimation" case. In the former, the mean of Q is estimated directly by a time average of the observed queue length. In the latter, Little's law makes use of the sample mean for W together with λ or an estimator for it, to estimate the mean of Q_r (or Q). The results of this paper relied on the authors' earlier work in [59], as well as other joint papers, which established a joint functional central limit theorem based on Little's law, which describes the weak convergence of both the queue length estimator and the waiting time estimator, to an appropriate limiting diffusion process. The results in [60] indicate that an indirect estimator is more efficient than a direct estimator in cases where the interarrival and waiting times are negatively correlated.

This line of research has been further extended in [99] in the context of manufacturing performance analysis. Further, in [57], Glynn et al. established a martingale central limit theorem which they then used to construct confidence intervals for estimators and perform statistical tests. In more recent work, in [84], Kim and Whitt surveyed previous results and refined the Little's law-based estimators to handle certain cases

such as removing bias due to interval edge effects. See also [85] for a treatment of the time-varying version of Little's law.

A different line of research that we would like to highlight here is the detection of stability or instability of queueing systems. To date not much work has been done towards this direction, but one notable publication is [91] where the authors deal with Monte Carlo simulation of systems for detecting stability or instability.

3.6 Queue inference engine problems

A paper published in 1990, *The queue inference engine: deducing queue statistics from transactional data*, by Larson [88], opened up a research direction associated with retrospective state estimation of a queue based on *transactional data*. This is what we call the (T) observation scheme; see Table 1. The key idea is that only observations of ordered service entry and service completion times are available. Under this observation scheme, the cumulative number of departures for $t \in [0, T]$, $D(t)$, is observable; however, the cumulative number of arrivals $A(t)$ is not. This occurs in applications such as automatic teller machines where the queue of customers waiting for the machine is not observable, but the transaction record is logged.

The only statistical assumption is a (homogeneous) Poisson arrival process and no further assumptions on the service processes. That is, the inference works for models such as M/G/c as well as more complicated models with Poisson arrivals. As was shown in [88], smart recursions utilizing the uniform order statistics property of Poisson processes can be utilized to infer retrospective mean queue length trajectories and other quantities, during congestion periods (duration during which all servers are busy), only based on transactional observations during that period.

As an elementary illustrative example, see Fig. 5 which assumes a single server case. Here, the congestion period begins at time 0 and then a completion of a transaction is recorded (a customer departing) at time t_1. Further, the server becomes idle at time t_2. Based on this information, it is clear that there were two arrivals for this specific congestion period with the first arrival at time 0 and the second at some unknown time $X_1 \in (0, t_1)$.

Under the Poisson arrival assumption, we know that $X_1 \sim \text{uniform}(0, t_1)$. This then allows us to deduce the expected number of customers in the waiting room as

$$\hat{Q}_r(t) = \mathbb{E}[Q_r(t)] = \begin{cases} \frac{t}{t_1} & t \in [0, t_1), \\ 0 & t \in [t_1, t_2]. \end{cases} \tag{28}$$

The computation of the estimator $\hat{Q}_r(\cdot)$ is based on the transactional observations and can only be made at time t_2 once it is known that the congestion period has ended.

In a more complex situation, there will be multiple transactions recorded during a congestion period. We illustrate such a situation in Fig. 6. At time $t = 0$, an arrival occurs and the server's state changes from idle to busy; hence, the congestion period starts. From transactional data, the system exhibits both service completion and service commencements at times t_1, t_2, t_3, and t_4. Further, there are customers waiting to be served at least at times t_1^-, t_2^-, t_3^-, and t_4^-. During the congestion period, a total of

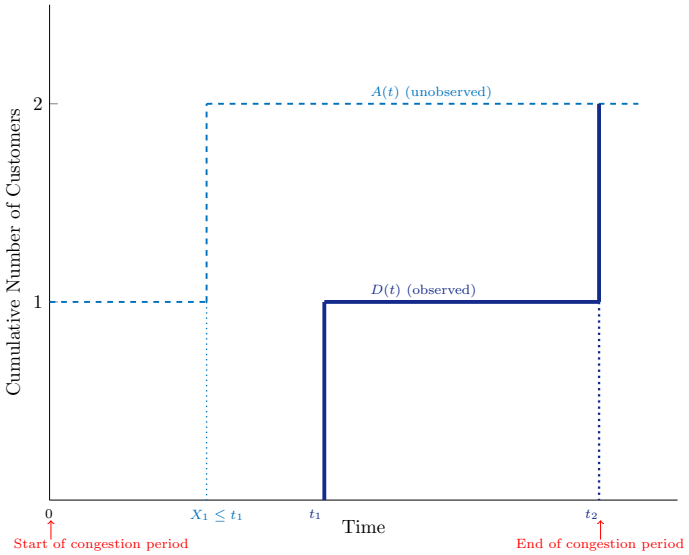

Fig. 5 The cumulative number of customers during a congestion time. Here, $A(t)$ is the cumulative number of arrivals and $D(t)$ is the cumulative number of departures during the congestion period. The arrival time of the ith customer to enter the queue and the departure time of the ith customer served are denoted by X_i and t_i, respectively

$N = 4$ customers are delayed in the queue, and at time t_5 the transactional data indicate a service completion but no service commencement, thus ending the congestion period and allowing the server to idle. It is evident that the service completion times within a congestion period impose a set of inequalities on the arrival times of other customers who waited in queue:

$$X_1 \leq t_1, \ldots \quad X_N \leq t_N. \tag{29}$$

The queue inference engine allows us to compute an estimator, $\hat{Q}_r(\cdot)$ based on this data with the main idea being a recursive computation that uses the uniform arrival property for Poisson arrival processes, similarly to the simple estimator in (28), as well as the inequalities in (29). Details are in [88].

Larson's [88] paper developed a system linear equations that can be solved at the end of each congestion period for computing $\hat{Q}_r(\cdot)$. Note that given n arrivals during a congestion period, the computation time needed to obtain mean queue lengths is of order $O(n^5)$. See [88] for numerical examples, compared to simulation.

