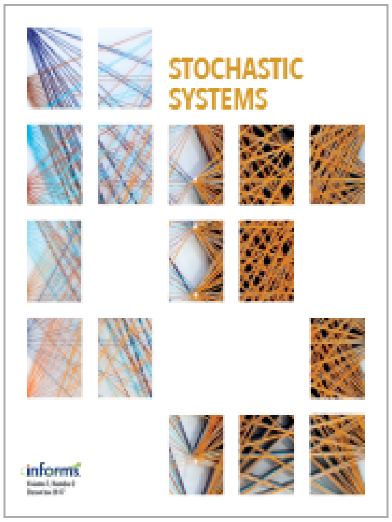


Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Big Jobs Arrive Early: From Critical Queues to Random Graphs

Gianmarco Bet, Remco van der Hofstad, Johan S. H. van Leeuwaarden

informs
http://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Vol. 10, No. 4, December 2020, pp. 310–334
ISSN 1946-5238 (online)

Big Jobs Arrive Early: From Critical Queues to Random Graphs

Gianmarco Bet,[a] Remco van der Hofstad,[b] Johan S. H. van Leeuwaarden[c]

[a] Dipartimento di Matematica e Informatica "Ulisse Dini," Università degli Studi di Firenze, 50134 Firenze, Italy; [b] Department of Mathematics and Computer Science, Eindhoven University of Technology, 5600 MB Eindhoven, Netherlands; [c] Department of Econometrics and Operations Research, Tilburg University, 5037 AB Tilburg, Netherlands

Contact: gianmarco.bet@unifi.it, https://orcid.org/0000-0001-8431-0636 (GB); r.w.v.d.hofstad@tue.nl, https://orcid.org/0000-0003-1331-9697 (RvdH); j.s.h.vanleeuwaarden@uvt.nl (JSHvL)

Received: April 11, 2017
Revised: February 2, 2020
Accepted: February 2, 2020
Published Online in Articles in Advance: October 6, 2020

https://doi.org/10.1287/stsy.2019.0057

Abstract. We consider a queue to which only a finite pool of n customers can arrive, at times depending on their service requirement. A customer with stochastic service requirement S arrives to the queue after an exponentially distributed time with mean $S^{-\alpha}$ for some $\alpha \in [0,1]$; therefore, larger service requirements trigger customers to join earlier. This finite-pool queue interpolates between two previously studied cases: $\alpha = 0$ gives the so-called $\Delta_{(i)}/G/1$ queue and $\alpha = 1$ is closely related to the exploration process for inhomogeneous random graphs. We consider the asymptotic regime in which the pool size n grows to infinity and establish that the scaled queue-length process converges to a diffusion process with a negative quadratic drift. We leverage this asymptotic result to characterize the head start that is needed to create a long period of activity. We also describe how this first busy period of the queue gives rise to a critically connected random forest.

Funding: This work is supported by the NWO Gravitation Networks [Grant 024.002.003]. The work of R. van der Hofstad is supported by the Netherlands Organisation for Scientific Research (NWO) [VICI Grant 639.033.806] and the NWO Gravitation Networks [Grant 024.002.003].

Keywords: queueing theory • transitory queueing systems • heavy-traffic approximation • directed random graphs

1. Introduction

This paper introduces the $\Delta^{\alpha}_{(i)}/G/1$ queue that models a situation in which only a finite pool of n customers will join the queue. These n customers are triggered to join the queue after independent exponential times, but the rates of their exponential clocks depend on their service requirements. When a customer requires S units of service, its exponential clock rings after an exponential time with mean $S^{-\alpha}$ with $\alpha \in [0,1]$. Depending on the value of the free parameter α, the arrival times are independent and identically distributed (i.i.d.) ($\alpha = 0$) or decrease with the service requirement ($\alpha \in (0,1]$). The queue is attended by a single server that starts working at time zero, works at unit speed, and serves the customers in a first-come-first-served manner (i.e., FIFO service discipline). At time zero, we allow for the possibility that i of the n customers have already joined the queue, waiting for service. We will take $i \ll n$, so that without loss of generality we can assume that at time zero there are still n customers waiting for service. These initial customers are numbered $1, \ldots, i$, and the customers that arrive later are numbered $i+1$, $i+2, \ldots$ in order of arrival. Let $A(k)$ denote the number of customers arriving during the service time of the kth customer. The busy periods of this queue will then be completely characterized by the initial number of customers i and the random variables $(A(k))_{k\geq 1}$. The random variables $(A(k))_{k\geq 1}$ are not i.i.d. because of the finite-pool effect and the service-dependent arrival rates. We will model and analyze this queue using the queue-length process embedded at service completions.

We consider the $\Delta^{\alpha}_{(i)}/G/1$ queue in the large-system limit $n \to \infty$ while imposing at the same time a heavy-traffic regime that will stimulate the occurrence of a substantial first busy period. By substantial we mean that the server can work without idling for quite a while, not only serving the initial customers but also those arriving somewhat later. Our main contribution is showing that the embedded queue-length process converges to a Brownian motion with negative quadratic drift. Both the drift coefficient and the variance of the limiting process depend crucially on the value of α. Therefore, from an operational perspective, our result makes clear the dependence of the queue-length process on the highly inhomogeneous arrival process. For the case $\alpha = 0$, referred to as the $\Delta_{(i)}/G/1$ queue with i.i.d. arrivals (Honnappa and Ward 2014, Honnappa et al. 2015), a similar regime was studied in Bet et al. (2018), whereas for $\alpha = 1$ it is closely related to the critical inhomogeneous random graph studied in Bhamidi et al. (2010) and Joseph (2014).

Although the queueing process consists of alternating busy periods and idle periods, in the $\Delta_{(i)}^{\alpha}/G/1$ queue, we naturally focus on the first busy period. After some time, the activity in the queue inevitably becomes negligible. The early phases of the process are therefore of primary interest, when the head start provided by the initial customers still matters and when the rate of newly arriving customers is still relatively high. The head start and strong influx together lead to a substantial first busy period and essentially determine the relevant time of operation of the system.

We also consider the structural properties of the first busy period in terms of a random graph. Let the random variable $H(i)$ denote the number of customers served in the first busy period, starting with i initial customers. We then associate a (directed) random graph to the queueing process as follows. Say $H(i) = N$ and consider a graph with vertex set $\{1, 2, \ldots, N\}$ and in which two vertices r and s are joined by an edge if and only if the r-th customer arrives during the service time of the s-th customer. If $i = 1$, then the graph is a rooted tree with N-labeled vertices, the root being labeled 1. If $i > 1$, then the graph is a forest consisting of i distinct rooted trees whose roots are labeled $1, \ldots, i$, respectively. The total number of vertices in the forest is N.

This random forest is exemplary for a deep relation between queues and random graphs, perhaps best explained by interpreting the embedded $\Delta_{(i)}^{\alpha}/G/1$ queue as an *exploration process*, a generalization of a branching process that can account for dependent random variables $(A(k))_{k\geq 1}$. Exploration processes arose in the context of random graphs as a recursive algorithm to investigate questions concerning the size and structure of the largest components (Aldous 1997). For a given random graph, the exploration process declares vertices active, neutral, or inactive. Initially, only one vertex is active, and all others are neutral. At each time step, one active vertex (e.g., the one with the smallest index) is explored, and it is declared inactive afterward. When one vertex is explored, its neutral neighbors become active for the next time step. As time progresses, and more vertices are already explored (inactive) or discovered (active), fewer vertices are neutral. This phenomenon is known as the *depletion-of-points effect* and plays an important role in the scaling limit of the random graph. Let $A(k)$ denote the neutral neighbors of the kth explored vertex. The exploration process then has increments $(A(k))_{k\geq 1}$ that each have a different distribution. The exploration process encodes useful information about the underlying random graph. For example, excursions above past minima are the sizes of the connected components. The critical behavior of random graphs connected with the emergence of a giant component has received tremendous attention (Bhamidi et al. 2010, 2012, 2014; van der Hofstad et al. 2010, 2016; Addario-Berry et al. 2012; Joseph 2014; Dhara et al. 2016, 2017). Interpreting active vertices as being in a queue, and vertices being explored as customers being served, we see that the exploration process and the (embedded) $\Delta_{(i)}^{\alpha}/G/1$ queue driven by $(A(k))_{k\geq 1}$ are identical.

The analysis of the $\Delta_{(i)}^{\alpha}/G/1$ queue and associated random forest is challenging because the random variables $(A(k))_{k\geq 1}$ are not i.i.d (Bet et al. 2020). In the case of i.i.d. $(A(k))_{k\geq 1}$, there exists an even deeper connection between queues and random graphs, established via branching processes instead of exploration processes (Kendall 1951). To see this, declare the initial customers in the queue to be the zeroth generation. The customers (if any) arriving during the total service time of the initial i customers form the first generation, and the customers (if any) arriving during the total service time of the customers in generation t form generation $t + 1$ for $t \geq 1$. The total progeny of this Galton-Watson branching process has the same distribution as the random variable $H(i)$ in the queueing process. Through this connection, properties of branching processes can be carried over to the queueing processes and associated random graphs (Takács 1988, 1993, 1995; Limic 2001; Duquesne and Le Gall 2005; Le Gall 2005). Takács (1988, 1993, 1995) proved several limit theorems for the case of i.i.d. $(A(k))_{k\geq 1}$, in which case the queue-length process and derivatives such as the first busy period weakly converge to (functionals of) the Brownian excursion process. In that classical line, the present paper can be viewed as an extension to exploration processes with more complicated dependency structures in $(A(k))_{k\geq 1}$.

In Section 2 we describe the $\Delta_{(i)}^{\alpha}/G/1$ queue and associated graphs in more detail and present our main results. The proof of the main theorem, the stochastic-process limit for the queue-length process in the large-pool heavy-traffic regime, is presented in Sections 3 and 4. Section 5 discusses some interesting questions related to the $\Delta_{(i)}^{\alpha}/G/1$ queue and associated random graphs that are left open.

2. Model Description

We consider a sequence of queueing systems, each with a finite (but growing) number n of potential customers labeled with indices $i \in [n] := \{1, \ldots, n\}$. Customers have i.i.d. service requirements with distribution $F_s(\cdot)$. We denote with S_i the service requirement of customer i and with S a generic random value, and S_i and S all have distribution $F_s(\cdot)$. In order to obtain meaningful limits as the system grows large, we scale the service speed by $n/(1+\beta n^{-1/3})$ with $\beta \in \mathbb{R}$ so that the service time of customer i is given by

$$\tilde{S}_i = \frac{S_i\big(1+\beta n^{-1/3}\big)}{n}. \tag{1}$$

We further assume that $\mathbb{E}[S^{2+\alpha}] < \infty$.

If the service requirement of customer i is S_i, then, conditionally on S_i, its arrival time T_i is assumed to be exponentially distributed with mean $1/(\lambda S_i^\alpha)$, with $\alpha \in [0,1]$ and $\lambda > 0$. Hence

$$T_i \overset{\mathrm{d}}{=} \mathrm{E}_i\big(\lambda S_i^\alpha\big), \tag{2}$$

with $\overset{\mathrm{d}}{=}$ denoting equality in distribution and $\mathrm{E}_i(c)$ an exponential random variable with mean $1/c$ independent across i. Conditionally on the service times, the arrival times are independent (but not identically distributed). We introduce $c(1), c(2), \ldots, c(n)$ as the indices of the customers in order of arrival, so that $T_{c(1)} \le T_{c(2)} \le T_{c(3)} \le \ldots$ almost surely.

We will study the queueing system in heavy traffic, in a similar heavy-traffic regime as in Bet et al. (2017, 2019). The initial traffic intensity ρ_n is kept close to one by imposing the relation

$$\rho_n := \lambda_n \mathbb{E}\big[S^{1+\alpha}\big]\big(1+\beta n^{-1/3}\big) = 1 + \beta n^{-1/3} + o_{\mathbb{P}}\big(n^{-1/3}\big), \tag{3}$$

where $\lambda = \lambda_n$ can depend on n, and $f_n = o_{\mathbb{P}}(n^{-1/3})$ is such that $\lim_{n\to\infty} f_n n^{1/3} \xrightarrow{\mathbb{P}} 0$. The traffic intensity is greater than 1 for $\beta > 0$, so that the system is initially overloaded, whereas the system is initially underloaded for $\beta < 0$.

Our main object of study is the queue-length process embedded at service completions, given by $Q_n(0) = i$ and

$$Q_n(k) = (Q_n(k-1) + A_n(k) - 1)^+, \tag{4}$$

with $x^+ = \max\{0, x\}$ and $A_n(k)$ the number of arrivals during the kth service given by

$$A_n(k) = \sum_{i \notin \nu_k} \mathbb{1}_{\{T_i \le \tilde{S}_{c(k)}\}}, \tag{5}$$

where $\nu_k \subseteq [n]$ denotes the set of customers who have been served or are in the queue at the start of the kth service. Note that

$$|\nu_k| = (k-1) + Q_n(k-1) + 1 = k + Q_n(k-1). \tag{6}$$

Given a process $t \mapsto X(t)$, we define its *reflected version* through the reflection map $\phi(\cdot)$ as

$$\phi(X)(t) := X(t) - \inf_{s \le t} X(s)^-. \tag{7}$$

The process $Q_n(\cdot)$ can alternatively be represented as the reflected version of a certain process $N_n(\cdot)$, that is

$$Q_n(k) = \phi(N_n)(k), \tag{8}$$

where $N_n(\cdot)$ is given by $N_n(0) = i$ and

$$N_n(k) = N_n(k-1) + A_n(k) - 1. \tag{9}$$

As a consequence of our assumptions, whenever the server finishes processing one customer, and the queue is empty, the customer to be placed into service is chosen according to the following size-biased distribution:

$$\mathbb{P}\big(\text{customer } j \text{ is placed in service} \mid \nu_{i-1}\big) = \frac{S_j^\alpha}{\sum_{l \notin \nu_{i-1}} S_l^\alpha}, \qquad j \notin \nu_{i-1}, \tag{10}$$

where we tacitly assumed that customer j is the ith customer to be served. With definitions (5) and (10), the process (4) describes the $\Delta^\alpha_{(i)}/G/1$ queue with exponential arrivals (2), embedded at service completions.

2.1. The Scaling Limit of the Embedded Queue

All the processes we consider are elements of the space $\mathcal{D} := \mathcal{D}([0,\infty))$ of càdlàg functions that admit left limits and are continuous from the right. To simplify notation, for a discrete-time process $X(\cdot) : \mathbb{N} \to \mathbb{R}$, we write $X(t)$, with $t \in [0,\infty)$, instead of $X(\lfloor t \rfloor)$. A process defined in this way has càdlàg paths. The space $\mathcal{D}$ is endowed with the usual Skorokhod J_1 topology. We then say that a process converges in distribution in $(\mathcal{D}, J_1)$ when it converges as a random measure on the space $\mathcal{D}$, when this is endowed with the J_1 topology. We are now able to state our main result. Recall that $Q_n(\cdot)$ is the embedded queue-length process of the $\Delta_{(i)}^{\alpha}/G/1$ queue and let

$$\mathbf{Q}_n(t) := n^{-1/3} Q_n\big(tn^{2/3}\big) \tag{11}$$

be the diffusion-scaled queue-length process.

Theorem 1 (Scaling Limit for the $\Delta_{(i)}^{\alpha}/G/1$ Queue)**.** *Assume that $\alpha \in [0,1]$, $\mathbb{E}[S^{2+\alpha}] < \infty$ and that the heavy-traffic condition (3) holds. Assume further that $\mathbf{Q}_n(0) = q$. Then, as $n \to \infty$,*

$$\mathbf{Q}_n(\cdot) \xrightarrow{\mathrm{d}} \phi(W)(\cdot) \qquad \text{in } (\mathcal{D}, J_1), \tag{12}$$

where $W(\cdot)$ is the diffusion process

$$W(t) = q + \beta t - \lambda \frac{\mathbb{E}\big[S^{1+2\alpha}\big]}{2\mathbb{E}[S^{\alpha}]} t^2 + \sigma B(t), \tag{13}$$

with $\sigma^2 = \lambda^2 \mathbb{E}[S^{\alpha}]\mathbb{E}[S^{2+\alpha}]$, and $B(\cdot)$ is a standard Brownian motion.

Surprisingly, the assumption that α lies in the interval $[0,1]$ plays no role in our proof. On the other hand, we see from (13) that

$$\max\big\{\mathbb{E}\big[S^{2+\alpha}\big], \mathbb{E}\big[S^{1+2\alpha}\big], \mathbb{E}[S^{\alpha}]\big\} < \infty \tag{14}$$

is a necessary condition for Theorem 1 to hold. For example, for S having exponential distribution we need to restrict to $\alpha > -1$. This suggests that Theorem 1 remains true as long as $\alpha \in \mathbb{R}$ is such that (14) is satisfied. From the modeling point of view, $\alpha > 0$ represents a situation in which customers with larger job sizes have a stronger incentive to join the queue. The larger α is (e.g., $\alpha > 1$), the stronger the incentive is. On the other hand, when $\alpha < 0$, customers with large job sizes are lazy and thus favor joining the queue later. In what follows, for simplicity, we will limit ourselves to the case $\alpha \in [0,1]$.

