Spatial Capacity Planning

Omar Besbes
Columbia Business School
Francisco Castro
Columbia Business School
Ilan Lobel
NYU Stern School of Business

We study the relationship between capacity and performance for a service firm with spatial operations, in the sense that requests arrive with origin-destination pairs. An example of such a system is a ride-hailing platform in which each customer arrives in the system with the need to travel from an origin to a destination. We propose a state-dependent queueing model that captures spatial frictions as well as spatial economies of scale through the service rate. In a classical $M/M/n$ queueing model, the square root safety (SRS) staffing rule is known to balance server utilization and customer wait times. By contrast, we find that the SRS rule does not lead to such a balance in spatial systems. In a spatial environment, pickup times increase the load in the system; furthermore, they are an endogenous source of extra workload that leads the system to only operate efficiently if there is sufficient imbalance between supply and demand. In heavy traffic, we derive the mapping from load to operating regimes and establish implications on various metrics of interest. In particular, to obtain a balance of utilization and wait times, the service firm should use a higher safety factor, proportional to the offered load to the power of 2/3. We also discuss implications of these results for general systems.

Key words: heavy traffic, queueing, capacity sizing, staffing, spatial operations, QED regime, ride-hailing, ride-sharing, asymptotic analysis

1. Introduction

Motivation. Many traditional service systems are characterized by static servers and customers that arrive stochastically and line up in a queue before receiving service. These include call centers, health-care facilities, and amusement parks, among others. In designing such systems, one faces a tradeoff between the cost of servers and the quality of service as measured through the characteristics of wait time. The prevalence of such systems has led to an extensive literature on capacity sizing that has provided important practical guidelines about how to set capacity levels in service systems. Typically, there is a fine balance between the two objectives. A central rule, the so-called

square root safety (SRS) staffing rule, emerges naturally from different performance considerations. In the SRS rule, the capacity is set at the nominal offered load plus a safety factor proportional to the square root of the offered load. If one considers a social planner's problem attempting to minimize the system's total cost measured by the aggregate of capacity and waiting costs, the SRS rule is optimal in large systems. Another central metric in the literature and in practice is the probability that a customer waits before being attended by a server, which has led to the coining of various terms to describe the regimes of interest. Quality driven (QD) is the regime where customer quality is paramount and, thus, the probability of waiting is vanishingly small. Efficiency driven (ED) refers to the regime where cost concerns prevail. In ED, a customer's probability of having to wait approaches one. Quality and Efficiency driven (QED) is the intermediate regime, where the probability that a customer waits is separated from both zero and one, leading to a fine balance between utilization and quality of service. The latter is achieved through the SRS staffing rule. The latter capacity is sufficient to ensure that a positive fraction of customers do not wait at all before receiving service. .

However, there are other service systems in which customers arrive to random locations in space and servers have to spend time not only servicing customers, but also reaching them before service starts. This includes, for example, ride-hailing systems such as Uber, Lyft, Via and DiDi. On these platforms, a customer requests a ride from a given location and a driver is then dispatched by the platform to pick him up and take him to his desired destination.[1] Automated warehouses powered by Kiva robots (Amazon robotics) or the Ocado smart platform provide another example. In these warehouses, products are arranged in a grid. As orders for different products arrive, robots are dispatched to collect the products and transport them to picking stations. In these spatial multi-server systems, workload is larger than in traditional systems because servers must reach customers before starting to service them, making it unclear whether the SRS rule of thumb is still valid. The central question of this paper is the following: *How should "capacity thinking" be adapted to spatial settings, where servers need to reach customers before service can start?*

We anchor our analysis around a *spatial multi-server system* in which arrivals to a two-dimensional region follow a Poisson process. A customer draws an origin and a destination uniformly and independently in the region. From a pool of n servers, a central platform dispatches a server that must reposition to the origin of her assigned customer and then take him to his desired destination. This spatial multi-server system is different from a traditional queueing $M/M/n$ service system in at least two dimensions. First, servers must "pick up" customers by repositioning to a customer location before starting service. This translates into extra workload added to the system

[1] For consistency, we refer to customers as males and servers/drivers as females throughout the paper.

compared to a traditional system. Second, as the imbalance of servers and customers increases, *spatial economies of scale* can make the system operate at a faster pace. For example, the larger the spatial density of idle servers, the more opportunities for better matches and the shorter the time it takes a server to pick up an arriving customer. Similarly, the larger the spatial density of waiting customers, the more opportunities for better matches and the shorter the time it takes an idling server to reach a customer. That is, in a spatial multi-server system, service rate is state-dependent and might improve with large supply-demand imbalances. This is illustrated in Figure 1.