Following Larson's paper, a variety of research papers generalized the basic idea and presented improvements to the computation time. In [21], Bertsimas and Servi improved the computational complexity and generalized the results to cases of non-homogeneous and renewal arrival processes. Note that for the homogeneous Poisson arrival process, the unordered arrival times are independent and uniformly distributed, but for the non-homogeneous case with the arrival rate $\lambda(t)$, the unordered arrival times are i.i.d. with a probability density proportional to $\lambda(t)$. In [42], Daley and Servi

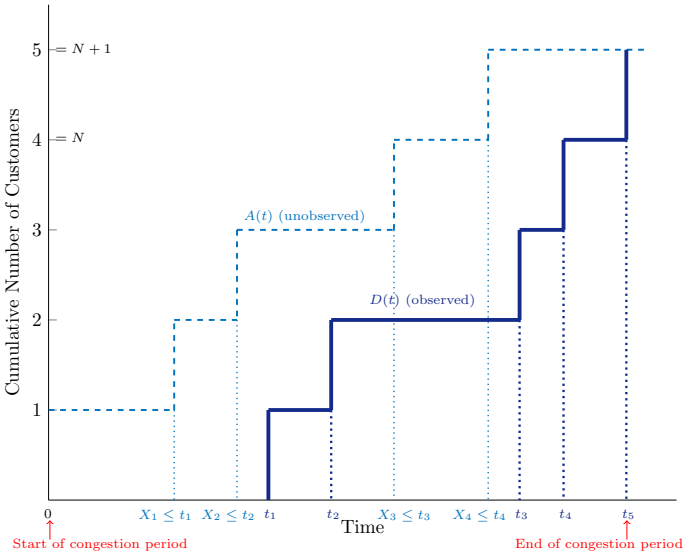

Fig. 6 Trajectories of $D(t)$ (observed) and $A(t)$ (unobserved) during a congestion period with $N = 5$ customers

extended the result to the case of Erlang-k interarrival times. They also considered the cases of having finite buffers and the real-time estimation problem where the arrival rate is known. Further, in [43], they considered systems where arriving customers can balk when the queue length is beyond a given threshold and the balking probability is constant. The results of that paper were extended by Jones in [77] where the balking probability increases to 1 when the queue length approaches ∞.

Meanwhile, Jones, and Larson [79] suggested additional algorithms with $O(n^3)$ computational complexity, improving the algorithms of [88]. In [44], Daley and Servi presented an additional algorithm by omitting queue lengths with low probabilities. Later on, in [45], the result was extended to customer balking and reneging. In [47], Dimitrijevic developed further algorithms which under special cases have complexity as low as $O(n^2)$. Moving onto queueing networks, in [90], Mandelbaum and Zeltyn applied the queue inference engine idea. This is one of the few papers in this survey that deal with complex queueing networks (as opposed to single pass queueing systems).

In all of these papers, calculating the likelihood of a congestion period is the most difficult task. This difficulty is due to the fact that the likelihood should be integrated over the all realisations of the unobserved arrival process and the number of terms in this sum increases exponentially with the number of departures. In [54], Fearnhead considered applying a likelihood recursion to test the likelihood efficiency of the estimator when used in the M/G/1 and E_k/G/1 cases. Here, the only observed variables were the interdeparture times.

In [56], Frey and Kaplan considered the case of periodic reporting data, where the arrivals follow a Poisson process with period-specific arrival rates and the data are the number of departures during each period. However, the results of this paper were challenged by Jones in [78] where he showed that queue inference cannot be carried

out without knowing service start or stop times. Further, in [78] Jones presented an extension of the analysis.

In [101], Park et al. presented a new complementary variant of the QIE problem. They considered the case where the number of servers is unknown and exact inferences about queueing and service times come from the arrival and departure times. Then, in [71], Heckmüller and Wolfinger considered the case of the G/D/1 queue for inference about characteristics of the arrival process from transactional data.

3.7 Bayesian approaches

A Bayesian approach to inference treats unknown parameters as random variables and the inference procedure is a process of refining the distribution of these parameters. We begin with a prior distribution on the parameter values, and once data are collected, Bayes' formula yields a posterior distribution. This posterior distribution and functionals of it are the main outcome of the inference.

In setting up a Bayesian estimation problem, prior distributions are often parameterized themselves by hyper-parameters, and in certain cases the resulting posterior distribution also has a parametric form. Such cases arise when the prior is a conjugate for the likelihood model, that is, when the distribution of the posterior has the same parametric form as the distribution of the prior. When there is not a nice parametric form for the posterior, computational methods are required, especially when the posterior distribution is high-dimensional with an intractable normalization constant. Common methods include Monte Carlo Markov chains (MCMC), as well as many modifications and adaptations such as approximate Bayesian computation (ABC) which is used in case of an intractable likelihood. See, for example, [28] for more details about Bayesian inference at large.

The Bayesian approach usually applies to queueing systems where, in addition to inference, prediction is of interest. Example applications include internet traffic analysis and risk theory. Let us return to a very elementary example of the M/D/1 queue and as with previous examples assume that the service duration m is known and the unknown arrival rate λ is the parameter of interest. Consider now the (F) observation scheme, see Table 1, where a full queue trajectory is observed, however, it is only the arrival process which is of interest. Now, the data collection is over n periods, each of duration T, and the data is a sequence $\{x_1, \ldots, x_n\}$ where x_i is the number of arrivals in period i. In this case, the likelihood model is that the distribution of the number of arrivals during a period is Poisson with parameter λT.

A common choice that works well with the Poisson likelihood is a Gamma prior with hyper-parameters α and β as shape and rate parameters, respectively. In this case, it is an elementary Bayesian calculation to show that gamma distribution is a conjugate prior.