By the continuous-mapping theorem and Theorem 1 we have the following.

Theorem 2 (Number of Customers Served in the First Busy Period)**.** *Assume that $\alpha \in [0,1]$, $\mathbb{E}[S^{2+\alpha}] < \infty$ and that the heavy-traffic condition (3) holds. Assume further that $\mathbf{Q}_n(0) = q$. Then, as $n \to \infty$, the number of customers served in the first busy period $\mathrm{BP}_n := H_{\mathbf{Q}_n}(0)$ converges to*

$$\mathrm{BP}_n \xrightarrow{\mathrm{d}} H_{\phi(W)}(0), \tag{15}$$

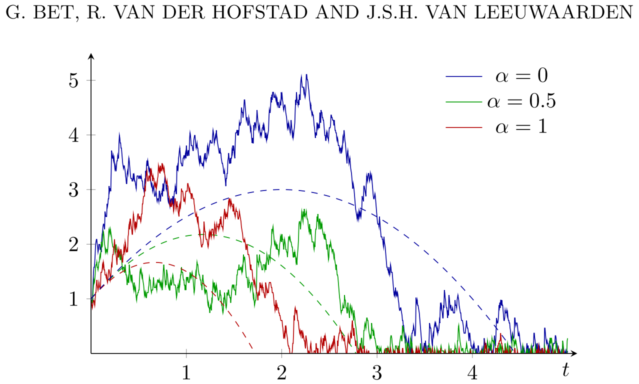
where $W(\cdot)$ is given in (13).

Theorem 1 implies that the typical queue length for the $\Delta_{(i)}^{\alpha}/G/1$ system in heavy traffic is $O_{\mathbb{P}}(n^{1/3})$ and that the typical busy period consists of $O_{\mathbb{P}}(n^{2/3})$ services. The linear drift $t \to \beta\lambda t$ describes the position of the system inside the critical window. For $\beta > 0$, the system is initially overloaded, and the process $W(\cdot)$ is more likely to cause a large initial excursion. For $\beta < 0$ the traffic intensity approaches 1 from below, so that the system is initially stable. Consequently, the process $W(\cdot)$ has a strong initial negative drift, so that $\phi(W)(\cdot)$ is close to zero also for small t. Finally, the negative quadratic drift $t \to -\lambda \frac{\mathbb{E}[S^{1+2\alpha}]}{2\mathbb{E}[S^{\alpha}]} t^2$ captures the *depletion-of-points effect*. Indeed, for large times, the process $W(t)$ is dominated by $-\lambda \frac{\mathbb{E}[S^{1+2\alpha}]}{2\mathbb{E}[S^{\alpha}]} t^2$, so that $\phi(W)(t)$ performs only small excursions away from zero (Figure 1).

Let us now compare Theorem 1 with two known results. For $\alpha = 0$, the limit diffusion simplifies to

$$W(t) = \beta t - \frac{1}{2} t^2 + \sigma B(t), \tag{16}$$

Figure 1. Sample Paths of the Process $\mathbf{Q}_n(\cdot)$ for Various Values of α and $n = 10^4$



Notes. The service times are taken unit-mean exponential. The dashed curves represent the drift $t \mapsto q + \beta t - \lambda\mathbb{E}([S^{1+2\alpha}])/(2\mathbb{E}[S^{\alpha}])t^2$. In all plots, $q = 1$, $\beta = 1$, and $\lambda = 1/\mathbb{E}[S^{1+\alpha}]$.

with $\sigma^2 = \lambda^2\mathbb{E}[S^2]$, in agreement with Bet et al. (2019, theorem 5). In Bhamidi et al. (2010), it is shown that, when $(\mathcal{W}_i)_{i\in[n]}$ are i.i.d. and further assuming that $\mathbb{E}[\mathcal{W}^2]/\mathbb{E}[\mathcal{W}] = 1$, the exploration process of the corresponding inhomogeneous random graph converges to

$$\overline{W}(t) = \beta t - \frac{\mathbb{E}\left[\mathcal{W}^3\right]}{2\mathbb{E}\left[\mathcal{W}^2\right]^2} t^2 + \frac{\sqrt{\mathbb{E}\left[\mathcal{W}\right]\mathbb{E}\left[\mathcal{W}^3\right]}}{\mathbb{E}\left[\mathcal{W}^2\right]} B(t). \tag{17}$$

For $\alpha = 1$, (13) can be rewritten using (3) as

$$W(t) = \beta t - \frac{\mathbb{E}[S^3]}{2\mathbb{E}[S^2]^2} t^2 + \frac{\sqrt{\mathbb{E}[S]\mathbb{E}[S^3]}}{\mathbb{E}[S^2]} B(t). \tag{18}$$

Therefore, the two processes coincide if $\mathcal{W}_i = S_i$, as expected.

2.2. Numerical Results

We now use Theorem 2 to obtain numerical results for the first busy period. We shall also use the explicit expression of the probability density function of the first passage time of zero of $\phi(W)$ obtained by Martin-Löf (1998) (see also van der Hofstad et al. 2010). Let $\mathrm{Ai}(x)$ and $\mathrm{Bi}(x)$ denote the classical Airy functions (Abramowitz and Stegun 1964). The first passage time of zero of $W(t) = q + \beta t - 1/2t^2 + \sigma B(t)$ has probability density (Martin-Löf 1998)

$$f\left(t;\beta,\sigma\right) = \mathrm{e}^{-\left(\left(t-\beta\right)^3 + \beta^3\right)/6\sigma^2 - \beta a} \int_{-\infty}^{+\infty} \mathrm{e}^{tu} \frac{\mathrm{Bi}(cu)\mathrm{Ai}(c(u-a)) - \mathrm{Ai}(cu)\mathrm{Bi}(c(u-a))}{\pi\left(\mathrm{Ai}(cu)^2 + \mathrm{Bi}(cu)^2\right)} \mathrm{d}u, \tag{19}$$

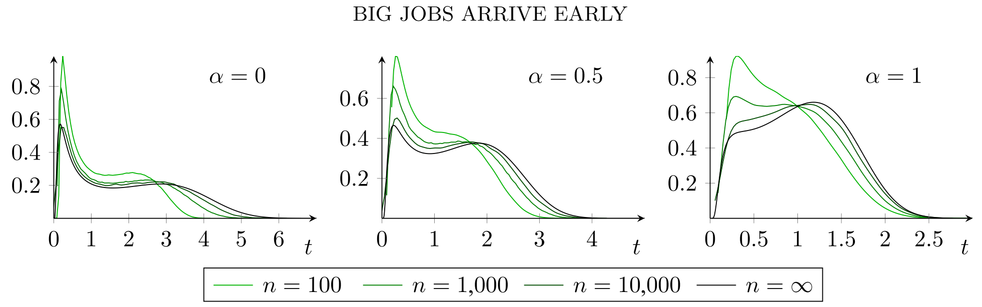
where $c = (2\sigma^2)^{1/3}$ and $a = q/\sigma^2 > 0$. The result (19) can be extended to a diffusion with a general quadratic drift through the scaling relation $W(\tau^2 t) = \tau(q/\tau + \beta\tau t - \tau^3 t^2/2 + \sigma B(t))$.

Figure 2 shows the empirical density of BP_n, for increasing values of n and various values of α, together with the exact limiting value (19).

Table 1 shows the mean busy period for different choices of α and different service time distributions. We computed the exact value for $n = \infty$ by numerically integrating (19). Observe that $\mathbb{E}[\mathrm{BP}_n]$ decreases with α. This might seem counterintuitive, because the larger α, the more likely customers with larger service join the queue early, who in turn might initiate a large busy period. Let us explain this apparent contradiction. When the arrival rate λ is fixed, assumption (3) does not necessarily hold, and $\mathbb{E}[\mathrm{BP}_n]$ increases with α, as can be seen in Table 2.

However, our heavy-traffic condition (3) implies that λ depends on α because $\lambda = 1/\mathbb{E}[S^{1+\alpha}]$. The interpretation of condition (3) is that, on average, one customer joins the queue during one service time. Notice that, because of the size biasing, the average service time is not $\mathbb{E}[S]$. Therefore, the number of customers that join during a (long) service is roughly equal to 1 as $\alpha \uparrow 1$. However, when customers with large services leave the system, they are not able to join any more. As $\alpha \uparrow 1$, customers with large services leave the system earlier.

Figure 2. Density Plot (Black) and Gaussian Kernel Density Estimates (Colored) Obtained by Running 10^6 Simulations of a $\Delta_{(i)}^{\alpha}/G/1$ Queue with $n = 100, 1{,}000, 10{,}000$ Customers and $\alpha = 0,\ 1/2,\ 1$



Note. In all cases, the service times are exponentially distributed, and $q = \beta = \mathbb{E}[S] = 1$.

Therefore, as $\alpha \uparrow 1$, the resulting second-order *depletion-of-points effect* causes shorter excursions as time progresses (Figure 1). In the limit process, this phenomenon is represented by the fact that the coefficient of the negative quadratic drift increases as $\alpha \uparrow 1$, as shown in the following lemma.

Lemma 1. *Let*

$$\alpha \mapsto f(\alpha) := \frac{\mathbb{E}\big[S^{1+2\alpha}\big]}{\mathbb{E}[S^{\alpha}]\mathbb{E}[S^{1+\alpha}]}. \tag{20}$$

Then $f'(\alpha) \geq 0$.

Proof. Because

$$f'(\alpha) = \frac{2\mathbb{E}\big[\log(S)S^{1+2\alpha}\big]}{\mathbb{E}[S^{\alpha}]\mathbb{E}[S^{1+\alpha}]} - \frac{\mathbb{E}\big[S^{1+2\alpha}\big]\mathbb{E}\big[\log(S)S^{\alpha}\big]}{\mathbb{E}[S^{\alpha}]^2\mathbb{E}[S^{1+\alpha}]} - \frac{\mathbb{E}\big[S^{1+2\alpha}\big]\mathbb{E}\big[\log(S)S^{1+\alpha}\big]}{\mathbb{E}[S^{\alpha}]\mathbb{E}[S^{1+\alpha}]^2}, \tag{21}$$

$f'(\alpha) \geq 0$ if and only if

$$2\mathbb{E}\big[\log(S)S^{1+2\alpha}\big]\mathbb{E}[S^{\alpha}]\mathbb{E}\big[S^{1+\alpha}\big] \geq \mathbb{E}\big[S^{1+\alpha}\big]\mathbb{E}\big[S^{1+2\alpha}\big]\mathbb{E}\big[\log(S)S^{\alpha}\big] + \mathbb{E}[S^{\alpha}]\mathbb{E}\big[S^{1+2\alpha}\big]\mathbb{E}\big[\log(S)S^{1+\alpha}\big]. \tag{22}$$

We split the lefthand side in two identical terms and show that each of them dominates one term on the righthand side. That is

$$\mathbb{E}\big[\log(S)S^{1+2\alpha}\big]\mathbb{E}[S^{\alpha}]\mathbb{E}\big[S^{1+\alpha}\big] \geq \mathbb{E}\big[S^{1+\alpha}\big]\mathbb{E}\big[S^{1+2\alpha}\big]\mathbb{E}\big[\log(S)S^{\alpha}\big], \tag{23}$$

the proof of the second bound being analogous. The inequality (23) is equivalent to

$$\frac{\mathbb{E}\big[\big(\log(S)S^{1+\alpha}\big)S^{\alpha}\big]}{\mathbb{E}[S^{\alpha}]} \geq \frac{\mathbb{E}\big[S^{1+\alpha}S^{\alpha}\big]}{\mathbb{E}[S^{\alpha}]}\frac{\mathbb{E}\big[\log(S)S^{\alpha}\big]}{\mathbb{E}[S^{\alpha}]}. \tag{24}$$

Table 1. Numerical Values of $n^{-2/3}\mathbb{E}[\mathrm{BP}_n]$ for Different Population Sizes and the Exact Expression for $n = \infty$ Computed Using (19)

	Deterministic			Exponential			Hyperexponential		
α	0	1/2	1	0	1/2	1	0	1/2	1
n									
10^1	1.1318	1.1318	1.1318	1.0359	0.8980	0.7429	0.8920	0.6356	0.5332
10^2	1.5842	1.5842	1.5842	1.3584	1.0924	0.8333	1.0959	0.7454	0.5525
10^3	1.9188	1.9188	1.9188	1.6387	1.2506	0.9284	1.2936	0.8352	0.6134
10^4	2.1474	2.1474	2.1474	1.8419	1.3925	1.0014	1.4960	0.9210	0.6554
∞	2.3374	2.3374	2.3374	2.0038	1.4719	1.0440	1.6242	0.9717	0.6881

Notes. The service requirements are displayed in order of increasing coefficient of variation. In all cases $q = \beta = \mathbb{E}[S] = 1$. The hyperexponential service times follow a rate $\lambda_1 = 0.501$ exponential distribution with probability $p_1 = 1/2$ and a rate $\lambda_2 = 250.5$ exponential distribution with probability $p_2 = 1 - p_1 = 1/2$. Each value for finite n is the average of 10^4 simulations.

Table 2. Expected Number of Customers Served in the First Busy Period of the Nonscaled $\Delta_{(i)}^{\alpha}/G/1$ Queue with Mean One Exponential Service Times and Arrival Rate $\lambda = 0.01$

	Exponential				
α	0	1/4	1/2	3/4	1
n					
10^1	1.0854	1.0922	1.1053	1.1118	1.1306
10^2	5.9515	8.1928	11.4478	16.3598	22.0381

Note. In all cases, $q = 1$. Each value is the average of 10^4 simulations.

The term on the left and the two terms on the right can be rewritten as the expectation of a size-biased random variable W, so that (24) is equivalent to

$$\mathbb{E}\big[\log(W)W^{1+\alpha}\big] \geq \mathbb{E}\big[\log(W)\big]\mathbb{E}\big[W^{1+\alpha}\big]. \tag{25}$$

Finally, the inequality (25) holds because W is positive with probability 1, and $x \mapsto \log(x)$ and $x \mapsto x^{1+\alpha}$ are increasing functions (van der Hofstad 2016, lemma 2.14). □

2.3. Component Sizes of Directed Random Graphs

The *directed Erdős-Rényi random graph* (directed ERRG) is obtained by taking n vertices and placing each of the possible $n(n-1)$ directed edges with probability p. This differs from the undirected ERRG, where each of the possible $n(n-1)/2$ undirected edges is present independently with probability p.

A *strongly connected component* of a directed graph (digraph) is a subgraph such that, for every two pairs of vertices i, j, there exists a directed path from i to j *and* one from j to i. We denote the strongly connected components, ordered by decreasing size, as $\mathscr{C}_1, \mathscr{C}_2, \mathscr{C}_3, \ldots$ It can be shown that the directed ERRG undergoes a *phase transition* when $p = \tilde{p} = 1/n$ (Luczak 1990). Indeed, when $p = c/n$ with $c < 1$, the largest strongly connected component $\mathscr{C}_1$ is of size $O_{\mathbb{P}}(\log(n))$, and when $p = c/n$ with $c > 1$, $\mathscr{C}_1$ is of size $O_{\mathbb{P}}(n)$ and $\mathscr{C}_2$ is of size $O_{\mathbb{P}}(\log(n))$. When p is in the so-called *critical window* between the two regimes $p = \tilde{p} = n^{-1} + \lambda n^{-4/3}$, $\lambda \in \mathbb{R}$, the strongly connected components $\mathscr{C}_1, \mathscr{C}_2, \mathscr{C}_3, \ldots$ are all of size $O_{\mathbb{P}}(n^{1/3})$ (Luczak and Seierstad 2009).

A crucial tool for the study of *undirected* random graphs in the critical regime is the so-called *depth-first exploration process*, defined as follows. Start with an arbitrary vertex, then reveal its neighbors and place them in a stack. Then discard the first vertex, consider the first vertex in the stack (called the *active vertex*), reveal its neighbors, and place them in the stack. If at any point the stack is empty and not all vertices have been activated, take one vertex uniformly at random and place it in the stack. This process continues by exploring the neighbors of each revealed vertex in order of appearance. The sizes of the connected components are encoded as the time between successive minima of the exploration process. Because of this, the exploration process has been extensively applied to the study of *undirected* random graphs in the critical regime (Bollobás et al. 2007; Bhamidi et al. 2014, 2010, 2012, 2017; van der Hofstad et al. 2018). However, until recently, it was not known if this approach was useful for the study of *directed* critical random graphs.

This issue has been solved when recently Goldschmidt and Stephenson (2019) used exploration-process techniques in their breakthrough paper to determine the scaling limit for the critical directed ERRG, extending an earlier result for the undirected ERRG (Addario-Berry et al. 2012). Roughly speaking, their scaling limit fully characterizes the structure of the strongly connected components of the digraph as $n \to \infty$. Their argument crucially relies on an exploration process that at each step follows the outgoing edges, revealing the so-called *forward exploration trees* of the directed graph. To obtain the original graph from the forward exploration tree, one has then to add *back edges* independently with probability p.[1] The proof in Goldschmidt and Stephenson (2019) consists of two main steps:

1. Ignoring edge directions, the forward exploration trees have the same distribution as the exploration trees of the undirected ERRG (Goldschmidt and Stephenson 2019, proposition 2.1). In Aldous (1997), it is shown that the rescaled component sizes of the undirected ERRG are distributed, in the limit, as the excursions of a Brownian motion with negative parabolic drift. Hence, the rescaled sizes of the forward exploration trees follow the same limiting distribution.

2. Conditionally on their sizes, the forward exploration trees are independent. Their topological structure is the same as in the undirected ERRG, and thus the scaling limit is the same as Addario-Berry et al. (2012).

Finally, to obtain the scaling limit of the directed ERRG, one has to prove that the process that adds the back edges is continuous in a suitable sense and identify the corresponding operation on the limit object.

Despite its mathematical significance, the ERRG model is inappropriate as a model for most concrete applications because it is *homogeneous*, that is, the vertices share the same degree distribution. This leads to considering the (directed) *inhomogeneous random graph* (IRG). In the directed IRG, each vertex i is assigned a (possibly random) weight $\mathscr{W}_i \geq 0$ and an edge from i pointing to j is present with probability p_{ij} proportional to $f(\mathscr{W}_i, \mathscr{W}_j)$, where $f: \mathbb{R}^2 \mapsto \mathbb{R}^+$ is a smooth function. A simple example of such a graph is obtained by taking $f(x, y) = xy^\alpha$ for $\alpha \in \mathbb{R}$, so that

$$p_{ij} \sim \mathscr{W}_i \mathscr{W}_j^\alpha. \tag{26}$$

The exponent α is a parameter that controls the *out-degree* distribution of the vertices. Indeed, if for example $\alpha = 0$, then each outgoing edge from a given vertex i is present with equal probability proportional to $\mathscr{W}_i$. For $\alpha = 1$, we have that $p_{ij} = p_{ji}$, so the graph is undirected and we retreive the classical IRG (van der Hofstad 2016). Note that (26) leads to an interesting dependence between the out-degree and in-degree of vertices.

The next natural step in the study of critical directed random graphs is then to extend the result (Goldschmidt and Stephenson 2019) to the *inhomogeneous* setting. The main challenges are that

1. the scaling limit of the size of the forward exploration trees is not known, and
2. the scaling limit of a forward exploration tree with fixed size m is not known.

Our main result (Theorem 1) gives an answer to the first question. In fact, the embedded queueing process (4) is distributionally equivalent to the depth-first exploration process of the directed IRG. To see this, associate a vertex i to customer i and let $c(1)$ be the root. Then, draw a directed edge from $c(1)$ to $c(2), \dots, c(A_n(1) + 1)$, that is, to all customers who joined during the service time of $c(1)$. Then, draw an edge from $c(2)$ to each of the customers who joined during the service time of $c(2)$, and so on. According to this construction, we associate to each busy period of the queue a different directed tree, and thus the queueing process corresponds to a directed random forest. The degree of vertex $c(i)$ is $1 + |A_n(i)|$ and the total number of vertices in the first tree (say) is given by

$$H_{Q_n}(0) = \inf\{k \geq 0 : Q_n(k) = 0\}, \tag{27}$$

the (first) hitting time of zero of the process $Q_n(\cdot)$. The directed random forest thus created is distributionally equivalent to the exploration process of the directed IRG, as we now show in detail. In the directed IRG, conditionally on the weights $(\mathscr{W}_i)_{i \in [n]}$, a directed edge from i to j is present with probability

$$p_{ij} = 1 - \exp\left(-\frac{\mathscr{W}_i \mathscr{W}_j^\alpha}{\sum_{l \in [n]} \mathscr{W}_l}\right). \tag{28}$$

We set $S_i := (1 + \beta n^{-1/3})^{-1/\alpha} \mathscr{W}_i$ for $i \in [n]$. Then the probability that an edge from i to j is present is equal to

$$p_{ij} = 1 - \exp\left(-\left(1 + \beta n^{-1/3}\right) \frac{S_i}{n} \frac{S_j^\alpha n}{\sum_{l \in [n]} S_l}\right) = 1 - \exp\left(-\tilde{S}_i S_j^\alpha \frac{n}{\sum_{l \in [n]} S_l}\right) = \mathbb{P}\left(T_j \leq \tilde{S}_i | (S_l)_{l \in [n]}\right), \tag{29}$$

where $\tilde{S}_i = (1 + \beta n^{-1/3}) S_i / n$ and T_j are distributed as

$$T_j \overset{\mathrm{d}}{=} \mathrm{E}_j\left(\lambda_n S_j^\alpha\right), \tag{30}$$

with $\lambda_n := n / \sum_{i \in [n]} S_i$ and $\mathrm{E}_j(c)$ being exponential random variables with mean $1/c$ independent across j. This implies that, for every fixed $n \in \mathbb{N}$, the directed forest associated to the $\Delta_{(i)}^\alpha / G / 1$ queue with service times $S_i = (1 + \beta n^{-1/3})^{-1/\alpha} \mathscr{W}_i$ and arrival parameter $\lambda_n = n / \sum_{i \in [n]} S_i$ is distributed as the exploration process of the directed IRG with weights $(\mathscr{W}_i)_{i \in [n]}$.

Theorem 1 then implies that the size of each tree in the forward depth-first forest is $O_{\mathbb{P}}(n^{2/3})$ and gives the limiting distribution of the rescaled sizes. This is only an upper bound on the size of strongly connected components. In fact, we expect from Goldschmidt and Stephenson (2019) that the strongly connected components have size $O_{\mathbb{P}}(n^{1/3})$.

2.3.1. The Critical Condition and the Heavy-Traffic Condition. The undirected IRG with weights $(\mathcal{W}_i)_{i=1}^n$ is said to be *critical* (see Bhamidi et al. 2010, equation 1.13) if

$$\frac{\sum_{i\in[n]}\mathcal{W}_i^2}{\sum_{i\in[n]}\mathcal{W}_i} = \frac{\mathbb{E}[\mathcal{W}^2]}{\mathbb{E}[\mathcal{W}]} + o_{\mathbb{P}}\big(n^{-1/3}\big) = 1 + o_{\mathbb{P}}\big(n^{-1/3}\big). \tag{31}$$

Recall also that, in the special case $\alpha = 1$, the heavy-traffic condition (3) for the $\Delta_{(i)}^{\alpha}/G/1$ reads

$$\lambda_n \mathbb{E}\big[S^2\big]\big(1 + \beta n^{-1/3}\big) = 1 + \beta n^{-1/3} + o_{\mathbb{P}}\big(n^{-1/3}\big). \tag{32}$$

Consequently, if $S_i = (1 + \beta n^{-1/3})^{-1}\mathcal{W}_i$ and $\lambda_n = n/\sum_{i\in[n]} S_i$, the heavy-traffic condition (32) for the $\Delta_{(i)}^{\alpha}/G/1$ queue implies the criticality condition (31) for the associated random graph and vice versa.

2.3.2. Extension to the Queue-Length Process. By definition, the embedded queue (4) neglects the idle time of the server. Via a time-change argument, it is possible to prove that, in the limit, the (cumulative) idle time is negligible, and the embedded queue is arbitrarily close to the queue-length process uniformly over compact intervals. This has been proven for the $\Delta_{(i)}/G/1$ queue in Bet et al. (2017), and the techniques developed there can be extended to the $\Delta_{(i)}^{\alpha}/G/1$ queue without additional difficulties.

3. Preliminaries

The proof of Theorem 1 extends the techniques we developed in Bet et al. (2017). However, the dependency structure of the arrival times complicates the analysis considerably. Customers with larger job sizes have a higher probability of joining the queue earlier, and this gives rise to a size-biased reordering of the service times. In the next section, we study this phenomenon in detail.

Given two sequences of random variables $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$, we say that X_n converges in probability to X, and we denote it by $X_n \xrightarrow{\mathbb{P}} X$, if $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ as $n \to 0$ for each $\varepsilon > 0$. We also write $X_n = o_{\mathbb{P}}(Y_n)$ if $X_n/Y_n \xrightarrow{\mathbb{P}} 0$ and $X_n = O_{\mathbb{P}}(Y_n)$ if $(X_n/Y_n)_{n\geq 1}$ is tight. Given two real-valued random variables X, Y, we say that X *stochastically dominates* Y and denote it by $Y \leq X$, if $\mathbb{P}(X \leq x) \leq \mathbb{P}(Y \leq x)$ for all $x \in \mathbb{R}$.

For our results, we condition on the entire sequence $(S_i)_{i\geq 1}$. More precisely, if the random variables that we consider are defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then we define a new probability space $(\Omega, \mathcal{F}_S, \mathbb{P}_S)$, with $\mathbb{P}_S(A) := \mathbb{P}(A|(S_i)_{i=1}^{\infty})$ and $\mathcal{F}_s := \sigma(\{\mathcal{F}, (S_i)_{i=1}^{\infty}\})$, the σ-algebra generated by $\mathcal{F}$ and $(S_i)_{i=1}^{\infty}$. Correspondingly, for any random variable X on Ω, we define $\mathbb{E}_s[X]$ as the expectation with respect to $\mathbb{P}_S$, and $\mathbb{E}[X]$ for the expectation with respect to $\mathbb{P}$. We say that a sequence of events $(\mathcal{E}_n)_{n\geq 1}$ holds with high probability (w.h.p.) if $\mathbb{P}(\mathcal{E}_n) \to 1$ as $n \to \infty$.

First, we recall a well-known result that will be useful on several occasions.

Lemma 2. *Assume $(X_i)_{i=1}^n$ is a sequence of positive i.i.d. random variables such that $\mathbb{E}[X_i] < \infty$. Then $\max_{i\in[n]} X_i = o_{\mathbb{P}}(n)$.*

Proof. We have the inclusion of events

$$\Big\{\max_{i\in[n]} X_i \geq \varepsilon n\Big\} \subseteq \bigcup_{i=1}^{n}\{X_i \geq \varepsilon n\}. \tag{33}$$

Therefore,

$$\mathbb{P}\Big(\max_{i\in[n]} X_i \geq \varepsilon n\Big) \leq \sum_{i=1}^{n}\mathbb{P}(X_i \geq \varepsilon n). \tag{34}$$

Because for any positive random variable Y, $\varepsilon\mathbb{1}_{\{Y\geq\varepsilon\}} \leq Y\mathbb{1}_{\{Y\geq\varepsilon\}}$ almost surely, it follows

$$\mathbb{P}\Big(\max_{i\in[n]} X_i \geq \varepsilon n\Big) \leq \frac{\sum_{i=1}^{n}\mathbb{E}\big[X_i\mathbb{1}_{\{X_i\geq\varepsilon n\}}\big]}{\varepsilon n} = \frac{\mathbb{E}\big[X_1\mathbb{1}_{\{X_1\geq\varepsilon n\}}\big]}{\varepsilon}. \tag{35}$$

The rightmost term tends to zero as $n \to \infty$ because $\mathbb{E}[X_1] < \infty$, and this concludes the proof.

Given a vector $\bar{x} = (x_1, x_2, \ldots, x_n)$ with deterministic, real-valued entries, the size-biased ordering of $\bar{x}$ is a *random* vector $X^{(s)} = (X_1^{(s)}, X_2^{(s)}, \ldots, X_n^{(s)})$ such that

$$\mathbb{P}\Big(X_1^{(s)} = x_j\Big) = \frac{x_j}{\sum_{l=1}^{n} x_l}, \quad \mathbb{P}\Big(x_2^{(s)} = x_j \mid X_1^{(s)}\Big) = \frac{x_j}{\sum_{l=1}^{n} x_l - X_1^{(s)}}, \quad \ldots. \quad \square \tag{36}$$

More generally, for any $\alpha \in \mathbb{R}$, the α size-biased ordering of $\bar{x}$ is given by a vector $\bar{X}^{(\alpha)} = (X_1^{(\alpha)}, X_2^{(\alpha)}, \ldots, X_n^{(\alpha)})$ such that

$$\mathbb{P}\left(X_1^{(\alpha)} = x_j\right) = \frac{x_j^\alpha}{\sum_{l=1}^n x_l^\alpha}, \quad \mathbb{P}\left(X_2^{(\alpha)} = x_j \mid X_1^{(\alpha)} = x_i\right) = \frac{x_j^\alpha}{\sum_{l=1}^n x_l^\alpha - x_i^\alpha}, \quad \ldots. \tag{37}$$

Finally, we define

$$\mathfrak{S}_k = \{c(1), \ldots, c(k)\} \tag{38}$$

as the set of the first k customers served. The following lemma is the first step in understanding the structure of the arrival process:

Lemma 3. (Size-Biased Reordering of the Arrivals). *The order of appearance of customers is the α size-biased ordering of their service times. In other words,*

$$\mathbb{P}_S\left(c(j) = i \mid \mathfrak{S}_{j-1}\right) = \frac{S_i^\alpha}{\sum_{l \notin \mathfrak{S}_{j-1}} S_l^\alpha}. \tag{39}$$

Proof. Conditionally on $(S_l)_{l=1}^n$, the arrival times are independent exponential random variables. By basic properties of exponentials, we have

$$\mathbb{P}_S\left(c(j) = i \mid \mathfrak{S}_{j-1}\right) = \mathbb{P}_S\left(\min\left\{T_l : \ l \notin \mathfrak{S}_{j-1}\right\} = T_i \mid \mathfrak{S}_{j-1}\right) = \frac{S_i^\alpha}{\sum_{l \notin \mathfrak{S}_{j-1}} S_l^\alpha}, \tag{40}$$

as desired. □

We remark that (40) differs from the classical size-biased reordering in that the weights are a *nonlinear* function of the $(S_i)_{i=1}^n$. In our definition of the queueing process (4) and (5), we do not keep track of the service requirements of the customers that join the queue but only of their arrival times (2). Therefore, at the start of service, a customer's service requirement is a random variable that depends on the arrival time relative to the remaining customers.

The next lemma is crucial, establishing stochastic domination between the service requirements of the customers in order of appearance. Recall that X stochastically dominates Y (with notation $Y \preceq X$) if and only if there exists a probability space $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}})$ and two random variables $\bar{X}$, $\bar{Y}$ defined on $\bar{\Omega}$ such that $\bar{X} \overset{d}{=} X$, $\bar{Y} \overset{d}{=} Y$, and $\bar{\mathbb{P}}(\bar{Y} \leq \bar{X}) = 1$.

Lemma 4. *Assume that $\alpha > 0$. Let $f : \mathbb{R}^+ \to \mathbb{R}$ be a function such that $\mathbb{E}[f(S)S^\alpha] < \infty$. Then there exists a constant $C_{f,s}$ such that almost surely, for n large enough,*

$$\mathbb{E}_s\left[f\left(S_{c(k)}\right)\right] \leq C_{f,s} < \infty, \tag{41}$$

uniformly in $k \leq cn$, for a fixed $c \in (0,1)$.