To see this note that the posterior is proportional to the product of the likelihood and the prior, and hence,

$$
\begin{aligned}
f(\lambda \mid \text{data}) &\propto \left(\prod_{i=1}^{n} e^{-\lambda T} \frac{(\lambda T)^{x_i}}{x_i!} \right) \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \\
&\propto e^{-n\lambda T} \ (\lambda T)^{\sum_{i=1}^{n} x_i} \ \lambda^{\alpha-1} \ e^{-\beta\lambda} \\
&\propto e^{-n\lambda T} \ \lambda^{\sum_{i=1}^{n} x_i} \ \lambda^{\alpha-1} \ e^{-\beta\lambda} \\
&= \lambda^{\alpha+(\sum_{i=1}^{n} x_i)-1} \, e^{-\lambda(\beta+nT)},
\end{aligned}
\tag{30}
$$

which is proportional to the density function of a gamma distribution with shape parameter $\alpha + \sum x_i$ and with rate parameter $\beta + nT$.

This example is of course just performing inference for the rate of the Poisson arrival process although it is posed here in the context of an M/D/1 queue. A similar line of reasoning has been employed in a significant body of the literature dealing with Bayesian inference for queues. For example, see early work of Armero et al. [4–7] for obtaining the posterior distribution of the traffic intensity, waiting time, number of customers, and length of idle and busy period for an M/M/1 queueing system. Later similar work, focusing on specific phase-type service times, can be found in [15] and [76]. See also [105], where additional specific distributions amenable for efficient Bayesian inference are employed and [93,94] where considerations of the subjective Bayesian paradigm for queueing inference are discussed.

An alternative computational line of research includes [119] by Sutton and Jordan where Bayesian inference for general queueing networks and service mechanisms was studied. Here, the queue is generally viewed as a transformation mechanism between exogenous processes and endogenous processes (although the authors did not use this terminology). They considered a variety of mechanisms and policies and presented an overview of the application of Bayesian inference for queueing networks where simulation of the queueing processes is part of the posterior procedures to sample latent variables. The computational procedures make use of the slice sampler, [97]. The computational paradigms introduced in [119] have influenced several other works in the computer science and Bayesian statistics communities, such as [125] where closed queueing networks were considered.

The Bayesian paradigm also extends to empirical Bayesian approaches as discussed for queues in [121] and to more recent work dealing with nonparametric Bayesian approaches. A notable paper in this direction is [40] focusing on discrete time queues where the inference is for the service time distribution.

3.8 Online prediction

In this paradigm, we observe some of the endogenous processes up to a given time and make prediction about future values. A common application is *delay prediction* where the queue length process or workload process is observed and used for predicting the waiting time of arriving customers. Using delay predictions to make *delay announcements* is common in call centres and other service operations. In certain cases, some of the model parameters are known.

Most of the literature considers the mean square error (MSE) criterion, under which the best predictor is the expected value. As an example, consider a GI/G/1 system with known arrival rate λ and known mean service time m. Assume that, at time t_0, we observe $Q(t_0) = q_0 > 0$. We may then require predictors for the waiting time of a customer arriving at time t_0 (not included in the count q_0) or for functions of the future queue length, $f\big(Q(t_0+u)\big)$. Such predictors also include the waiting time of future customers that arrive at $t_0 + u$.

Under a FCFS policy, predicting the waiting time of a customer arriving to find q_0 customers already in the system is straightforward. The expected service time of each of those yet to commence service is m, and hence, the expected delay is,

$$(q_0 - 1)m + R,$$

where R is the residual service time of the customer currently in service. The value of R may either be observed, or estimated. In a case such as the GI/M/1 queue, the expected delay is simply $q_0 m$ due to the memory-less property of the exponential distribution. Further, it is straightforward to provide quantiles or other measures of the delay time, as the waiting time of the customer is Erlang (Gamma) distributed.

If we are looking for predictors for $f\big(Q(t_0+u)\big)$, then explicit expressions require more stringent assumptions. For example, in an M/M/1 queue, $Q(t_0)$ describes the full state information and the predictor that minimizes the MSE is

$$\widehat{f}(Q(t_0+u)) = \mathbb{E}[f\big(Q(u)\big) \mid Q(0) = q_0] = \sum_{j=0}^{\infty} f(j)\, p_{q_0 j}(u),$$

where $p_{ij}(u)$ is the transition probability of a birth–death continuous time Markov chain, from state i to state j in u time steps. In the case of M/M/1, expressions involving Bessel functions for $p_{ij}(\cdot)$ are known, [38], and hence, in principle, closed-form predictors can be computed. However, in more general models, predictors for $f\big(Q(t_0+u)\big)$ quickly become intractable and hence approximations are involved.

The classic literature dealing with such cases includes [118] where transition probabilities for the GI/M/1 embedded Markov chain are used, [129] where extensions to multi-server GI/M/c queues are considered, and [100] where predictors associated with the M/G/1 queue are considered. In these papers, explicit transition probabilities of certain endogenous processes were used along with the embedded Markov chain structure of GI/M/c- and M/G/1-type queues. In general, there does not appear to be a mechanism for generalizing this type of analysis beyond GI/M/c and M/G/1. That is, systems such as M/G/c or GI/G/1 queues or more complex systems require a different set of tools.

For more general settings, one can consider approximations. A general entry point focused on operations management of call centres is [126] deriving predictors for the waiting time of customers currently in the system. The analysis focuses on multi-server systems with multiple customer classes. Further work was carried out in [75] where the realistic scenario of time-varying demand and time-varying service capacity was considered. In such cases, fluid approximations were employed to derive several types of predictors. See also [120].

3.9 Implicit models

In many of the paradigms described above, queueing models are explicitly used to model real-life situations and data are used for parameter estimation or state prediction. However, queueing models may also be used implicitly, without requiring an "exact fit" between model and reality. Towards that end, several different research directions have been pursued. One direction is the application of *information mining* based on *event logs* for creating queueing models directly from the data. Another direction is *grey-box modelling* where queueing-like processes are used to describe the data, without requiring an exact fit.