Proof. We compute explicitly

$$\begin{aligned} \mathbb{E}_s\left[f\left(S_{c(k)}\right)\right] &= \mathbb{E}_s\left[\frac{\sum_{j \notin \mathfrak{S}_{k-1}} f\left(S_j\right) S_j^\alpha}{\sum_{j \notin \mathfrak{S}_{k-1}} S_j^\alpha}\right] \\ &= \mathbb{E}_s\left[\frac{\sum_{j \in [n]} f\left(S_j\right) S_j^\alpha - \sum_{j \in \mathfrak{S}_k} f\left(S_j\right) S_j^\alpha}{\sum_{j \notin \mathfrak{S}_{k-1}} S_j^\alpha}\right] \\ &\leq \mathbb{E}_s\left[\frac{1}{\sum_{j \notin \mathfrak{S}_{k-1}} S_j^\alpha}\right] \sum_{j \in [n]} f\left(S_j\right) S_j^\alpha. \end{aligned} \tag{42}$$

We have the almost sure bound

$$\frac{1}{\sum_{j \notin \mathfrak{S}_{k-1}} S_j^\alpha} = \frac{1}{\sum_{j \in [n]} S_j^\alpha - \sum_{j \in \mathfrak{S}_{k-1}} S_j^\alpha} \leq \frac{1}{\sum_{j \in [n]} S_j^\alpha - \sum_{j \in \mathfrak{S}_{k-1}} S_j^\alpha} \leq \frac{1}{\sum_{j \in [n]} S_j^\alpha - \sum_{j=1}^{k-1} S_{(n-j+1)}^\alpha} = \frac{1}{\sum_{j=1}^{n-k+1} S_{(j)}^\alpha}, \tag{43}$$

where $S^{\alpha}_{(1)} \le S^{\alpha}_{(2)} \le \ldots \le S^{\alpha}_{(n)}$ denote the order statistics of the finite sequence $(S^{\alpha}_i)_{i\in[n]}$. There exists $p \in (0,1)$ such that $n - k + 1 \ge pn$, for large enough n. Consequently,

$$\frac{1}{\sum_{j\notin\mathfrak{S}_{k-1}} S^{\alpha}_j} \le \frac{1}{\sum_{j=1}^{\lfloor pn\rfloor} S^{\alpha}_{(j)}}, \tag{44}$$

so that we have

$$\mathbb{E}_s\big[f\big(S_{c(k)}\big)\big] \le \frac{\sum_{j\in[n]} f\big(S_j\big) S^{\alpha}_j}{\sum_{j=1}^{\lfloor pn\rfloor} S^{\alpha}_{(j)}}. \quad \square \tag{45}$$

Let us denote by ξ_p the pth quantile of the distribution $F_s(\cdot)$, and let us assume, without loss of generality, that $f_s(\xi_p) > 0$. Note that $S_{(\lfloor np\rfloor)} = F^{-1}_{n,s}(\lfloor np\rfloor/n)$, where $F_{n,s}(t) = \sum_{i=1}^n \mathbb{1}_{\{S_i\le t\}}/n$ is the empirical distribution function of the $(S_i)_{i=1}^n$, and $\xi_p = F^{-1}_s(p)$. Indeed, the assumption $f_s(\xi_p) > 0$ implies that $F_s(\cdot)$ is invertible in a neighborhood of ξ_p. We have that, as $n \to \infty$,

$$S_{(\lfloor np\rfloor)} \overset{\text{a.s.}}{\longrightarrow} \xi_p. \tag{46}$$

In particular, as $n \to \infty$,

$$\frac{1}{n}\left|\sum_{j\in[n]} S_j \mathbb{1}_{\{S_j\le\xi_p\}} - \sum_{j\in[n]} S_j \mathbb{1}_{\left\{S_j\le S_{(\lfloor pn\rfloor)}\right\}}\right| \overset{\text{a.s.}}{\longrightarrow} 0. \tag{47}$$

Therefore, by the strong law of large numbers, as $n \to \infty$,

$$\frac{\sum_{j=1}^{\lfloor pn\rfloor} S_{(j)}}{n} \overset{\text{a.s.}}{\longrightarrow} \mathbb{E}\Big[S\mathbb{1}_{\{S\le\xi_p\}}\Big]. \tag{48}$$

Then, choosing $C_{n,f,s} = \mathbb{E}[f(S)S^{\alpha}]/\mathbb{E}[S\mathbb{1}_{\{S\le\xi_p\}}] + \varepsilon$, for an arbitrary $\varepsilon > 0$, gives the desired result.

If $\alpha > 0$, as is the case in our setting, the proof of Lemma 4 shows that, uniformly in $k = O(n^{2/3})$,

$$\mathbb{E}_s\big[f\big(S_{c(k)}\big)\big] \le \frac{\sum_{j\in[n]} f\big(S_j\big) S^{\alpha}_j}{\sum_{j=1}^{\lfloor pn\rfloor} S^{\alpha}_{(j)}} = \frac{\sum_{j\in[n]} f\big(S_j\big) S^{\alpha}_j}{\sum_{j=1}^{n} S^{\alpha}_{(j)}}\left(1 + \frac{\sum_{j=\lfloor pn\rfloor}^{n} S^{\alpha}_{(j)}}{\sum_{j=1}^{\lfloor pn\rfloor} S^{\alpha}_{(j)}}\right), \tag{49}$$

and therefore

$$\mathbb{E}_s\big[f\big(S_{c(k)}\big)\big] \le \mathbb{E}_s\big[f\big(S_{c(1)}\big)\big]\big(1 + O_{\mathbb{P}_S}(1)\big). \tag{50}$$

If $f(\cdot)$ is an increasing function, (50) makes precise the intuition that, if $\alpha > 0$, customers with larger job sizes join the queue earlier. We will often make use of the expression (50)

The following lemma will often prove useful in dealing with sums over a random index set.

Lemma 5. (Uniform Convergence of Random Sums). *Let $(S_j)_{j=1}^n$ be a sequence of positive random variables such that $\mathbb{E}[S^{2+\alpha}] < +\infty$, for $\alpha \in (0,1)$. Then,*

$$\sup_{\substack{\mathscr{X}\subseteq[n] \\ |\mathscr{X}|=O_{\mathbb{P}}\left(n^{2/3}\right)}} \frac{1}{n}\sum_{j\in\mathscr{X}} S^{\alpha}_j = o_{\mathbb{P}}(1). \tag{51}$$

Proof. By Lemma 2, $\max_{j\in[n]} S^{\alpha}_j = o_{\mathbb{P}}(n^{\alpha/(2+\alpha)})$. This gives

$$\sup_{\substack{\mathscr{X}\subseteq[n] \\ |\mathscr{X}|=O_{\mathbb{P}}\left(n^{2/3}\right)}} \frac{1}{n}\sum_{j\in\mathscr{X}} S^{\alpha}_j \le \frac{\max_{j\in[n]} S^{\alpha}_j}{n^{1/3}} O_{\mathbb{P}}(1) = o_{\mathbb{P}}\left(n^{\frac{\alpha-2/3-\alpha/3}{2+\alpha}}\right) = o_{\mathbb{P}}\left(n^{\frac{2\alpha-1}{32+\alpha}}\right). \tag{52}$$

Because $\alpha - 1 \le 0$ by assumption, the claim is proven. $\square$

We now focus on the i-th customer joining the queue (for i large) and characterize the distribution of its service time. In particular, for $\alpha > 0$, this is different from S_i.

Lemma 6. (Size-Biased Distribution of the Service Times). *For every bounded, real-valued continuous function $f(\cdot)$, as $n \longrightarrow \infty$,*

$$\mathbb{E}_s\left[f\left(S_{c(i)}\right) \mid \mathscr{F}_{i-1}\right] \xrightarrow{\mathbb{P}} \frac{\mathbb{E}\left[f(S)S^\alpha\right]}{\mathbb{E}[S^\alpha]}, \tag{53}$$

uniformly for $i = O_{\mathbb{P}_S}(n^{2/3})$. Moreover, as $n \longrightarrow \infty$,

$$\mathbb{E}_s\left[f\left(S_{c(i)}\right)\right] \rightarrow \frac{\mathbb{E}\left[f(S)S^\alpha\right]}{\mathbb{E}[S^\alpha]}, \qquad \text{for } i = O_{\mathbb{P}_S}\left(n^{2/3}\right). \tag{54}$$

Proof. First, note that

$$\mathbb{E}_s\left[f\left(S_{c(i)}\right) \mid \mathscr{F}_{i-1}\right] = \sum_{j \notin \mathfrak{S}_{i-1}} f\left(S_j\right)\mathbb{P}_S\left(c(i) = j \mid \mathscr{F}_{i-1}\right) = \sum_{j \notin \mathfrak{S}_{i-1}} \frac{f\left(S_j\right)S_j^\alpha}{\sum_{l \notin \mathfrak{S}_{i-1}} S_l^\alpha}. \tag{55}$$

This can be further decomposed as

$$\mathbb{E}_s\left[f\left(S_{c(i)}\right) \mid \mathscr{F}_{i-1}\right] = \frac{\sum_{j=1}^n f\left(S_j\right)S_j^\alpha - \sum_{j \in \mathfrak{S}_{i-1}} f\left(S_j\right)S_j^\alpha}{\sum_{l=1}^n S_l^\alpha - \sum_{l \in \mathfrak{S}_{i-1}} S_l^\alpha}. \tag{56}$$

Because $|\mathfrak{S}_{i-1}| = i - 1$ and $i = O_{\mathbb{P}}(n^{2/3})$, by the law of large numbers and Lemma 5,

$$\frac{\sum_{j \notin \mathfrak{S}_{i-1}} f\left(S_j\right)S_j^\alpha}{n} \xrightarrow{\mathbb{P}} \mathbb{E}\left[f(S)S^\alpha\right], \qquad \frac{\sum_{l \notin \mathfrak{S}_{i-1}} S_l^\alpha}{n} \xrightarrow{\mathbb{P}} \mathbb{E}[S^\alpha]. \tag{57}$$

uniformly in $i = O_{\mathbb{P}}(n^{2/3})$. This gives the first claim.

Furthermore, we bound $\mathbb{E}_s[f(S_{c(i)}) \mid \mathscr{F}_{i-1}]$ as

$$\mathbb{E}_s\left[f\left(S_{c(i)}\right) \mid \mathscr{F}_{i-1}\right] = \sum_{j \notin \mathfrak{S}_{i-1}} \frac{f\left(S_j\right)S_j^\alpha}{\sum_{l \notin \mathfrak{S}_{i-1}} S_l^\alpha} \leq \sup_{x \geq 0} f(x) < \infty. \tag{58}$$

Because $\mathbb{E}_s[f(S_{c(i)})] = \mathbb{E}_s[\mathbb{E}_s[f(S_{c(i)}) \mid \mathscr{F}_{i-1}]]$, using (53) and the dominated convergence theorem, the second claim follows. □

In Lemma 6, we studied the distribution of the service time of the ith customer, and we now focus on its (conditional) moments. The following lemma should be interpreted as follows: Because of the size-biased reordering of the customer arrivals, the service time of the ith customer being served (for i large) is highly concentrated.

Lemma 7. *For any fixed $\gamma \in [-1, 1]$,*

$$\mathbb{E}_s\left[S_{c(i)}^{1+\gamma} \mid \mathscr{F}_{i-1}\right] = \frac{\mathbb{E}\left[S^{1+\gamma+\alpha}\right]}{\mathbb{E}[S^\alpha]} + o_{\mathbb{P}}(1) \quad \text{for } i = O_{\mathbb{P}_S}\left(n^{2/3}\right), \tag{59}$$

where the error term is uniform in $i = O_{\mathbb{P}_S}(n^{2/3})$. Moreover, the convergence holds in L^1, that is,

$$\mathbb{E}_s\left[\left|\mathbb{E}_s\left[S_{c(i)}^{1+\gamma} \mid \mathscr{F}_{i-1}\right] - \frac{\mathbb{E}\left[S^{1+\gamma+\alpha}\right]}{\mathbb{E}[S^\alpha]}\right|\right] = o_{\mathbb{P}}(1), \tag{60}$$

uniformly in $i = O_{\mathbb{P}_S}(n^{2/3})$.

Proof. In order to apply Lemma 6, we first split

$$S_{c(i)}^{1+\gamma} = \left(S_{c(i)} \wedge K\right)^{1+\gamma} + \left(\left(S_{c(i)} - K\right)^+\right)^{1+\gamma}, \tag{61}$$

where $K > 0$ is arbitrary, so that

$$\mathbb{E}_S\Big[S_{c(i)}^{1+\gamma} \mid \mathscr{F}_{i-1}\Big] = \mathbb{E}_S\Big[\big(S_{c(i)} \wedge K\big)^{1+\gamma} \mid \mathscr{F}_{i-1}\Big] + \mathbb{E}_S\Big[\big((S_{c(i)} - K)^+\big)^{1+\gamma} \mid \mathscr{F}_{i-1}\Big]. \tag{62}$$

The first term is bounded and therefore converges to $\mathbb{E}[(S \wedge K)^{1+\gamma} S^\alpha]/\mathbb{E}[S^\alpha]$ by Lemma 6. The second term is bounded through Markov's inequality, as

$$\mathbb{P}_S\Big(\mathbb{E}_S\Big[\big((S_{c(i)} - K)^+\big)^{1+\gamma} \mid \mathscr{F}_{i-1}\Big] \geq \varepsilon\Big) \leq \frac{\mathbb{E}_S\Big[\big((S_{c(i)} - K)^+\big)^{1+\gamma}\Big]}{\varepsilon}. \tag{63}$$

Next we apply Lemma 4 with $f(x) = f_K(x) = ((x - K)^+)^{1+\gamma}$,

$$\mathbb{E}_S\Big[\big((S_{c(i)} - K)^+\big)^{1+\gamma}\Big] \leq C_{f_K,S}. \tag{64}$$

Therefore,

$$\left|\mathbb{E}_S\Big[S_{c(i)}^{1+\gamma} \mid \mathscr{F}_{i-1}\Big] - \frac{\mathbb{E}\big[S^{1+\gamma+\alpha}\big]}{\mathbb{E}[S^\alpha]}\right| \leq \left|\mathbb{E}_S\Big[\big(S_{c(i)} \wedge K\big)^{1+\gamma} \mid \mathscr{F}_{i-1}\Big] - \frac{\mathbb{E}\big[S^{1+\gamma+\alpha}\big]}{\mathbb{E}[S^\alpha]}\right| + C_{f_K,S}. \tag{65}$$

The proof of Lemma 4 shows that, for any $\varepsilon > 0$, $\lim_{K\to\infty} C_{f_K,S} \leq \varepsilon$, and thus $\lim_{K\to\infty} C_{f_K,S} = 0$. Therefore, by letting $K \to \infty$ in (65), (59) follows. Next, we split

$$\mathbb{E}_S\left[\left|\mathbb{E}_S\Big[S_{c(i)}^{1+\gamma} \mid \mathscr{F}_{i-1}\Big] - \frac{\mathbb{E}\big[S^{1+\gamma+\alpha}\big]}{\mathbb{E}[S^\alpha]}\right|\right] \leq \mathbb{E}_S\left[\left|\big(S_{c(i)} \wedge K\big)^{1+\gamma} - \frac{\mathbb{E}\big[S^{1+\gamma+\alpha}\big]}{\mathbb{E}[S^\alpha]}\right|\right] + \mathbb{E}_S\Big[\big((S_{c(i)} - K)^+\big)^{1+\gamma}\Big]. \tag{66}$$

The second term can be bounded as in (64). For the first term,

$$\begin{aligned}\mathbb{E}_S\left[\left|\big(S_{c(i)} \wedge K\big)^{1+\gamma} - \frac{\mathbb{E}\big[S^{1+\gamma+\alpha}\big]}{\mathbb{E}[S^\alpha]}\right|\right] &\leq \left|\frac{\sum_{j=1}^n \big(S_j \wedge K\big)^{1+\gamma} S_j^\alpha}{\sum_{j=1}^n S_j^\alpha} - \frac{\mathbb{E}\big[S^{1+\gamma+\alpha}\big]}{\mathbb{E}[S^\alpha]}\right| \\ + \mathbb{E}_S\left[\left|\frac{\sum_{j=1}^n \big(S_j \wedge K\big)^{1+\gamma} S_j^\alpha \sum_{l\in\mathfrak{S}_{i-1}} S_l^\alpha}{\big(\sum_{j=1}^n S_j^\alpha\big)^2}\right|\right] &+ \mathbb{E}_S\left[\left|\frac{\sum_{l=1}^n S_l^\alpha \sum_{j\in\mathfrak{S}_{i-1}} \big(S_j \wedge K\big)^{1+\gamma} S_j^\alpha}{\big(\sum_{j=1}^n S_j^\alpha\big)^2}\right|\right],\end{aligned} \tag{67}$$

where we have used that $|(a - b)/(c - d) - a/c| \leq ad/c^2 + bc/c^2$, for positive a, b, c, and d. The second and third terms converge uniformly over $i = O_{\mathbb{P}_S}(n^{2/3})$ by Lemma 5. Summarizing,

$$\mathbb{E}_S\left[\left|\mathbb{E}_S\Big[S_{c(i)}^{1+\gamma} \mid \mathscr{F}_{i-1}\Big] - \frac{\mathbb{E}\big[S^{1+\gamma+\alpha}\big]}{\mathbb{E}[S^\alpha]}\right|\right] \leq \left|\frac{\sum_{j=1}^n \big(S_j \wedge K\big)^{1+\gamma} S_j^\alpha}{\sum_{j=1}^n S_j^\alpha} - \frac{\mathbb{E}\big[S^{1+\gamma+\alpha}\big]}{\mathbb{E}[S^\alpha]}\right| + \frac{\sum_{l=1}^n ((S_l - K)^+)^{1+\gamma}}{\sum_{j=1}^n S_j^\alpha} + o_{\mathbb{P}}(1). \tag{68}$$

Letting first $n \to \infty$ and then $K \to \infty$, (60) follows. □

We will make use of Lemma 7 several times throughout the proof, with the specific choices $\gamma \in \{0, \alpha, 1\}$. The following lemma is of central importance in the proof of the uniform convergence of the quadratic part of the drift.

Lemma 8. *As $n \to \infty$,*

$$n^{-2/3} \sup_{j\leq tn^{2/3}} \left|\sum_{i=1}^j \left(S_{c(i)}^{1+\alpha} - \frac{\mathbb{E}\big[S^{1+2\alpha}\big]}{\mathbb{E}[S]}\right)\right| \xrightarrow{\mathbb{P}} 0. \tag{69}$$

Proof. By Lemma 7, (68) is equivalent to

$$n^{-2/3} \sup_{j\leq tn^{2/3}} \left|\sum_{i=1}^j \left(S_{c(i)}^{1+\alpha} - \mathbb{E}\Big[S_{c(i)}^{1+\alpha} \mid \mathscr{F}_{i-1}\Big]\right)\right| \xrightarrow{\mathbb{P}} 0. \tag{70}$$

We split the event space and separately bound

$$n^{-2/3}\sup_{j\le tn^{2/3}}\left|\sum_{i=1}^{j}\left(S_{c(i)}^{1+\alpha}\mathbb{1}_{\left\{S_{c(i)}^{1+\alpha}\le K_n\right\}}-\mathbb{E}\left[S_{c(i)}^{1+\alpha}\mathbb{1}_{\left\{S_{c(i)}^{1+\alpha}\le K_n\right\}}\mid\mathscr{F}_{i-1}\right]\right)\right| \tag{71}$$

and

$$n^{-2/3}\sup_{j\le tn^{2/3}}\left|\sum_{i=1}^{j}\left(S_{c(i)}^{1+\alpha}\mathbb{1}_{\left\{S_{c(i)}^{1+\alpha}> K_n\right\}}-\mathbb{E}\left[S_{c(i)}^{1+\alpha}\mathbb{1}_{\left\{S_{c(i)}^{1+\alpha}> K_n\right\}}\mid\mathscr{F}_{i-1}\right]\right)\right|, \tag{72}$$

for a sequence $(K_n)_{n\ge1}$ that we choose later on and is such that $K_n\to\infty$. We start with (71). Because the sum inside the absolute value is a martingale as a function of j, (71) can be bounded through Doob's L^p inequality (Klenke 2008, theorem 11.2) with $p=2$ as

$$\begin{aligned}
&\mathbb{P}_S\left(\sup_{j\le tn^{2/3}}\left|\sum_{i=1}^{j}\left(S_{c(i)}^{1+\alpha}\mathbb{1}_{\left\{S_{c(i)}^{1+\alpha}\le K_n\right\}}-\mathbb{E}_S\left[S_{c(i)}^{1+\alpha}\mathbb{1}_{\left\{S_{c(i)}^{1+\alpha}\le K_n\right\}}\mid\mathscr{F}_{i-1}\right]\right)\right|\ge\varepsilon n^{2/3}\right)\\
&\qquad\le\frac{1}{\varepsilon n^{4/3}}\mathbb{E}_S\left[\left(\sum_{i=1}^{tn^{2/3}}\left(S_{c(i)}^{1+\alpha}\mathbb{1}_{\left\{S_{c(i)}^{1+\alpha}\le K_n\right\}}-\mathbb{E}_S\left[S_{c(i)}^{1+\alpha}\mathbb{1}_{\left\{S_{c(i)}^{1+\alpha}\le K_n\right\}}\mid\mathscr{F}_{i-1}\right]\right)\right)^2\right]\\
&\qquad\le\frac{2}{\varepsilon n^{4/3}}\sum_{i=1}^{tn^{2/3}}\mathbb{E}_S\left[S_{c(i)}^{2+2\alpha}\mathbb{1}_{\left\{S_{c(i)}^{1+\alpha}\le K_n\right\}}\right]\le\frac{2}{\varepsilon n^{4/3}}\sum_{i=1}^{tn^{2/3}}K_n^{2\alpha}\mathbb{E}_S\left[S_{c(i)}^{2}\right].
\end{aligned} \tag{73}$$

Lemma 7 allows us to approximate $\mathbb{E}_S[S_{c(i)}^2]$ uniformly by $\frac{\mathbb{E}[S^{2+\alpha}]}{\mathbb{E}[S^\alpha]}$. Thus, we get

$$\frac{2}{\varepsilon n^{4/3}}\sum_{i=1}^{tn^{2/3}}\left(K_n^{2\alpha}\frac{\mathbb{E}\left[S^{2+\alpha}\right]}{\mathbb{E}[S^\alpha]}+o_{\mathbb{P}}(1)\right)=\frac{tK_n^{2\alpha}}{\varepsilon n^{2/3}}O_{\mathbb{P}}(1), \tag{74}$$

which converges to zero as $n\to\infty$ if and only if $K_n^\alpha/n^{1/3}$ does. We now turn to (72) and apply Doob's L^1 martingale inequality (Klenke 2008, theorem 11.2) to obtain

$$\begin{aligned}
&\mathbb{P}_S\left(\sup_{j\le tn^{2/3}}\left|\sum_{i=1}^{j}\left(S_{c(i)}^{1+\alpha}\mathbb{1}_{\left\{S_{c(i)}^{1+\alpha}> K_n\right\}}-\mathbb{E}_S\left[S_{c(i)}^{1+\alpha}\mathbb{1}_{\left\{S_{c(i)}^{1+\alpha}> K_n\right\}}\mid\mathscr{F}_{i-1}\right]\right)\right|\ge\varepsilon n^{2/3}\right)\\
&\le\frac{1}{\varepsilon n^{2/3}}\mathbb{E}_S\left[\left|\sum_{i=1}^{tn^{2/3}}\left(S_{c(i)}^{1+\alpha}\mathbb{1}_{\left\{S_{c(i)}^{1+\alpha}> K_n\right\}}-\mathbb{E}_S\left[S_{c(i)}^{1+\alpha}\mathbb{1}_{\left\{S_{c(i)}^{1+\alpha}> K_n\right\}}\mid\mathscr{F}_{i-1}\right]\right)\right|\right]\\
&\le\frac{2}{\varepsilon n^{2/3}}\sum_{i=1}^{tn^{2/3}}\mathbb{E}_S\left[S_{c(i)}^{1+\alpha}\mathbb{1}_{\left\{S_{c(i)}^{1+\alpha}> K_n\right\}}\right]\le\frac{2}{\varepsilon n^{2/3}}\sum_{i=1}^{tn^{2/3}}\mathbb{E}_S\left[S_{c(1)}^{1+\alpha}\mathbb{1}_{\left\{S_{c(1)}^{1+\alpha}> K_n\right\}}\right]\left(1+O_{\mathbb{P}_S}(1)\right)\\
&=\frac{2t}{\varepsilon}\mathbb{E}_S\left[S_{c(1)}^{1+\alpha}\mathbb{1}_{\left\{S_{c(1)}^{1+\alpha}> K_n\right\}}\right]\left(1+O_{\mathbb{P}_S}(1)\right)=o_{\mathbb{P}}(1).
\end{aligned} \tag{75}$$

We have used Lemma 7 in the second inequality, and Lemma 4 with $f(x)=x^{1+\alpha}\mathbb{1}_{\{x^{1+\alpha}>K_n\}}$ in the third. The rightmost term in (75) is $o_{\mathbb{P}}(1)$ as $n\to\infty$ by the strong law of large numbers. This side of the bound does not impose additional conditions on K_n, so that, if we take $K_n=n^c$, it is sufficient that $c<\frac{1}{3\alpha}$, with the convention that $\frac{1}{0}=\infty$. □

We conclude this section with a technical lemma concerning error terms in the computations of quadratic variations. Denote the density (respectively, distribution function) of a rate λ exponential random variable by $f_E(\cdot)$ (respectively, $F_E(\cdot)$).

Lemma 9. *We have*

$$\mathbb{E}_S\left[\sum_{h,q\in[n]}\left|F_E\left(\frac{S_{c(i)}S_h^\alpha}{n}\right)-\frac{\lambda S_{c(i)}S_h^\alpha}{n}\right|\left|F_E\left(\frac{S_{c(i)}S_q^\alpha}{n}\right)-\frac{\lambda S_{c(i)}S_q^\alpha}{n}\right| \mid \mathcal{F}_{i-1}\right]=o_{\mathbb{P}}(1) \tag{76}$$

uniformly in $i=O(n^{2/3})$.

Proof. Because $|F_E(x)-x|=O(x^2)$, the bound $|\lambda S_{c(i)}S_h^\alpha/n-F_E(S_{c(i)}S_h^\alpha/n)|\le C(S_{c(i)}S_h^\alpha/n)^{1+\varepsilon}$ holds almost surely for $0<\varepsilon<1$ and $C>0$, which gives

$$\lambda^2\sum_{h,q\in[n]}\mathbb{E}_S\left[\left(\frac{S_{c(i)}S_h^\alpha}{n}\right)^{1+\varepsilon}\left(\frac{S_q^\alpha S_{c(i)}}{n}\right)^{1+\varepsilon} \mid \mathcal{F}_{i-1}\right]=\frac{\lambda^2}{n^{2+2\varepsilon}}\sum_{h,q\in[n]}\mathbb{E}_S\left[S_{c(i)}^{2+2\varepsilon}\mid \mathcal{F}_{i-1}\right]S_h^{\alpha(1+\varepsilon)}S_q^{\alpha(1+\varepsilon)}. \tag{77}$$

Therefore,

$$\begin{aligned}\lambda^2\sum_{h,q\in[n]}&\mathbb{E}_S\left[\left(\frac{S_{c(i)}S_h^\alpha}{n}\right)^{1+\varepsilon}\left(\frac{S_q^\alpha S_{c(i)}}{n}\right)^{1+\varepsilon} \mid \mathcal{F}_{i-1}\right]\\ &\le\frac{\lambda^2}{n^{2+2\varepsilon}}\max_{j\in[n]}S_j^{2\varepsilon}\mathbb{E}_S\left[S_{c(i)}^2\mid\mathcal{F}_{i-1}\right]\sum_{h,q\in[n]}S_h^{\alpha(1+\varepsilon)}S_q^{\alpha(1+\varepsilon)}\\ &\le\frac{\lambda^2\mathbb{E}\left[S^{2+\alpha}\right]}{\mathbb{E}[S^\alpha]}\frac{\max_{j\in[n]}S_j^{2\varepsilon}}{n^{2\varepsilon}}\frac{1}{n^2}\sum_{h,q\in[n]}S_h^{\alpha(1+\varepsilon)}S_q^{\alpha(1+\varepsilon)}+o_{\mathbb{P}}(1),\end{aligned} \tag{78}$$

where in the last step we used Lemma 7. Because $\mathbb{E}[S^{2+\alpha}]<\infty$, by Lemma 2, $\max_{j\in[n]}S_j^{2\varepsilon}=o_{\mathbb{P}}(n^{2\varepsilon/(2+\alpha)})$. The rightmost term in (78) then tends to zero as n tends to infinity as long as $0<\varepsilon<\min\{1,2/\alpha\}$. □

4. Proving the Scaling Limit

We first establish some preliminary estimates on $N_n(\cdot)$ that will be crucial for the proof of convergence. We will upper bound the process $N_n(\cdot)$ by a simpler process $N_n^u(\cdot)$ in such a way that the increments of $N_n^u(\cdot)$ almost surely dominate the increments of $N_n(\cdot)$. We also show that, after rescaling, $N_n^u(\cdot)$ converges in distribution to $W(\cdot)$. In fact, we introduce the upper bound $N_n^U(\cdot)$ to deal with the complicated index set for the summation in (5). The difficulty arises as follows: in order to estimate $N_n(\cdot)$, one has to estimate $A_n(\cdot)$. To do this, one has to separately (uniformly) bound each element in the sum, and also estimate the number of elements in the sum. The first goal is accomplished, for example, through Lemma 7, whereas for the second the crude upper bound, n is too loose. However, estimating $|\nu_k|$ requires an estimate on $N_n(\cdot)$ itself, as (6) shows. To solve this circularity, we introduce a bootstrap argument: first, we upper bound $N_n(\cdot)$ and we obtain estimates on the upper bound; from this follows an estimate on $|\nu_k|$, and this in turn allows us to estimate $N_n(\cdot)$.

This technique can be applied to solve a recently found technical issue in the proof of the main result of Bhamidi et al. (2010). Bhamidi et al. (2010) prove convergence of a process that upper bounds the exploration process of the graph. Therefore, their main result is analogous to Theorem 3. However, a further step is required to complete the proof of convergence of the exploration process, and this is provided by our approach.

The process $N_n^u(\cdot)$ is defined as $N_n^u(0)=N_n(0)$, and

$$N_n^u(k)=N_n^u(k-1)+A_n^u(k)-1, \tag{79}$$

where

$$A_n^u(k)=\sum_{i\notin\mathfrak{S}_k}\mathbb{1}_{\{T_i\le\tilde{S}_{c(k)}\}}, \tag{80}$$

with

$$T_i\overset{\mathrm{d}}{=}\mathrm{E}_i\left(\lambda S_i^\alpha\right), \tag{81}$$

and where we have used notation (1). Moreover, $\mathrm{E}_i(c)$ denotes a family of exponential random variables with mean $1/c$ independent across i. The process $N_n^u(\cdot)$ can be interpreted as follows. Each customer is replaced by an ON/OFF source of arriving customers. The source is initially ON and is turned OFF as soon as the first customer it generated has been served. Once a source is OFF, it remains in that state indefinitely. We couple the processes $N_n(\cdot)$ and $N_n^u(\cdot)$ as follows. Consider a sequence of arrival times $(T_i)_{i=1}^{\infty}$ and of service times $(S_i)_{i=1}^{\infty}$ and then define $A_n(\cdot)$ as (5) and $A_n^u(\cdot)$ as (80) With this coupling we have that, almost surely,

$$A_n(k) \le A_n^u(k) \qquad \forall\ k \ge 1. \tag{82}$$

Consequently,

$$N_n(k) \le N_n^u(k) \qquad \forall k \ge 0, \tag{83}$$

and

$$Q_n(k) = \phi(N_n)(k) \le \phi\big(N_n^u\big)(k) =: Q_n^u(k) \qquad \forall k \ge 0, \tag{84}$$

almost surely. Crucially, the complicated set ν_k does not appear in the definition of $N_n^u(\cdot)$. The random variable $A_n^u(\cdot)$ depends instead on the set $\mathfrak{S}_k$.

Although, in general, only the upper bounds (83) and (84) hold, the processes $N_n(\cdot)$ and $N_n^u(\cdot)$ (respectively, $Q_n(\cdot)$ and $Q_n^u(\cdot)$) turn out to be very close to each other. We start by proving results for $N_n^u(\cdot)$ and $Q_n^u(\cdot)$ because they are easier to treat, and only then, we are able to prove that identical results hold for $N_n(\cdot)$ and $Q_n(\cdot)$.

Theorem 3 (Convergence of the Upper Bound)**.** We have

$$n^{-1/3}N_n^u\big(tn^{2/3}\big) \xrightarrow{\mathrm{d}} W(t) \qquad \text{in } (\mathcal{D}, J_1) \text{ as } n \to \infty, \tag{85}$$

where $W(\cdot)$ is the diffusion process in (13). In particular,

$$n^{-1/3}\phi\big(N_n^u\big)\big(tn^{2/3}\big) \xrightarrow{\mathrm{d}} \phi(W)(t) \qquad \text{in } (\mathcal{D}, J_1) \text{ as } n \to \infty. \tag{86}$$

The next section is dedicated to the proof of Theorem 3.

4.1. Convergence of the Upper Bound

We use a classical martingale decomposition followed by a martingale FCLT. The process $N_n^u(\cdot)$ in (79) can be decomposed as $N_n^u(k) = M_n^u(k) + C_n^u(k)$, where $M_n^u(\cdot)$ is a martingale and $C_n^u(\cdot)$ is a drift term, as follows:

$$M_n^u(k) = \sum_{i=1}^{k}\big(A_n^u(i) - \mathbb{E}_s\big[A_n^u(i) \mid \mathcal{F}_{i-1}\big]\big), \quad C_n^u(k) = \sum_{i=1}^{k}\big(\mathbb{E}_s\big[A_n^u(i) \mid \mathcal{F}_{i-1}\big] - 1\big). \tag{87}$$

Moreover, $(M_n^u(k))^2$ can be written as $(M_n^u(k))^2 = Z_n^u(k) + B_n^u(k)$ with $Z_n^u(k)$ a martingale and $B_n^u(k)$ the compensator, or quadratic variation, of $M_n^u(k)$ given by

$$B_n^u(k) = \sum_{i=1}^{k}\Big(\mathbb{E}_s\Big[\big(A_n^u(i)\big)^2 \mid \mathcal{F}_{i-1}\Big] - \mathbb{E}_s\big[A_n^u(i) \mid \mathcal{F}_{i-1}\big]^2\Big). \tag{88}$$

In order to prove convergence of $N_n^u(\cdot)$, we separately prove convergence of $C_n^u(\cdot)$ and of $M_n^u(\cdot)$. We prove the former directly and the latter by applying the following martingale FCLT (Ethier and Kurtz 1989, theorem 7.1.4).