Information mining The general idea here is to use an extensive event log dataset to dynamically create queueing models that describe the underlying processes. This is quite different from classical modelling where the modeller observes the process and suggests a mathematical model. This idea has been explored in a series of recent papers. In [115] Sendrovich, Weidlich, Gal, and Mandelbaum use the developed field of *business process mining* based on event logs, see [123], for queues. They adapted ideas from this field to queues and developed the method of *queue mining*. In [113], the work was extended to handle the queue mining paradigm in view of partial information. In [114], customers with different priorities were incorporated as part of the queue mining process. Further, in [116] a resource-driven perspective was employed with an application to an outpatient clinic.

Grey-box models There are several classes of stochastic processes that are often used in explicit queueing models. These include birth and death processes and other structured Markov chains. One may develop *statistical queueing models* which are based on similar underlying processes but do not attempt to utilize a mechanistic relationship between models for the exogenous and endogenous processes. As an analogy, consider time-series models where common stochastic processes, such as autoregressive integrated moving average (ARIMA) models, are used without an explicit description of how the underlying random variables are related to the physical world. This idea can be used with birth and death processes or with any other queueing-based stochastic process in the hope that the queueing-like stochastic process can model queueing phenomena well.

In [49], Dong and Whitt considered a stationary birth and death process fitted to a sample path of an arbitrary queueing system. General birth and death parameters were allowed. This differs from an explicit queueing model such as M/M/c where the birth rate is assumed to be constant λ and the death rate at level k is μk for $k \leq c$ and μc for $k > c$. In the latter scenario, an exact queueing model could be fitted to estimate the parameters (parameters are λ, μ, and c), whereas in the grey-box approach of Dong and Whitt, an arbitrary birth and death process allows us to compensate for potential model misspecification. A similar approach was applied to health-care data in [14]. Another grey-box-type paper is [130] where queueing networks are approximately fitted to network data.

The fitting of birth and death processes is also interesting in its own right and, as shown in [128], different fitting methods are possible. See also [50] dealing with time-varying periodic queues.

3.10 Control, design, and uncertainty quantification

Most of the paradigms surveyed above deal with parameter or state estimation. However, related problems deal with how to control queueing systems in the presence of uncertainty, how to design such systems, and how to deal with uncertainty quantification when carrying out such control or design. There is an extensive literature for control, design, and architecture selection for queueing systems. However, the literature mainly focuses on cases which assume that the probability laws of arrival and service processes are precisely known and the state of the system is fully observable.

In the realm of stochastic control, there are two general paradigms for dealing with such uncertainty. In one paradigm, a controller wishes to optimally control a system in which parameters are not known. This is sometimes called adaptive control. The field of reinforcement learning suggests a variety of methods for dealing with such a setting. An alternative case is that in which the state observation is not fully available. The field of partially observable Markov decision processes (POMDPs) deals with this setting. To the best of our knowledge, in the specific context of queueing control, both of these areas have not received extensive attention.

An early paper dealing with adaptive control of queues was [74] where, for an M/G/1 queueing system with an unknown arrival rate and an average cost criterion, the controller chooses the service rate to minimize long term costs. In [87], the celebrated $c\mu_i$ scheduling rule was analysed in the case where the service rates μ_i are estimated online. In this case, a regret-based analysis was performed. As for POMDPs, some recent work was presented in [10] where the interaction between partially observable queues and stability was explored. Also related is [8] where supply systems were considered and the effect of not being aware of duplicate orders is analysed. Beyond these adaptive control and POMDP papers, we are not aware of further significant work.

In addition to control, there is the problem of how to design queueing systems. This often refers to offline specification of quantities such as the number of servers, server rate allocation, and the queueing discipline. In contrast, control of a system typically considers the problem of online decision making based on state measurements or estimates. An interesting aspect dealing with design arises when parameter uncertainty is present. As an illustration of the trade-offs inherently involved, in [48] the authors considered a single-pass loss-less queueing system in steady state with an unknown arrival rate. They analysed several trade-offs dealing with architecture selection for such systems.

In general, design of queueing systems is often based on performance analysis which includes computing functionals of the endogenous processes. In recent years, there has been much work on the robust evaluation of such performance measures. In this setting, parameters are assumed to not be known exactly, but rather to lie within specified uncertainty sets. See, for example, [17] and references within.

4 Conclusion

The queuing theory literature spans multiple journals, dozens of books, and thousands of publications. However, within that, the literature dealing with parameter and state

estimation is much more limited. We have done our best to list this comprehensively in the annotated bibliography [9]. Our purpose in this survey is to present an up to date account of this more narrow aspect of queueing research. While our discussion is not exhaustive, we have attempted to present a comprehensive view of the estimation paradigms that have been investigated to date.

When attempting to classify a body of work, one approach is to consider the several dimensions that specify the problems at hand. As outlined in Sect. 2 for parameter and state estimation in queues, these dimensions include the inference activity, the models, the observation scheme, and the statistical methods and principles. We have attempted to describe the field using this viewpoint and the ten major estimation paradigms that appear in Table 2 and are surveyed in the subsections of Sect. 3.

In considering the estimation paradigms outlined in Sect. 2.1, we believe that there is room for extensive further research that will connect some of the paradigms. Specifically, the joint application of inverse problem methods, Bayesian approaches, implicit models, and control of systems may be of interest. The past decade has witnessed an explosive growth in data-driven statistical learning applications. Some of the application areas that have benefited from this growth include systems of congestion, resource scarcity, and queues. It remains a challenge to connect queueing estimation paradigms with modern machine learning applications and methods.

Acknowledgements Azam Asanjarani's and Peter Taylor's research is supported by the Australian Research Council (ARC) Centre of Excellence for the Mathematical and Statistical Frontiers (ACEMS). Yoni Nazarathy is supported by ARC grant DP180101602. We are grateful to Liron Ravner for feedback on an early version of the manuscript. We also thank Ross McVinish and an anonymous referee for useful feedback. We thank Phil Pollett for contributions to an early version of the associated annotated bibliography [9].