Theorem 4. *Let $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ be an increasing filtration and $\{\bar{M}_n\}_{n\in\mathbb{N}}$ be a sequence of continuous-time, real-valued, square-integrable martingales, each with respect to $\mathcal{F}_n$, such that $\bar{M}_n(0) = 0$. Assume that $\bar{V}_n(\cdot)$, the predictable quadratic variation process associated with $\bar{M}_n(\cdot)$, and $\bar{M}_n(\cdot)$ satisfy the following conditions:*

a. $\bar{V}_n(t) \xrightarrow{\mathbb{P}} \sigma^2 t, \qquad \forall t \in \mathbb{R}^+$;
b. $\lim_{n\to\infty} \mathbb{E}[\sup_{t\le\bar{t}} |\bar{V}_n(t) - \bar{V}_n(t^-)|] = 0, \qquad \forall \bar{t} \in \mathbb{R}^+$; *and*
c. $\lim_{n\to\infty} \mathbb{E}[\sup_{t\le\bar{t}} |\bar{M}_n(t) - \bar{M}_n(t^-)|^2] = 0, \qquad \forall \bar{t} \in \mathbb{R}^+$.

Then, as $n \to \infty$, $\bar{M}_n(\cdot)$ converges in distribution in $\mathcal{D}([0,\infty))$ to a centered Brownian motion with variance $\sigma^2 t$.

For this, we need to verify the following conditions:

i. $\sup_{t\le\bar{t}} |n^{-1/3}C_n^u(tn^{2/3}) - \beta t + \lambda \frac{\mathbb{E}[S^{1+2\alpha}]}{2\mathbb{E}[S^\alpha]} t^2| \xrightarrow{\mathbb{P}} 0, \qquad \forall \bar{t} \in \mathbb{R}^+$;
ii. $n^{-2/3}B_n^u(tn^{2/3}) \xrightarrow{\mathbb{P}} \sigma^2 t, \qquad \forall t \in \mathbb{R}^+$;

iii. $\lim_{n\to\infty} n^{-2/3}\mathbb{E}_s[\sup_{t\le\bar t}|B_n^u(tn^{2/3}) - B_n^u(tn^{2/3}-)|] = 0, \qquad \forall \bar t \in \mathbb{R}^+$; and
iv. $\lim_{n\to\infty} n^{-2/3}\mathbb{E}_s[\sup_{t\le\bar t}|M_n^u(tn^{2/3}) - M_n^u(tn^{2/3}-)|^2] = 0, \qquad \forall \bar t \in \mathbb{R}^+$.

4.1.1. Proof of (i) for the Upper Bound. First we obtain an explicit expression for $\mathbb{E}[A_n^u(i) \mid \mathscr{F}_{i-1}]$, as

$$\begin{aligned}
\mathbb{E}_s\big[A_n^u(i) \mid \mathscr{F}_{i-1}\big] &= \sum_{j\notin\mathfrak{S}_{i-1}} \mathbb{P}_S\big(c(i)=j \mid \mathscr{F}_{i-1}\big) \sum_{l\notin\mathfrak{S}_{i-1}\cup\{j\}} F_{\scriptscriptstyle E}\left(\frac{c_{n,\beta}S_jS_l^\alpha}{n}\right)\\
&= \sum_{j\notin\mathfrak{S}_{i-1}} \mathbb{P}_S\big(c(i)=j \mid \mathscr{F}_{i-1}\big) \sum_{l=1}^{n} \frac{c_{n,\beta}\lambda S_jS_l^\alpha}{n}\\
&\quad - \sum_{j\notin\mathfrak{S}_{i-1}} \mathbb{P}_S\big(c(i)=j \mid \mathscr{F}_{i-1}\big) \sum_{l\in\mathfrak{S}_{i-1}\cup\{j\}} \frac{c_{n,\beta}\lambda S_jS_l^\alpha}{n}\\
&\quad + \sum_{j\notin\mathfrak{S}_{i-1}} \mathbb{P}_S\big(c(i)=j \mid \mathscr{F}_{i-1}\big) \sum_{l\notin\mathfrak{S}_{i-1}\cup\{j\}} \left(F_{\scriptscriptstyle E}\left(\frac{c_{n,\beta}S_jS_l^\alpha}{n}\right) - \frac{c_{n,\beta}\lambda S_jS_l^\alpha}{n}\right).
\end{aligned} \tag{89}$$

The third term is an error term. Indeed, for some $\zeta_n \in [0, S_{c(i)}S_l/n]$,

$$\begin{aligned}
&\mathbb{E}_s\left[\left|\sum_{l\notin\mathfrak{S}_{i-1}\cup\{j\}} F_{\scriptscriptstyle E}\left(\frac{S_{c(i)}S_l^\alpha}{n}\right) - \frac{\lambda S_{c(i)}S_l^\alpha}{n}\right|\middle|\mathscr{F}_{i-1}\right]\\
&\le \sum_{l=1}^{n} \mathbb{E}_s\left[\left|F_{\scriptscriptstyle E}\left(\frac{S_{c(i)}S_l^\alpha}{n}\right) - \lambda\frac{S_{c(i)}S_l^\alpha}{n}\right|\middle|\mathscr{F}_{i-1}\right]\\
&= \frac{1}{2n^2}\mathbb{E}_s\left[\left|F_{\scriptscriptstyle E}''(\zeta_n)S_{c(i)}^2\right|\middle|\mathscr{F}_{i-1}\right]\sum_{l=1}^{n} S_l^{2\alpha} \le \frac{\lambda^2}{2n^2}\mathbb{E}_s\left[S_{c(i)}^2\middle|\mathscr{F}_{i-1}\right]\sum_{l=1}^{n} S_l^{2\alpha},
\end{aligned} \tag{90}$$

because $|F_{\scriptscriptstyle E}''(x)| \le \lambda^2$ for all $x \ge 0$. By Lemma 7, this can be bounded by

$$\frac{\lambda^2}{2n^2}(C_n + o_{\mathbb{P}}(1))\sum_{l\in[n]} S_l^{2\alpha}, \tag{91}$$

where C_n is bounded w.h.p. and the $o_{\mathbb{P}}(1)$ term is uniform in $i = O(n^{2/3})$. Therefore, the third term in (89) is $o_{\mathbb{P}}(n^{-1/3})$. Inserting this into (89) and splitting the summation in the second term gives

$$\begin{aligned}
\mathbb{E}_s\big[A_n^u(i) \mid \mathscr{F}_{i-1}\big] - 1 &= \sum_{j\notin\mathfrak{S}_{i-1}} \mathbb{P}_S\big(c(i)=j \mid \mathscr{F}_{i-1}\big)c_{n,\beta}\lambda S_j\frac{\sum_{l=1}^n S_l^\alpha}{n}\\
&\quad - \sum_{j\notin\mathfrak{S}_{i-1}} \mathbb{P}_S\big(c(i)=j \mid \mathscr{F}_{i-1}\big) \sum_{l\in\mathfrak{S}_{i-1}} \frac{c_{n,\beta}\lambda S_jS_l^\alpha}{n}\\
&\quad - c_{n,\beta}\lambda \sum_{j\notin\mathfrak{S}_{i-1}} \mathbb{P}_S\big(c(i)=j \mid \mathscr{F}_{i-1}\big)\frac{S_j^{1+\alpha}}{n} - 1 + o_{\mathbb{P}}\big(n^{-1/3}\big)\\
&= \left(c_{n,\beta}\lambda\frac{\sum_{l\in[n]} S_l^\alpha}{n}\mathbb{E}\big[S_{c(i)} \mid \mathscr{F}_{i-1}\big] - 1\right) - c_{n,\beta}\mathbb{E}_s\big[S_{c(i)} \mid \mathscr{F}_{i-1}\big]\sum_{l\in\mathfrak{S}_{i-1}}\lambda\frac{S_l^\alpha}{n}\\
&\quad - c_{n,\beta}\frac{\lambda}{n}\mathbb{E}_s\left[S_{c(i)}^{1+\alpha} \mid \mathscr{F}_{i-1}\right] + o_{\mathbb{P}}\big(n^{-1/3}\big).
\end{aligned} \tag{92}$$

Let us focus on the first term of (92). Using $\frac{c}{a-b} = \frac{c}{a} + \frac{c}{a-b}\frac{b}{a}$, with $a = \sum_{l\in[n]} S_l^\alpha$ and $b = \sum_{l\in\mathfrak{S}_{i-1}} S_l^\alpha$, we get

$$\begin{aligned}
&c_{n,\beta}\lambda\frac{\sum_{l=1}^n S_l^\alpha}{n}\mathbb{E}_s\big[S_{c(i)} \mid \mathscr{F}_{i-1}\big] - 1\\
&= c_{n,\beta}\lambda\frac{\sum_{l\in[n]} S_l^\alpha}{n}\sum_{j\notin\mathfrak{S}_{i-1}}\frac{S_j^{1+\alpha}}{\sum_{l\in[n]} S_l^\alpha} - 1 + c_{n,\beta}\lambda\frac{\sum_{l\in[n]} S_l^\alpha}{n}\sum_{j\notin\mathfrak{S}_{i-1}}\frac{S_j^{1+\alpha}}{\sum_{l\notin\mathfrak{S}_{i-1}} S_l^\alpha}\frac{\sum_{s\in\mathfrak{S}_{i-1}} S_s^\alpha}{\sum_{l\in[n]} S_l^\alpha}\\
&= \left(c_{n,\beta}\frac{\lambda}{n}\sum_{j\notin\mathfrak{S}_{i-1}} S_j^{1+\alpha} - 1\right) + c_{n,\beta}\mathbb{E}_s\big[S_{c(i)} \mid \mathscr{F}_{i-1}\big]\sum_{s\in\mathfrak{S}_{i-1}}\lambda\frac{S_s^\alpha}{n}.
\end{aligned} \tag{93}$$

The second term in (92) is canceled out by the rightmost term in (93). We emphasize that this cancellation is what makes the analysis of $N_n^u(\cdot)$ considerably easier than the analysis of $N_n(\cdot)$.

Finally, Lemma 7 implies that the third term in (92) is $o_{\mathbb{P}}(n^{-1/3})$. Piecing together the computations, we obtain

$$\begin{aligned}\mathbb{E}_s\big[A_n^u(i)\mid\mathscr{F}_{i-1}\big]-1&=c_{n,\beta}\frac{\lambda}{n}\sum_{j\notin\mathfrak{S}_{i-1}}S_j^{1+\alpha}-1+o_{\mathbb{P}}\big(n^{-1/3}\big)\\&=\left(c_{n,\beta}\frac{\lambda}{n}\sum_{j=1}^{n}S_j^{1+\alpha}-1\right)-c_{n,\beta}\frac{\lambda}{n}\sum_{j\in\mathfrak{S}_{i-1}}S_j^{1+\alpha}+o_{\mathbb{P}}\big(n^{-1/3}\big)\\&=\left(c_{n,\beta}\frac{\lambda}{n}\sum_{j=1}^{n}S_j^{1+\alpha}-1\right)-c_{n,\beta}\frac{\lambda}{n}\sum_{j=1}^{i-1}S_{c(j)}^{1+\alpha}+o_{\mathbb{P}}\big(n^{-1/3}\big),\end{aligned}\tag{94}$$

and the $o_{\mathbb{P}}(n^{-1/3})$ term is uniform in $i=O(n^{2/3})$. We are now able to compute

$$\begin{aligned}n^{-1/3}C_n^u\big(tn^{2/3}\big)&=n^{-1/3}\sum_{i=1}^{tn^{2/3}}\big(\mathbb{E}_s\big[A_n^u(i)\mid\mathscr{F}_{i-1}\big]-1\big)\\&=tn^{1/3}\left(c_{n,\beta}\frac{\lambda}{n}\sum_{j=1}^{n}S_j^{1+\alpha}-1\right)-c_{n,\beta}\frac{\lambda}{n^{4/3}}\sum_{i=1}^{tn^{2/3}}\sum_{j=1}^{i-1}S_{c(j)}^{1+\alpha}+o_{\mathbb{P}}(1).\end{aligned}\tag{95}$$

Because $\mathbb{E}[(S^{1+\alpha})^{\frac{2+\alpha}{1+\alpha}}]<\infty$, by the Marcinkiewicz and Zygmund theorem (Durrett 2010, theorem 2.5.8), if $\alpha\in(0,1]$,

$$c_{n,\beta}\frac{\lambda}{n}\sum_{j=1}^{n}S_j^{1+\alpha}=c_{n,\beta}\lambda\mathbb{E}\big[S^{1+\alpha}\big]+o_{\mathbb{P}}\Big(n^{-\frac{1}{2+\alpha}}\Big)=1+\beta n^{-1/3}+o_{\mathbb{P}}\Big(n^{-\frac{1}{2+\alpha}}\Big).\tag{96}$$

For $\alpha=0$, by a similar result (Durrett 2010, theorem 2.5.7), for all $\varepsilon>0$,

$$\frac{1}{n}\sum_{j=1}^{n}S_j=\mathbb{E}[S]+o_{\mathbb{P}}\big(n^{-1/2}\log(n)^{1/2+\varepsilon}\big).\tag{97}$$

Summarizing the two results, for any $\alpha\in[0,1]$ we have

$$tn^{1/3}\left(c_{n,\beta}\frac{\lambda}{n}\sum_{j=1}^{n}S_j^{1+\alpha}-1\right)=t\big(\beta+o_{\mathbb{P}}(1)\big).\tag{98}$$

Because this expression is monotone in t, we also have

$$\sup_{t\le T}\left|tn^{1/3}\left(c_{n,\beta}\frac{\lambda}{n}\sum_{j=1}^{n}S_j^{1+\alpha}-1\right)-\beta t\right|\xrightarrow{\mathbb{P}}0,\tag{99}$$

so that, for $\alpha\in[0,1]$,

$$n^{-1/3}C_n^u\big(tn^{2/3}\big)=\beta t-c_{n,\beta}\frac{\lambda}{n^{4/3}}\sum_{i=1}^{tn^{2/3}}\sum_{j=1}^{i-1}S_{c(j)}^{1+\alpha}+o_{\mathbb{P}}(1).\tag{100}$$

Because $c_{n,\beta}=1+O(n^{-1/3})$, the second term in (100) converges uniformly to $-t^2\lambda\mathbb{E}[S^{1+2\alpha}]/2\mathbb{E}[S^\alpha]$ by Lemma 8.

4.1.2. Proof of (ii) for the Upper Bound.

Rewrite $B_n^u(k)$, for $k=O(n^{2/3})$, as

$$\begin{aligned}B_n^u(k)&=\sum_{i=1}^{k}\Big(\mathbb{E}_s\big[A_n^u(i)^2\mid\mathscr{F}_{i-1}\big]-\mathbb{E}_s\big[A_n^u(i)\mid\mathscr{F}_{i-1}\big]^2\Big)\\&=\sum_{i=1}^{k}\big(\mathbb{E}_s\big[A_n^u(i)^2\mid\mathscr{F}_{i-1}\big]-1\big)+O_{\mathbb{P}}\big(kn^{-1/3}\big),\end{aligned}\tag{101}$$

where we have used the asymptotics for $\mathbb{E}_s[A^u_n(i) \mid \mathscr{F}_{i-1}]$ in (94). Moreover, we can compute $\mathbb{E}_s[A^u_n(i)^2 \mid \mathscr{F}_{i-1}]$ as

$$\begin{aligned}\mathbb{E}_s\big[A^u_n(i)^2 \mid \mathscr{F}_{i-1}\big] &= \mathbb{E}_s\Bigg[\Bigg(\sum_{h\notin\mathfrak{S}_i}\mathbb{1}_{\{T_h\le c_{n,\beta}S_{c(i)}S_h/n\}}\Bigg)^2 \mid \mathscr{F}_{i-1}\Bigg]\\ &= \mathbb{E}_s\big[A^u_n(i) \mid \mathscr{F}_{i-1}\big] + \mathbb{E}_s\Bigg[\sum_{h,q\notin\mathfrak{S}_i}\mathbb{1}_{\{T_h\le c_{n,\beta}S_{c(i)}S_h/n\}}\mathbb{1}_{\{T_q\le c_{n,\beta}S_{c(i)}S_q/n\}} \mid \mathscr{F}_{i-1}\Bigg].\end{aligned} \tag{102}$$

By (94), $\mathbb{E}_s[A_n(i) \mid \mathscr{F}_{i-1}] = 1 + o_{\mathbb{P}}(1)$, uniformly in $i = O(n^{2/3})$, so that (101) simplifies to

$$B_n(k) = \sum_{i=1}^{k}\mathbb{E}_s\Bigg[\sum_{h,q\notin\mathfrak{S}_i}\mathbb{1}_{\{T_h\le c_{n,\beta}S_{c(i)}S^\alpha_h/n\}}\mathbb{1}_{\{T_q\le c_{n,\beta}S_{c(i)}S^\alpha_q/n\}} \mid \mathscr{F}_{i-1}\Bigg] + O_{\mathbb{P}}\big(kn^{-1/3}\big). \tag{103}$$

We then focus on the second term in (102), which we compute as

$$\begin{aligned}&\sum_{\substack{h,q\notin\mathfrak{S}_i\\ h\neq q}}\mathbb{E}_s\Big[\mathbb{1}_{\{T_h\le c_{n,\beta}S_{c(i)}S^\alpha_h/n\}}\mathbb{1}_{\{T_q\le c_{n,\beta}S_{c(i)}S^\alpha_q/n\}} \mid \mathscr{F}_{i-1}\Big]\\ &= \sum_{j\notin\mathfrak{S}_{i-1}}\mathbb{P}_S\big(c(i) = j \mid \mathscr{F}_{i-1}\big)\sum_{\substack{h,q\notin\mathfrak{S}_{i-1}\cup\{j\}\\ h\neq q}}\mathbb{E}_s\Big[\mathbb{1}_{\{T_h\le c_{n,\beta}S_jS^\alpha_h/n\}}\mathbb{1}_{\{T_q\le c_{n,\beta}S_jS^\alpha_q/n\}} \mid \mathscr{F}_{i-1}\Big].\end{aligned} \tag{104}$$

By Lemma 9,

$$\begin{aligned}\text{r.h.s. (4.26)} &= \sum_{j\notin\mathfrak{S}_{i-1}}\mathbb{P}_S\big(c(i) = j \mid \mathscr{F}_{i-1}\big)\sum_{\substack{h,q\notin\mathfrak{S}_{i-1}\cup\{j\}\\ h\neq q}}\left(\frac{c^2_{n,\beta}\lambda^2S^2_jS^\alpha_hS^\alpha_q}{n^2} + o_{\mathbb{P}}\big(n^{-2}\big)\right)\\ &= \big(c_{n,\beta}\lambda\big)^2\mathbb{E}_s\Big[S^2_{c(i)} \mid \mathscr{F}_{i-1}\Big]\frac{1}{n^2}\sum_{\substack{h,q\notin\mathfrak{S}_{i-1}\cup\{c(i)\}\\ h\neq q}}S^\alpha_hS^\alpha_q + o_{\mathbb{P}}(1)\\ &= \frac{\big(c_{n,\beta}\lambda\big)^2}{n^2}\mathbb{E}_s\Big[S^2_{c(i)} \mid \mathscr{F}_{i-1}\Big]\sum_{1\le h,q\le n}S^\alpha_hS^\alpha_q\\ &\quad - \frac{\big(c_{n,\beta}\lambda\big)^2}{n^2}\mathbb{E}_s\Bigg[S^2_{c(i)}\sum_{\substack{h,q\in\mathfrak{S}_{i-1}\cup\{c(i)\}\\ \cup\{h=q\}}}S^\alpha_hS^\alpha_q \mid \mathscr{F}_{i-1}\Bigg] + o_{\mathbb{P}}(1).\end{aligned}$$

The leading contribution to $B^u_n(k)$ is given by the first term, wheras the second term is an error term by Lemma 5. We have shown that $B^u_n(\cdot)$ can be rewritten as

$$B^u_n(k) = \left(\frac{\lambda}{n}\sum_{h\in[n]}S^\alpha_h\right)^2\sum_{i=1}^{k}\mathbb{E}_s\Big[S^2_{c(i)} \mid \mathscr{F}_{i-1}\Big] + o_{\mathbb{P}}(k). \tag{105}$$

Thus,

$$n^{-2/3}B^u_n\big(n^{2/3}u\big)\xrightarrow{\mathbb{P}}\lambda^2\mathbb{E}[S^\alpha]\mathbb{E}\big[S^{2+\alpha}\big]u, \tag{106}$$

which concludes the proof of (ii).

4.1.3. Proof of (iii) for the Upper Bound. The jumps of $B^u_n(k)$ are given by

$$\begin{aligned}B^u_n(i) - B^u_n(i-1) &= \mathbb{E}_s\big[A^u_n(i)^2 \mid \mathscr{F}_{i-1}\big] - \mathbb{E}_s\big[A^u_n(i) \mid \mathscr{F}_{i-1}\big]^2\\ &= \mathbb{E}_s\Bigg[\sum_{\substack{h,q\notin\mathfrak{S}_i\\ h\neq q}}\mathbb{1}_{\{T_h\le c_{n,\beta}S_{c(i)}S^\alpha_h/n\}}\mathbb{1}_{\{T_q\le c_{n,\beta}S_{c(i)}S^\alpha_q/n\}} \mid \mathscr{F}_{i-1}\Bigg]\\ &\quad + \Big(\mathbb{E}_s\big[A^u_n(i) \mid \mathscr{F}_{i-1}\big] - \mathbb{E}_s\big[A^u_n(i) \mid \mathscr{F}_{i-1}\big]^2\Big).\end{aligned} \tag{107}$$

Because $\mathbb{E}_s[A_n^u(i) \mid \mathcal{F}_{i-1}] = 1 + O_{\mathbb{P}}(n^{-1/3})$ for $i = O_{\mathbb{P}}(n^{2/3})$ by (94), the second term is of order $O_{\mathbb{P}}(n^{-1/3})$, uniformly in $i = O_{\mathbb{P}}(n^{2/3})$. The first term was computed in (104). Therefore,

$$\begin{aligned} &B_n^u(i) - B_n^u(i-1) \\ &\quad = \frac{(c_{n,\beta}\lambda)^2}{n^2}\mathbb{E}_s\Big[S_{c(i)}^2 \mid \mathcal{F}_{i-1}\Big]\sum_{h,q\in[n]} S_h^\alpha S_q^\alpha - \frac{(c_{n,\beta}\lambda)^2}{n^2}\mathbb{E}_s\Bigg[S_{c(i)}^2 \sum_{\substack{h,q\in\mathfrak{S}_{i-1}\cup\{c(i)\} \\ \cup\{h=q\}}} S_h^\alpha S_q^\alpha \mid \mathcal{F}_{i-1}\Bigg] + o_{\mathbb{P}}(1) \\ &\quad \le \frac{(c_{n,\beta}\lambda)^2}{n^2}\mathbb{E}_s\Big[S_{c(i)}^2 \mid \mathcal{F}_{i-1}\Big]\sum_{h,q\in[n]} S_h^\alpha S_q^\alpha. \end{aligned} \tag{108}$$

After rescaling and taking the expectation, we obtain the bound

$$n^{-2/3}\mathbb{E}_s\Bigg[\sup_{i\le\bar{t}n^{2/3}}\big|B_n^u(i) - B_n^u(i-1)\big|\Bigg] \le \frac{(c_{n,\beta}\lambda)^2}{n^{2/3}}\mathbb{E}_s\Bigg[\sup_{i\le\bar{t}n^{2/3}} S_{c(i)}^2\Bigg]\Bigg(\frac{1}{n}\sum_{h,q\in[n]} S_h^\alpha\Bigg)^2. \tag{109}$$

Lemma 10. *If* $\mathbb{E}[S^{2+\alpha}] < \infty$,

$$\mathbb{E}_s\Bigg[\sup_{k\le tn^{2/3}} S_{c(k)}^2\Bigg] = o_{\mathbb{P}}\big(n^{2/3}\big), \tag{110}$$

for each $t \in \mathbb{R}^+$

Proof. For $\varepsilon > 0$, split the expectation as

$$\frac{1}{n^{2/3}}\mathbb{E}_s\Bigg[\Bigg(\sup_{k\le tn^{2/3}} S_{c(k)}\Bigg)^2\Bigg] \le \frac{1}{n^{2/3}}\mathbb{E}_s\Bigg[\sup_{k\le tn^{2/3}} S_{c(k)}^2 \mathbb{1}_{\{S_{c(k)}>\varepsilon n^{1/3}\}}\Bigg] + \varepsilon^2. \tag{111}$$

We bound the expected value in the first term as

$$\begin{aligned} \mathbb{E}_s\Bigg[\sup_{k\le tn^{2/3}} S_{c(k)}^2 \mathbb{1}_{\{S_{c(k)}>\varepsilon n^{1/3}\}}\Bigg] &\le \sum_{k\le tn^{2/3}} \mathbb{E}_s\Big[S_{c(k)}^2 \mathbb{1}_{\{S_{c(k)}>\varepsilon n^{1/3}\}}\Big] \\ &\le n^{2/3}t\mathbb{E}_s\Big[S_{c(1)}^2 \mathbb{1}_{\{S_{c(1)}>\varepsilon n^{1/3}\}}\Big]\big(1 + O_{\mathbb{P}_S}(1)\big), \end{aligned} \tag{112}$$

where we used Lemma 4 with $f(x) = x^2\mathbb{1}_{\{x>\varepsilon n^{1/3}\}}$. Computing the expectation explicitly we get

$$\begin{aligned} t\mathbb{E}_s\Big[S_{c(1)}^2 \mathbb{1}_{\{S_{c(1)}>\varepsilon n^{1/3}\}}\Big] &= t\sum_{i\in[n]} S_i^2 \mathbb{1}_{\{S_i>\varepsilon n^{1/3}\}}\mathbb{P}(c(1) = i) \\ &= t\sum_{i\in[n]} S_i^2 \mathbb{1}_{\{S_i>\varepsilon n^{1/3}\}}\frac{S_i^\alpha}{\sum_{j\in[n]} S_j^\alpha}, \end{aligned} \tag{113}$$

so that the lefthand side of (111) is bounded by

$$\frac{t}{\sum_{j\in[n]} S_j^\alpha}\sum_{i\in[n]} S_i^{2+\alpha}\mathbb{1}_{\{S_i>\varepsilon n^{1/3}\}} + \varepsilon^2, \tag{114}$$

which tends to zero as $n \to \infty$ because $\mathbb{E}[S^{2+\alpha}] < \infty$ and $\varepsilon > 0$ is arbitrary.

By Lemma 10 the righthand side of (109) converges to zero, and this concludes the proof.

4.1.4. Proof of (iv) for the Upper Bound. First, we split

$$\begin{aligned}\mathbb{E}_s\left[\sup_{k\le tn^{2/3}}\left(M_n^u(k)-M_n^u(k-1)\right)^2\right] &= \mathbb{E}_s\left[\sup_{k\le tn^{2/3}}\left(A_n^u(k)-\mathbb{E}_s\left[A_n^u(k)\mid\mathcal{F}_{k-1}\right]\right)^2\right]\\ &\le \mathbb{E}_s\left[\sup_{k\le tn^{2/3}}|A_n^u(k)|^2\right]+\mathbb{E}_s\left[\sup_{k\le tn^{2/3}}\mathbb{E}\left[A_n^u(k)\mid\mathcal{F}_{k-1}\right]^2\right]\\ &\le 2\mathbb{E}_s\left[\sup_{k\le tn^{2/3}}|A_n^u(k)|^2\right]. \qquad (115)\end{aligned}$$

We then stochastically dominate $(A_n^u(k))_{k\le tn^{2/3}}$ by a sequence of Poisson processes $(\Pi_k)_{k\le tn^{2/3}}$, according to

$$A_n^u(k)\le\Pi_k\left(c_{n,\beta}S_{c(k)}\sum_{i\in[n]}\frac{S_i^\alpha}{n}\right)=:A_n'(k). \qquad (116)$$

Indeed, if $E_1,E_2,\ldots,E_n$ are exponential random variables with parameters $\lambda_1,\lambda_2,\ldots,\lambda_n$, there exists a coupling with a Poisson process $\Pi(\cdot)$ such that $\sum_{i\le n}\mathbb{1}_{\{E_i\le t\}}\le\Pi(\sum_{i\le n}\lambda_i t)$. The coupling is constructed as follows. Each random variable E_i is coupled with a Poisson process $\Pi^{(i)}$ with intensity λ_i in such a way that $\mathbb{1}_{\{E_i\le t\}}\le\Pi^{(i)}(\lambda_i t)$. Moreover, by basic properties of the Poisson process $\sum_{i\le n}\Pi^{(i)}(\lambda_i t)\overset{\mathrm{d}}{=}\Pi(\sum_{i\le n}\lambda_i t)$.

We bound (115) via martingale techniques. First, we decompose it as

$$\begin{aligned}n^{-2/3}\mathbb{E}_s\left[\sup_{k\le tn^{2/3}}|A_n^u(k)|^2\right] &\le 2n^{-2/3}\mathbb{E}_s\left[\left(\sup_{k\le tn^{2/3}}\left|A_n'(k)-c_{n,\beta}S_{c(k)}\sum_{i\in[n]}\frac{S_i^\alpha}{n}\right|\right)^2\right]\\ &\quad+2n^{-2/3}\mathbb{E}_s\left[\left(c_{n,\beta}\sup_{k\le tn^{2/3}}S_{c(k)}\sum_{i\in[n]}\frac{S_i^\alpha}{n}\right)^2\right]. \qquad (117)\end{aligned}$$

Applying Doob's L^2 martingale inequality (Klenke 2008, theorem 11.2) to the first term, we see that it converges to zero, because

$$\begin{aligned}n^{-2/3}\mathbb{E}_s\left[\left(\sup_{k\le tn^{2/3}}\left|A_n'(k)-S_{c(k)}\sum_{i\in[n]}\frac{S_i^\alpha}{n}\right|\right)^2\right] &\le 4n^{-2/3}\mathbb{E}_s\left[\left|A_n'(tn^{2/3})-S_{c(tn^{2/3})}\sum_{i\in[n]}\frac{S_i^\alpha}{n}\right|^2\right]\\ &= 4n^{-2/3}\mathbb{E}_s\left[S_{c(tn^{2/3})}\sum_{i\in[n]}\frac{S_i^\alpha}{n}\right]. \qquad (118)\end{aligned}$$

The last equality follows from the expression for the variance of a Poisson random variable. The rightmost term converges to zero by Lemma 7. We now bound the second term in (117), as

$$n^{-2/3}\mathbb{E}_s\left[\left(\sup_{k\le tn^{2/3}}S_{c(k)}\sum_{i\in[n]}\frac{S_i^\alpha}{n}\right)^2\right]=n^{-2/3}\left(\sum_{i\in[n]}\frac{S_i^\alpha}{n}\right)^2\mathbb{E}_s\left[\left(\sup_{k\le tn^{2/3}}S_{c(k)}\right)^2\right]. \qquad (119)$$

By Lemma 10, the righthand side of (119) converges to zero, concluding the proof of (iv).

4.2. Convergence of the Scaling Limit

As a consequence of (84) and Theorem 3, we have that $Q_n(k)=O_{\mathbb{P}}(n^{1/3})$ for $k=O(n^{2/3})$. In fact, $n^{-1/3}Q_n(k)$ is tight when $k=O(n^{2/3})$, as the following lemma shows.

Lemma 11. *Fix $\bar{t}>0$. The sequence $n^{-1/3}\sup_{t\le\bar{t}}Q_n(tn^{2/3})$ is tight.*

Proof. The supremum function $f(\cdot)\mapsto\sup_{t\le\bar{t}}f(t)$ is continuous in $(\mathcal{D},J_1)$ by Whitt (2002, theorem 13.4.1). In particular,

$$n^{-1/3}\sup_{t\le\bar{t}}Q_n^u(tn^{2/3})\xrightarrow{\mathrm{d}}\sup_{t\le\bar{t}}W(t),\qquad\text{in }(\mathcal{D},J_1). \qquad (120)$$

Because $Q_n(k) \le Q_n^u(k)$, the conclusion follows. □

As an immediate consequence of (6) and Lemma 11, we have the following important corollary. Recall that ν_i is the set of customers who have left the system or are in the queue at the beginning of the ith service, so that $|\nu_i| = i + Q_n(i)$.

Corollary 1. *As $n \to \infty$,*

$$|\nu_i| = i + o_{\mathbb{P}}(i), \qquad \text{uniformly in } i = O_{\mathbb{P}}\big(n^{2/3}\big). \tag{121}$$

Intuitively, this implies that the main contribution to the downward drift in the queue-length process comes from the customers that have left the system and not from the customers in the queue. Alternatively, the order of magnitude of the queue length, that is $n^{1/3}$, is negligible with respect to the order of magnitude of the customers who have left the system, which is $n^{2/3}$.

In order to prove Theorem 1, we proceed as in the proof of Theorem 3, but we now need to deal with the more complicated drift term. As before, we decompose $N_n(k) = M_n(k) + C_n(k)$, where

$$\begin{aligned} M_n(k) &= \sum_{i=1}^{k} \big(A_n(i) - \mathbb{E}_s[A_n(i) \mid \mathcal{F}_{k-1}]\big), \\ C_n(k) &= \sum_{i=1}^{k} \big(\mathbb{E}_s[A_n(i) \mid \mathcal{F}_{k-1}] - 1\big), \\ B_n(k) &= \sum_{i=1}^{k} \Big(\mathbb{E}_s\big[A_n(i)^2 \mid \mathcal{F}_{i-1}\big] - \mathbb{E}_s[A_n(i) \mid \mathcal{F}_{i-1}]^2\Big). \end{aligned} \tag{122}$$

As before, we separately prove the convergence of the drift $C_n(k)$ and of the martingale $M_n(k)$ by verifying conditions (i)–(iv) in Section 4.1. Verifying (i) proves to be the most challenging task, whereas the estimates for (ii)–(iv) in Section 4.1 carry over without further complications.