References

1. Alouf, S., Nain, P., Towsley, D.: Inferring network characteristics via moment-based estimators. In: INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, vol. 2, pp. 1045–1054. IEEE (2001)
2. Antunes, N., Jacinto, G., Pacheco, A.: Probing a M/G/1 queue with general input and service times. ACM SIGMETRICS Perform Eval Rev **41**(3), 34–36 (2014)
3. Antunes, N., Jacinto, G., Pacheco, A., Wichelhaus, C.: Estimation of the traffic intensity in a piecewise-stationary $M_t/G_t/1$ queue with probing. ACM SIGMETRICS Perform Eval Rev **44**(2), 3–5 (2016)
4. Armero, C.: Bayesian analysis of M/M/1/∞/FIFO queues. Bayesian Stat. **2**, 613–618 (1985)
5. Armero, C.: Bayesian inference in Markovian queues. Queueing Syst. **15**(1), 419–426 (1994)
6. Armero, C., Bayarri, M.J.: Bayesian prediction in M/M/1 queues. Queueing Syst. **15**(1), 401–417 (1994)
7. Armero, C., Bayarri, M.J.: Prior assessments for prediction in queues. Stat. **43**(1), 139–153 (1994)
8. Armony, M., Plambeck, E.L.: The impact of duplicate orders on demand estimation and capacity investment. Manag. Sci. **51**(10), 1505–1518 (2005)
9. Asanjarani, A., Nazarathy, Y.: Parameter and State Estimation in Queues and Related Stochastic Models: A Bibliography. arXiv:1701.08338 (2020)
10. Asanjarani, A., Nazarathy, Y.: The role of information in system stability with partially observable servers. Methodol. Comput. Appl. Probab. **22**, 949–968 (2020)
11. Asanjarani, A., Nazarathy, Y., Taylor, G.P.: Queueing Estimation Survey. https://github.com/yoninazarathy/QueueingEstimationSurvey (2020)

12. Asmussen, S.: Applied Probability and Queues (Stochastic Modeling and Applied Probability Series), vol. 51. Springer, Berlin (2010)
13. Asmussen, S., Nerman, O., Olsson, M.: Fitting phase-type distributions via the EM algorithm. Scand. J. Stat. **23**(4), 419–441 (1996)
14. Au, L., Byrnes, G.B., Bain, C.A., Fackrell, M., Brand, C., Campbell, D.A., Taylor, P.G.: Predicting overflow in an emergency department. IMA J. Manag. Math. **20**(1), 39–49 (2009)
15. Ausín, M.C., Wiper, M.P., Lillo, R.E.: Bayesian estimation for the M/G/1 queue using a phase-type approximation. J. Stat. Plan. Inference **118**(1–2), 83–101 (2004)
16. Baccelli, F., Kauffmann, B., Veitch, D.: Inverse problems in queueing theory and internet probing. Queueing Syst. **63**(1), 59–107 (2009)
17. Bandi, C., Bertsimas, D., Youssef, N.: Robust queueing theory. Oper. Res. **63**(3), 676–700 (2015)
18. Basawa, I.V., Bhat, U.N., Lund, R.: Maximum likelihood estimation for single server queues from waiting time data. Queueing Syst. **24**(1), 155–167 (1996)
19. Basawa, I.V., Bhat, U.N., Zhou, J.: Parameter estimation using partial information with applications to queueing and related models. Stat. Prob. Lett. **78**(12), 1375–1383 (2008)
20. Basawa, I.V., Prabhu, N.U.: Large sample inference from single server queues. Queueing Syst. **3**(4), 289–304 (1988)
21. Bertsimas, D.J., Servi, L.D.: Deducing queueing from transactional data: the queue inference engine, revisited. Oper. Res. **40**, S217–S228 (1992)
22. Bhat, U.N., Miller, G.K., Rao, S.S.: Statistical analysis of queueing systems. In: Dshalalow, J.H. (ed) Chapter in: Frontiers in Queueing, pp. 351–394 (1997)
23. Bhat, U.N., Rao, S.S.: Statistical analysis of queueing systems. Queueing Syst. **1**(3), 217–247 (1987)
24. Bingham, N.H., Pitts, S.M.: Non-parametric estimation for the M/G/∞ queue. Ann. Inst. Stat. Math. **51**(1), 71–97 (1999)
25. Bingham, N.H., Pitts, S.M.: Nonparametric inference from M/G/l busy periods. Stoch. Models **15**(2), 247–272 (1999)
26. Bladt, M., Sørensen, M.: Statistical inference for discretely observed Markov jump processes. J. R. Stat. Soc. Ser. B (Statistical Methodology) **67**(3), 395–410 (2005)
27. Blanghaps, N., Nov, Y., Weiss, G.: Sojourn time estimation in an M/G/∞ queue with partial information. J. Appl. Prob. **50**(4), 1044–1056 (2013)
28. Bolstad, W.M., Curran, J.M.: Introduction to Bayesian statistics. Wiley, Hoboken (2016)
29. Boxma, O.J., Vlasiou, M.: On queues with service and interarrival times depending on waiting times. Queueing Syst. **56**(3–4), 121–132 (2007)
30. Bramson, M.: Stability of Queueing Networks. Springer, Berlin (2008)
31. Brillinger, D.R.: Cross-spectral analysis of processes with stationary increments including the stationary G/G/∞ queue. Ann. Prob. **2**(5), 815–827 (1974)
32. Brockwell, P.J., Davis, R.A., Fienberg, S.E.: Time Series: Theory and Methods. Springer, Berlin (1991)
33. Brown, M.: An M/G/∞ estimation problem. Ann. Math. Stat. **41**(2), 651–654 (1970)
34. Carson, J.S., Law, A.M.: Conservation equations and variance reduction in queueing simulations. Oper. Res. **28**(3–part–i), 535–546 (1980)
35. Chen, H., Yao, D.D.: Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization (stochastic Modelling and Applied Probability Series), vol. 46. Springer, Berlin (2013)
36. Chen, T.M., Walrand, J., Messerschmitt, D.G.: Parameter estimation for partially observed queues. IEEE Trans. Commun. **42**(9), 2730–2739 (1994)
37. Clarke, A.B.: Maximum likelihood estimates in a simple queue. Ann. Math. Stat. **28**(4), 1036–1040 (1957)
38. Cohen, J.W.: The Single Server Queue (Applied Mathematics and Mechanics), vol. 8. North-Holland, Amsterdam (1982)
39. Comert, G., Cetin, M.: Queue length estimation from probe vehicle location and the impacts of sample size. Eur. J. Oper. Res. **197**(1), 196–202 (2009)
40. Conti, P.L.: Large sample Bayesian analysis for Geo/G/1 discrete-time queueing models. Ann. Stat. **27**(6), 1785–1807 (1999)
41. Cox, D.R.: Some problems of statistical analysis connected via congestion. In: Proceedings of Symposium on Congestion Theory, pp. 289–316 (1965)
42. Daley, D.J., Servi, L.D.: Exploiting Markov chains to infer queue length from transactional data. J. Appl. Prob. **29**(3), 713–732 (1992)

43. Daley, D.J., Servi, L.D.: A two-point Markov chain boundary-value problem. Adv. Appl. Prob. **25**(3), 607–630 (1993)
44. Daley, D.J., Servi, L.D.: Estimating waiting times from transactional data. INFORMS J. Comput. **9**(2), 224–229 (1997)
45. Daley, D.J., Servi, L.D.: Moment estimation of customer loss rates from transactional data. J. Appl. Math. Stoch. Anal. **11**(3), 301–310 (1998)
46. den Boer, A.V., Mandjes, M.: Convergence rates of Laplace-transform based estimators. Bernoulli **23**(4A), 2533–2557 (2017)
47. Dimitrijevic, D.D.: Inferring most likely queue length from transactional data. Oper. Res. Lett. **19**(4), 191–199 (1996)
48. Dinh, T.V., Andrew, L.L.H., Nazarathy, Y.: Architecture and robustness tradeoffs in speed-scaled queues with application to energy management. Int. J. Syst. Sci. **45**(8), 1728–1739 (2014)
49. Dong, J., Whitt, W.: Stochastic grey-box modeling of queueing systems: fitting birth-and-death processes to data. Queueing Syst. **79**(3–4), 391–426 (2015)
50. Dong, J., Whitt, W.: Using a birth-and-death process to estimate the steady-state distribution of a periodic queue. Nav. Res. Logist. (NRL) **62**(8), 664–685 (2015)
51. Duffie, D., Glynn, P.: Estimation of continuous-time Markov processes sampled at random time intervals. Econometrica **72**(6), 1773–1808 (2004)
52. Duffy, K.R., Meyn, S.P.: Estimating Loynes' exponent. Queueing Syst. **68**(3–4), 285–293 (2011)
53. Edelmann, D., Wichelhaus, C.: Nonparametric inference for queueing networks of $\mathrm{Geom}_X/\mathrm{G}/\infty$ queues in discrete time. Adv. Appl. Prob. **46**(3), 790–811 (2014)
54. Fearnhead, P.: Filtering recursions for calculating likelihoods for queues based on inter-departure time data. Stat. Comput. **14**(3), 261–266 (2004)
55. Foss, S., Konstantopoulos, T.: An overview of some stochastic stability methods (special issue: network design, control and optimization). J. Oper. Res. Soc. Jpn. **47**(4), 275–303 (2004)
56. Frey, J.C., Kaplan, E.H.: Queue inference from periodic reporting data. Oper. Res. Lett. **38**(5), 420–426 (2010)
57. Glynn, P.W., Melamed, B., Whitt, W.: Estimating customer and time averages. Oper. Res. **41**(2), 400–408 (1993)
58. Glynn, P.W., Torres, M.: Parametric estimation of tail probabilities for the single-server queue. In: Dshalalow, J.H. (ed.) Chapter in Frontiers in Queueing: Models and Applications in Science and Engineering, pp. 449–462 (1997)
59. Glynn, P.W., Whitt, W.: A central-limit-theorem version of L= λW. Queueing Syst. **1**(2), 191–215 (1986)
60. Glynn, P.W., Whitt, W.: Indirect estimation via $L = \lambda W$. Oper. Res. **37**(1), 82–103 (1989)
61. Glynn, P.W., Zeevi, A.J.: Estimating tail probabilities in queues via extremal statistics. Analysis of communication networks: call centres, traffic and performance. Fields Inst. Commun. **28**, 135–158 (2000)
62. Goldenschluger, A.: Nonparametric estimation of the service time distribution in the $\mathrm{M}/\mathrm{G}/\infty$ queue. Adv. Appl. Prob. **48**(4), 1117–1138 (2016)
63. Goldenshluger, A.: The $\mathrm{M}/\mathrm{G}/\infty$ estimation problem revisited. Bernoulli **24**(4A), 2531–2568 (2018)
64. Goldenshluger, A., Koops, D.T.: Nonparametric estimation of service time characteristics in infinite-server queues with nonstationary Poisson input. Stoch. Syst. **9**(3), 183–207 (2019)
65. Grübel, R.: Stochastic models as functionals: some remarks on the renewal case. J. Appl. Prob. **26**(2), 296–303 (1989)
66. Grübel, R., Pitts, S.M.: A functional approach to the stationary waiting time and idle period distributions of the GI/G/1 queue. Ann. Prob. **20**(4), 1754–1778 (1992)
67. Grübel, R., Wegener, H.: Matchmaking and testing for exponentiality in the $\mathrm{M}/\mathrm{G}/\infty$ queue. J. Appl. Prob. **48**(1), 131–144 (2011)
68. Hall, P., Park, J.: Nonparametric inference about service time distribution from indirect measurements. J. R. Stat. Soc. Ser. B (Statistical Methodology) **66**(4), 861–875 (2004)
69. Hansen, M.B., Pitts, S.M.: Nonparametric inference from the M/G/1 workload. Bernoulli **12**(4), 737–759 (2006)
70. Heckmuller, S., Wolfinger, B.E.: Reconstructing arrival processes to G/D/1 queueing systems and tandem networks. In: International Symposium on Performance Evaluation of Computer & Telecommunication Systems, 2009. SPECTS 2009, vol. 41, pp. 361–368. IEEE (2009)