4.2.1. Proof of (i) for the Embedded Queue. By expanding $\mathbb{E}_s[A_n(i) \mid \mathcal{F}_{i-1}] - 1$ as in (92), we get

$$\begin{aligned} \mathbb{E}_s[A_n(i) \mid \mathcal{F}_{i-1}] - 1 = {} & \left(c_{n,\beta}\lambda \frac{\sum_{l=1}^{n} S_l^{\alpha}}{n} \mathbb{E}_s\big[S_{c(i)} \mid \mathcal{F}_{i-1}\big] - 1\right) - c_{n,\beta}\mathbb{E}_s\big[S_{c(i)} \mid \mathcal{F}_{i-1}\big] \sum_{l \in \nu_i \setminus \{c(i)\}} \lambda \frac{S_l^{\alpha}}{n} \\ & - c_{n,\beta}\frac{\lambda}{n}\mathbb{E}_s\Big[S_{c(i)}^{1+\alpha} \mid \mathcal{F}_{i-1}\Big] + o_{\mathbb{P}}\big(n^{-1/3}\big). \end{aligned} \tag{123}$$

By further expanding the first term in (123) as in (93), we get

$$\begin{aligned} \mathbb{E}_s[A_n(i) \mid \mathcal{F}_{i-1}] - 1 = {} & \left(c_{n,\beta}\frac{\lambda}{n} \sum_{j \notin \mathfrak{S}_{i-1}} S_j^{1+\alpha} - 1\right) - c_{n,\beta}\mathbb{E}_s\big[S_{c(i)} \mid \mathcal{F}_{i-1}\big] \sum_{l=i+1}^{i+1+Q_n(i-1)} \lambda \frac{S_{c(l)}^{\alpha}}{n} \\ & - c_{n,\beta}\frac{\lambda}{n}\mathbb{E}_s\Big[S_{c(i)}^{1+\alpha} \mid \mathcal{F}_{i-1}\Big] + o_{\mathbb{P}}\big(n^{-1/3}\big), \end{aligned} \tag{124}$$

where in the first equality we have used (6). Comparing Equation (124) with Equation (94), we rewrite the drift as

$$C_n(k) = C_n^u(k) - c_{n,\beta}\lambda \sum_{i=1}^{k} \mathbb{E}_s\big[S_{c(i)} \mid \mathcal{F}_{i-1}\big] \sum_{l=i+1}^{i+1+Q_n(i-1)} \frac{S_{c(l)}^{\alpha}}{n}. \tag{125}$$

Therefore, to conclude the proof of (i) it is enough to show that the second term vanishes, after rescaling. We do this in the following lemma.

Lemma 12. *As $n \to \infty$,*

$$n^{-1/3} c_{n,\beta}\lambda \sum_{i=1}^{\bar{t}n^{2/3}} \mathbb{E}_s\big[S_{c(i)} \mid \mathcal{F}_{i-1}\big] \sum_{l=i+1}^{i+1+Q_n(i-1)} \frac{S_{c(l)}^{\alpha}}{n} \xrightarrow{\mathbb{P}} 0. \tag{126}$$

Proof. By Lemma 11, $\sup_{i\le \bar{t}n^{2/3}} Q_n(i)\le C_1 n^{1/3}$ w.h.p. for a large constant C_1, and by Lemma 7, $\sup_{i\le \bar{t}n^{2/3}} \mathbb{E}_s[S_{c(i)}\mid \mathcal{F}_{i-1}]\le C_2$ w.h.p. for another large constant C_2. This implies that, w.h.p.,

$$n^{-1/3}c_{n,\beta}\lambda\sum_{i=1}^{\bar{t}n^{2/3}}\mathbb{E}_s\big[S_{c(i)}\mid \mathcal{F}_{i-1}\big]\sum_{l=i+1}^{i+1+Q_n(i-1)}\frac{S^{\alpha}_{c(l)}}{n}\le c_{n,\beta}\lambda C_2\sum_{i=1}^{\bar{t}n^{2/3}}\sum_{l=i+1}^{i+1+C_1n^{1/3}}\frac{S^{\alpha}_{c(l)}}{n^{4/3}}. \tag{127}$$

The double sum can be rewritten as

$$\begin{aligned} c_{n,\beta}\lambda C_2\sum_{i=1}^{\bar{t}n^{2/3}}\sum_{l=i+1}^{i+1+C_1n^{1/3}}\frac{S^{\alpha}_{c(l)}}{n^{4/3}} &\le c_{n,\beta}\lambda C_2\sum_{j=1}^{\bar{t}n^{2/3}+C_1n^{1/3}}\min\{j,C_1n^{1/3}\}\frac{S^{\alpha}_{c(j)}}{n^{4/3}}\\ &\le c_{n,\beta}\lambda C_1C_2\sum_{j=1}^{(\bar{t}+C_1)n^{2/3}}\frac{S^{\alpha}_{c(j)}}{n}. \end{aligned} \tag{128}$$

The rightmost term converges to zero in probability as $n\to\infty$ by Lemma 8. This concludes the proof. □

Because

$$n^{-1/3}C_n(tn^{2/3}) = n^{-1/3}C^{u}_n(tn^{2/3}) - n^{-1/3}c_{n,\beta}\lambda\sum_{i=1}^{tn^{2/3}}\mathbb{E}_s\big[S_{c(i)}\mid \mathcal{F}_{i-1}\big]\sum_{l=i+1}^{i+1+Q_n(i-1)}\frac{S^{\alpha}_{c(l)}}{n}, \tag{129}$$

Lemma 12 and the convergence result (100) for $n^{-1/3}C^{u}_n(tn^{2/3})$ conclude the proof of (i). □

4.2.2. Proof of (ii), (iii), and (iv) for the Embedded Queue. Proceeding as before, we find that

$$\begin{aligned} B_n(k) &= \sum_{i=1}^{k}\big(\mathbb{E}_s[A_n(i)^2\mid \mathcal{F}_{i-1}] - \mathbb{E}_s[A_n(i)\mid \mathcal{F}_{i-1}]^2\big)\\ &= \sum_{i=1}^{k}\big(\mathbb{E}_s[A_n(i)^2\mid \mathcal{F}_{i-1}] - 1\big) + O_{\mathbb{P}}(kn^{-1/3}), \end{aligned} \tag{130}$$

where

$$\mathbb{E}_s\big[A_n(i)^2\mid \mathcal{F}_{i-1}\big] = \mathbb{E}_s[A_n(i)\mid \mathcal{F}_{i-1}] + \mathbb{E}_s\left[\sum_{\substack{h,q\notin \nu_{i-1}\\ h\neq q}}\mathbb{1}_{\{T_h\le S_{c(i)}S_h/n\}}\mathbb{1}_{\{T_q\le S_{c(i)}S_q/n\}}\mid \mathcal{F}_{i-1}\right]. \tag{131}$$

Similarly as in Section 4.1.2, we get

$$\begin{aligned} &\sum_{\substack{h,q\notin \nu_{i-1}\\ h\neq q}}\mathbb{E}_s\Big[\mathbb{1}_{\{T_h\le S_{c(i)}S^{\alpha}_h/n\}}\mathbb{1}_{\{T_q\le S_{c(i)}S^{\alpha}_q/n\}}\mid \mathcal{F}_{i-1}\Big]\\ &\quad = \mathbb{E}_s\Big[S^2_{c(i)}\mid \mathcal{F}_{i-1}\Big]\frac{\lambda^2}{n^2}\left(\sum_{h=1}^{n}S^{\alpha}_h\right)^2 - \mathbb{E}_s\left[S^2_{c(i)}\frac{\lambda^2}{n^2}\sum_{\substack{h,q\in \nu_{i-1}\cup\{c(i)\}\\ \cup\{h=q\}}}S^{\alpha}_hS^{\alpha}_q\mid \mathcal{F}_{i-1}\right] + o_{\mathbb{P}}(1). \end{aligned} \tag{132}$$

The second term is an error term by Lemma 5 and Corollary 1. This implies that $B_n(\cdot)$ can be rewritten as

$$B_n(k) = \left(\frac{\lambda}{n}\sum_{h=1}^{n}S^{\alpha}_h\right)^2\sum_{i=1}^{k}\mathbb{E}_s\Big[S^2_{c(i)}\mid \mathcal{F}_{i-1}\Big] + o_{\mathbb{P}}(k), \tag{133}$$

so that

$$n^{-2/3}B_n(n^{2/3}u)\xrightarrow{\mathbb{P}}\lambda^2\mathbb{E}[S^{\alpha}]\mathbb{E}[S^{2+\alpha}]u, \tag{134}$$

which concludes the proof of (ii). □

To conclude the proof of Theorem 1, we are left to verify (iii) and (iv). However, the estimates in Sections 4.1.3 and 4.1.4 also hold for $B_n(\cdot)$ and $M_n(\cdot)$, because they rely, respectively, on (109) and (115) to bound the lower-order contributions to the drift. This concludes the proof of Theorem 1.

5. Conclusions and Discussion

In this paper, we considered a generalization of the $\Delta_{(i)}/G/1$ queue, which we coined the $\Delta^{\alpha}_{(i)}/G/1$ queue, a model for the dynamics of a queueing system in which only a finite number of customers can join. In our model, the arrival time of a customer depends on its service requirement through a parameter $\alpha \in [0,1]$. We have proved that, under a suitable heavy-traffic assumption, the diffusion-scaled queue-length process embedded at service completions converges to a stochastic process $W(\cdot)$. A distinctive characteristic of our results is the so-called *depletion-of-points effect*, represented by a quadratic drift in $W(\cdot)$. A (directed) tree is associated to the $\Delta^{\alpha}_{(i)}/G/1$ queue in a natural way, and the heavy-traffic assumption corresponds to *criticality* of the associated random tree. Our result interpolates between two already known results. For $\alpha = 0$ the arrival clocks are i.i.d., and the analysis simplifies significantly. In this context, Bet et al. (2019) proves an analogous heavy-traffic diffusion approximation result. Theorem 1 can then be seen as a generalization of Bet et al. (2019, theorem 5). If $\alpha = 1$, the $\Delta^{\alpha}_{(i)}/G/1$ queue has a natural interpretation as an exploration process of an inhomogeneous random graph. In this context, Bhamidi et al. (2010) proves that the ordered component sizes converge to the excursion of a reflected Brownian motion with parabolic drift. Our result can then also be seen as a generalization of Bhamidi et al. (2010) to the *directed* components of directed inhomogeneous random graphs.

Lemma 6 implies that the distribution of the service time of the first $O(n^{2/3})$ customers to join the queue converges to the *α-size-biased* distribution of S, irrespective of the precise time at which the customers arrive. This suggests that it is possible to prove Theorem 1 by approximating the $\Delta^{\alpha}_{(i)}/G/1$ queue via a $\Delta_{(i)}/G/1$ queue with service time distribution S^* such that

$$\mathbb{P}(S^* \in \mathscr{A}) = \mathbb{E}[S^{\alpha}\mathbb{1}_{\{S\in\mathscr{A}\}}]/\mathbb{E}[S^{\alpha}], \tag{135}$$

and i.i.d. arrival times distributed as $T_i \sim \exp(\lambda\mathbb{E}[S^{\alpha}])$. This conjecture is supported by two observations. First, the heavy-traffic conditions for the two queues coincide. Second, the standard deviation of the Brownian motion is the same in the two limiting diffusions. However, this approximation fails to capture the higher-order contributions to the queue-length process. As a result, the coefficients of the negative quadratic drift in the two queues are different, and thus the approximation of the $\Delta^{\alpha}_{(i)}/G/1$ queue with a $\Delta_{(i)}/G/1$ queue is insufficient to prove Theorem 1.

Endnote

[1] For brevity, we ignore the so-called *surplus edges* because they do not contribute to the sizes of the strongly connected component (Goldschmidt and Stephenson 2019, proposition 5.7).

References

Abramowitz M, Stegun IA (1964) *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables* (Dover Publications, New York).

Addario-Berry L, Broutin N, Goldschmidt C (2012) The continuum limit of critical random graphs. *Probability Theory Related Fields* 152(3–4):367–406.

Aldous D (1997) Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Probability* 25(2):812–854.

Bet G (2018) An alternative approach to heavy-traffic limits for finite-pool queues. Preprint, submitted November 23, https://arxiv.org/abs/1811.09576.

Bet G, Selen J, Zocca A (2020) Weighted Dyck paths for nonstationary queues. Preprint, submitted February 9, https://arxiv.org/abs/2002.03424.

Bet G, van der Hofstad R, van Leeuwaarden JSH (2017) Finite-pool queueing with heavy-tailed services. *J. Appl. Probab.* 54(3):921–942.

Bet G, van der Hofstad R, van Leeuwaarden JSH (2019) Heavy-traffic analysis through uniform acceleration of queues with diminishing populations. *Math. Oper. Res.* 44(3):821–864

Bhamidi S, Budhiraja A, Wang X (2014) The augmented multiplicative coalescent, bounded size rules and critical dynamics of random graphs. *Probability Theory Related Fields* 160(3–4):733–796.

Bhamidi S, Sen S, Wang X (2017) Continuum limit of critical inhomogeneous random graphs. *Probab. Theory Related Fields* 169(1–2):565–641.

Bhamidi S, van der Hofstad R, van Leeuwaarden JSH (2010) Scaling limits for critical inhomogeneous random graphs with finite third moments. *Electronic J. Probability* 15:1682–1702.

Bhamidi S, van der Hofstad R, van Leeuwaarden JSH (2012) Novel scaling limits for critical inhomogeneous random graphs. *Ann. Probability* 40(6):2299–2361.

Bollobás B, Janson S, Riordan O (2007) The phase transition in inhomogeneous random graphs. *Random Structures Algorithms* 31(1):3–122.

Dhara S, van der Hofstad R, van Leeuwaarden JSH, Sen S (2017) Critical window for the configuration model: finite third moment degrees. *Electron. J. Probab.* 22(2017):Paper No. 16.

Dhara S, van der Hofstad R, van Leeuwaarden JSH, Sen S (2016) Heavy-tailed configuration models at criticality. Preprint, submitted December 2, https://arxiv.org/abs/1612.00650.

Duquesne T, Le Gall J-F (2005) Random trees, Lévy processes and spatial branching processes. Preprint, submitted September 23, https://arxiv.org/abs/math/0509558.

Durrett R (2010) *Probability—Theory and Examples* (Cambridge University Press, MA).

Ethier SN, Kurtz TG (1986) *Markov Processes: Characterization and Convergence* (Wiley, New York).

Goldschmidt C, Stephenson R (2019) The scaling limit of a critical random directed graph. Preprint, submitted October 29, https://arxiv.org/abs/1905.05397.

Honnappa H, Ward AR (2014) On transitory queueing. Preprint, submitted December 7, https://arxiv.org/abs/1412.2321.

Honnappa H, Jain R, Ward AR (2015) A queueing model with independent arrivals, and its fluid and diffusion limits. *Queueing Systems* 80:71–103.

Joseph A (2014) The component sizes of a critical random graph with a given degree sequence. *Ann. Appl. Probability* 24(6):2560–2594.

Kendall DG (1951) Some problems in the theory of queues. *J. Roy. Statist. Soc. B* 13(2):151–185.

Klenke A (2008) *Probability Theory: A Comprehensive Course* (Springer, London).

Le Gall J-F (2005) Random trees and applications. *Probability Survey* 2:245–311.

Limic V (2001) A LIFO queue in heavy traffic. *Ann. Appl. Probability* 11(2):301–331.

Luczak T (1990) The phase transition in the evolution of random digraphs. *J. Graph Theory* 14(2):217–223.

Luczak T, Seierstad TG (2009) The critical behavior of random digraphs. *Random Structures Algorithms* 35(3):271–293.

Martin-Löf A (1998) The final size of a nearly critical epidemic, and the first passage time of a Wiener process to a parabolic barrier. *J. Appl. Probability* 35(3):671–682.

Takács L (1988) Queues, random graphs and branching processes. *J. Appl. Math. Simulation* 1(3):223–243.

Takács L (1993) Limit distributions for queues and random rooted trees. *J. Appl. Math. Stochastic Anal.* 6(3):189–216.

Takács L (1995) Queueing methods in the theory of random graphs. Dshalalow JH, ed. *Advances in Queueing: Theory, Methods and Open Problems* (CRC Press, Boca Raton, FL), 45–78.

van der Hofstad R (2016) *Random Graphs and Complex Networks* (Cambridge University Press, MA).

van der Hofstad R, Janssen A, van Leeuwaarden JSH (2010) Critical epidemics, random graphs, and Brownian motion with a parabolic drift. *Adv. Appl. Probability* 42(4):1187–1206.

van der Hofstad R, Kliem S, van Leeuwaarden JSH (2018) Cluster tails for critical power-law inhomogeneous random graphs. *J. Statist. Physics* 171(1):38–95.

van der Hofstad R, van Leeuwaarden JSH, Stegehuis C (2016) Mesoscopic scales in hierarchical configuration models. Preprint, submitted December 8, https://arxiv.org/abs/1612.02668.

Whitt W (2002) *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues* (Springer, New York).