71. Heckmüller, S., Wolfinger, B.E.: Reconstructing arrival processes to discrete queueing systems by inverse load transformation. Simulation **87**(12), 1033–1047 (2011)
72. Hei, X., Bensaou, B., Tsang, D.H.K.: A light-weight available bandwidth inference methodology in a queueing analysis approach. In: 2005 IEEE International Conference on Communications, 2005. ICC 2005, vol. 1, pp. 120–124. IEEE (2005)
73. Hei, X., Bensaou, B., Tsang, D.H.K.: Model-based end-to-end available bandwidth inference using queueing analysis. Comput. Netw. **50**(12), 1916–1937 (2006)
74. Hernandez-Lerma, O., Marcus, S.I.: Adaptive control of service in queueing systems. Syst. Control Lett. **3**(5), 283–289 (1983)
75. Ibrahim, R., Whitt, W.: Wait-time predictors for customer service systems with time-varying demand and capacity. Oper. Res. Baltimore **59**(5), 1106–1118 (2011)
76. Insua, D.R., Wiper, M., Ruggeri, F.: Bayesian analysis of M/Er/1 and M/H_k/1 queues. Queueing Syst. **30**(3), 289–308 (1998)
77. Jones, L.K.: Inferring balking behavior from transactional data. Oper. Res. **47**(5), 778–784 (1999)
78. Jones, L.K.: Remarks on queue inference from departure data alone and the importance of the queue inference engine. Oper. Res. Lett. **40**(6), 503–505 (2012)
79. Jones, L.K., Larson, R.C.: Efficient computation of probabilities of events described by order statistics and applications to queue inference. ORSA J. Comput. **7**(1), 89–100 (1995)
80. Kauffmann, B.: Inverse problems in bandwidth sharing networks. In: Proceedings of the 24th International Teletraffic Congress, p. 6. International Teletraffic Congress (2012)
81. Ke, J.C., Chu, Y.K.: Comparison on five estimation approaches of intensity for a queueing system with short run. Comput. Stat. **24**(4), 567–582 (2009)
82. Kelly, F.P.: Reversibility and Stochastic Networks. Cambridge University Press, Cambridge (2011)
83. Kiessler, P.C., Lund, R.: Technical note: traffic intensity estimation. Nav. Res. Logist. (NRL) **56**(4), 385–387 (2009)
84. Kim, S.H., Whitt, W.: Estimating waiting times with the time-varying Little's law. Prob. Eng. Inform. Sci. **27**(4), 471–506 (2013)
85. Kim, S.H., Whitt, W.: Statistical analysis with Little's law. Oper. Res. **61**(4), 1030–1045 (2013)
86. Kim, S.H., Whitt, W., Cha, W.C.: A data-driven model of an appointment-generated arrival process at an outpatient clinic. INFORMS J. Comput. **30**(1), 181–199 (2018)
87. Krishnasamy, S., Arapostathis, A., Johari, R., Shakkottai, S.: On learning the cμ rule in single and parallel server networks. In: 2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 153–154. IEEE (2018)
88. Larson, R.C.: The queue inference engine: deducing queue statistics from transactional data. Manag. Sci. **36**(5), 586–601 (1990)
89. Latouche, G., Ramaswami, V., Sethuraman, J., Sigman, K., Squillante, M.S., Yao, D.: Matrix-Analytic Methods in Stochastic Models (Springer Proceedings in Mathematics Statistics), vol. 27. Springer, Berlin (2012)
90. Mandelbaum, A., Zeltyn, S.: Estimating characteristics of queueing networks using transactional data. Queueing Syst. **29**(1), 75–127 (1998)
91. Mandjes, M., Patch, B., Walton, N.S.: Detecting Markov chain instability: a Monte Carlo approach. Stoch. Syst. **7**(2), 289–314 (2017)
92. Mandjes, M., Ravner, L.: Hypothesis testing for a Lévy-driven storage system by Poisson sampling. Stoch. Process. Appl. **133**, 41–73 (2020)
93. Mcgrath, M.F., Gross, D., Singpurwalla, N.D.: A subjective Bayesian approach to the theory of queues I—modeling. Queueing Syst. **1**(4), 317–333 (1987)
94. McGrath, M.F., Singpurwalla, N.D.: A subjective Bayesian approach to the theory of queues II—inference and information in M/M/1 queues. Queueing Syst. **1**(4), 335–353 (1987)
95. McVinish, R., Pollett, P.K.: Constructing estimating equations from queue length data. Preprint at https://people.smp.uq.edu.au/PhilipPollett/papers/2011dMcVinishPollett.pdf (2011)
96. Nam, S.Y., Kim, S., Sung, D.K.: Estimation of available bandwidth for an M/G/1 queueing system. Appl. Math. Model. **33**(8), 3299–3308 (2009)
97. Neal, R.M.: Slice sampling. Ann. Stat. **31**(3), 705–741 (2003)
98. Novak, A., Watson, R.: Determining an adequate probe separation for estimating the arrival rate in an M/D/1 queue using single-packet probing. Queueing Syst. **61**(4), 255–272 (2009)
99. Nozari, A., Whitt, W.: Estimating average production intervals using inventory measurements: Little's law for partially observable processes. Oper. Res. **36**(2), 308–323 (1988)

100. Pagurek, B., Stanford, D.A., Woodside, C.M.: Optimal prediction of times and queue lengths in the M/G/1 queue. J. Oper. Res. Soc. **39**(6), 585–593 (1988)
101. Park, J., Kim, Y.B., Willemain, T.R.: Analysis of an unobservable queue using arrival and departure times. Comput. Ind. Eng. **61**(3), 842–847 (2011)
102. Pickands, J., Stine, R.A.: Estimation for an M/G/∞ queue with incomplete information. Biometrika **84**(2), 295–308 (1997)
103. Pin, F., Veitch, D., Kauffmann, B.: Statistical estimation of delays in a multicast tree using accelerated EM. Queueing Syst. **66**(4), 369–412 (2010)
104. Pitts, S.M.: Nonparametric estimation of the stationary waiting time distribution function for the GI/G/1 queue. Ann. Stat. **22**(3), 1428–1446 (1994)
105. Ramirez-Cobo, P., Lillo, R.E., Wilson, S., Wiper, M.P.: Bayesian inference for double Pareto lognormal queues. Ann. Appl. Stat. **4**(3), 1533–1557 (2010)
106. Ravner, L., Boxma, O., Mandjes, M.: Estimating the input of a Lévy-driven queue by Poisson sampling of the workload process. Bernoulli **25**(4B), 3734–3761 (2019)
107. Reynolds, J.F.: The covariance structure of queues and related processes—a survey of recent work. Adv. Appl. Probab. **7**(2), 383–415 (1975)
108. Rodrigo, A., Vazquez, M.: Large sample inference in retrial queues. Math. Comput. Modell. **30**(3–4), 197–206 (1999)
109. Ross, J.V., Taimre, T., Pollett, P.K.: Estimation for queues from queue length data. Queueing Syst. **55**(2), 131–138 (2007)
110. Schruben, L., Kulkarni, R.: Some consequences of estimating parameters for the M/M/1 queue. Oper. Res. Lett. **1**(2), 75–78 (1982)
111. Schweer, S., Wichelhaus, C.: Nonparametric estimation of the service time distribution in the discrete-time GI/G/∞ queue with partial information. Stoch. Process. Appl. **125**(1), 233–253 (2015)
112. Schweer, S., Wichelhaus, C.: Nonparametric estimation of the service time distribution in discrete-time queueing networks. Stoch. Process. Appl. **130**(8), 4643–4666 (2020)
113. Senderovich, A., Leemans, S.J.J., Harel, S., Gal, A., Mandelbaum, A., van der Aalst, W.M.P.: Discovering queues from event logs with varying levels of information. In: International Conference on Business Process Management, pp. 154–166. Springer (2016)
114. Senderovich, A., Weidlich, M., Gal, A., Mandelbaum, A.: Queue mining for delay prediction in multi-class service processes. Inform. Syst. **53**, 278–295 (2019)
115. Senderovich, A., Weidlich, M., Gal, A., Mandelbaum, A.: Queue mining–predicting delays in service processes. In: International Conference on Advanced Information Systems Engineering, pp. 42–57. Springer (2014)
116. Senderovich, A., Weidlich, M., Gal, A., Mandelbaum, A., Kadish, S., Bunnell, C.A.: Discovery and validation of queueing networks in scheduled processes. In: Advanced Information Systems Engineering, p. 417–433. Springer (2015)
117. Sharma, V., Mazumdar, R.: Estimating traffic parameters in queueing systems with local information. Perform. Eval. **32**(3), 217–230 (1998)
118. Stanford, D.A., Pagurek, B., Woodside, C.M.: Optimal prediction of times and queue lengths in the GI/M/1 queue. Oper. Res. **31**(2), 322–337 (1983)
119. Sutton, C., Jordan, M.I.: Bayesian inference for queueing networks and modeling of internet services. Ann. Appl. Stat. **5**(1), 254–282 (2011)
120. Thiongane, M., Chan, W., l'Ecuyer, P.: New history-based delay predictors for service systems. In: 2016 Winter Simulation Conference (WSC), pp. 425–436. IEEE (2016)
121. Thiruvaiyaru, D., Basawa, I.V.: Empirical Bayes estimation for queueing systems and networks. Queueing Syst. **11**(3), 179–202 (1992)
122. Thiruvaiyaru, D., Basawa, I.V., Bhat, U.N.: Estimation for a class of simple queueing networks. Queueing Syst. **9**(3), 301–312 (1991)
123. Van der Aalst, W., Weijters, T., Maruster, L.: Workflow mining: discovering process models from event logs. IEEE Trans. Knowl. Data Eng. **16**(9), 1128–1142 (2004)
124. Wang, W., Casale, G.: Maximum likelihood estimation of closed queueing network demands from queue length data. ACM SIGMETRICS Perform. Eval. Rev. **43**(2), 45–47 (2015)
125. Wang, W., Casale, G., Kattepur, A., Nambiar, M.: Maximum likelihood estimation of closed queueing network demands from queue length data. In: Proceedings of the 7th ACM/SPEC on International Conference on Performance Engineering, pp. 3–14 (2016)
126. Whitt, W.: Predicting queueing delays. Manag. Sci. **45**(6), 870–888 (1999)

127. Whitt, W.: Stochastic-Process Limits: An Introduction to Stochastic-process Limits and their Application to Queues. Springer, Berlin (2002)
128. Whitt, W.: Fitting birth-and-death queueing models to data. Stat. Prob. Lett. **82**, 998–1004 (2012)
129. Woodside, C.M., Stanford, D.A., Pagurek, B.: Optimal prediction of queue lengths and delays in G_i/M/M multiserver queues. Oper. Res. **32**(4), 809–817 (1984)
130. Zhang, L., Xia, C.H., Squillante, M.S., Mills, W.N.: Workload service requirements analysis: A queueing network optimization approach. In: 10th IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunications Systems, 2002. MASCOTS 2002. Proceedings, pp. 23–32. IEEE (2002)
131. Zheng, S., Seila, A.F.: Some well-behaved estimators for the M/M/1 queue. Oper. Res. Lett. **26**(5), 231–235 (2000)