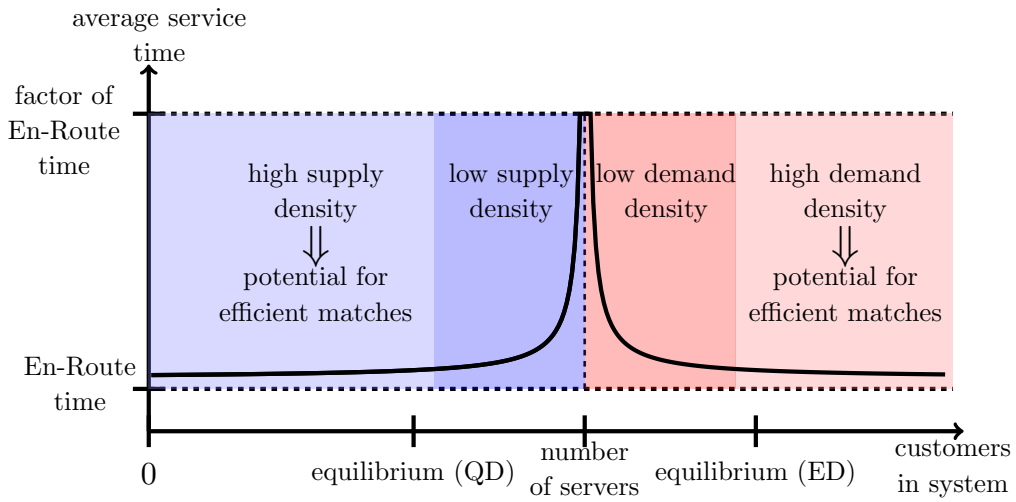


Figure 1 **Illustration of the potential for matches and the impact on pickup times.**

In order to shed light on the capacity sizing question of interest, we take a macro view of the spatial system by focusing on the key features that dictate its dynamics. More concretely, we consider a Markovian stochastic model that captures the key characteristics of input and output rates in the spatial multi-server system. Our Markovian model is a standard queueing system with n servers, but with a *state-dependent* service rate that adequately reproduces the spatial economies of scale of spatial systems. We analyze this queueing model in heavy traffic. On the one hand, the queueing setting provides guidelines for how the spatial system will behave. On the other hand, the spatial setting provides a physical interpretation of the queueing model results.

Main contributions. Our first contribution lies in the modeling domain. We develop a Markovian model that captures fundamental aspects of capacity planning in dynamic spatial environments. The system we analyze features both service speedups and service slowdowns that emerge due to the presence of spatial economies of scale. In addition, we ground our analysis on near-optimal dispatch rules derived from the vehicle routing literature.

Our second contribution lies in the set of insights and fundamental results we obtain for this class of problems. We first analyze a fluid model that highlights some of the key properties of

such systems. We characterize in closed form the two possible stable equilibria of this deterministic model. These equilibria correspond to two types of potential operating regimes: the first one with a high density of waiting customers and the second one with a high density of idle servers. These equilibria are depicted in Figure 1. In both of these operating points, the system is able to match customers to servers efficiently since supply and demand are fairly imbalanced.

We then analyze the stochastic system in heavy traffic. In this setting we first establish that, in stark contrast with a standard multi-server system, the SRS rule will always bring the spatial multi-server system to the efficiency-driven (ED) regime, in which customers will wait for a server to be dispatched with probability approaching one. In other words, the added workload due to pickups is substantial enough and cannot be compensated by simply increasing capacity levels on the order of the square root of the offered load.

In turn, we fully characterize the asymptotic system's performance under a range of scalings. If the capacity buffer is of lower order than the offered load to the power of 2/3, then the system is in the efficiency-driven (ED) regime. The system operates around the ED equilibrium depicted in Figure 1. If the capacity buffer is of higher order than the offered load to the power of 2/3, then the system is in the quality-driven (QD) regime. The system operates around the QD equilibrium depicted in Figure 1. Hence, in a spatial environment, the QED regime may only emerge if the safety capacity is of order the offered load to the power of 2/3. We furthermore establish that the QED regime can indeed be achieved. The QED regime does not correspond to a new stable operating point of the system, but to a system that oscillates stochastically between the ED and QD equilibrium points. Reaching the QED regime is more subtle in a spatial environment, as now it does not only depend on the order of the safety capacity but also on second order terms. Furthermore, as a by-product of this analysis, we can approximate the system cost and establish that the power of 2/3 scaling is optimal in the sense that it minimizes a sum of server costs and waiting costs, which is a natural social planner's objective.

We show that the approximation method used, which greatly simplifies the analysis of an otherwise highly non-tractable system, captures the fundamental features of the true system. We validate our approach via a series of numerical simulations that show that the heavy-traffic behavior of our Markovian system closely captures that of a simulated spatial multi-server system.

In sum, our model and results imply that common rules of thumb such as the SRS rule will no longer be valid for spatial operations and, therefore, new staffing rules of thumb are necessary. This has implications for how to think about such trade offs in automated warehouses and, with the advent of fleets of self-driving cars, in ride-hailing platforms. Our results derive new rules of thumb for the implications of capacity levels on the type of service regime they induce.

2. Related Literature

Our paper relates to several streams of literature.

Staffing. Our goal is to analyze the performance of a system with customers arriving and being served in a spatial setting as measured by the steady-state probability of waiting in heavy traffic. The seminal work of Halfin and Whitt (1981) introduces the so-called Halfin-Whitt regime in which the system is taken to heavy traffic by scaling the number of servers as $R + \beta \cdot \sqrt{R}$ where R is the offered load. This is also known as the square root staffing (SRS) rule. Under this regime, the authors show that in an $M/M/n$ or $GI/M/n$, the system the probability of delay is strictly between zero and one–the QED regime. Garnett et al. (2002) and Whitt (2004) study the Erlang-A case. For more on the QED regime with applications to call centers, we refer the reader to the survey papers by Gans et al. (2003) and Aksin et al. (2007). We also refer the reader to Whitt (2007) for related work, and Reed (2009) for the more general case of the $G/GI/n$ system. Bassamboo et al. (2010) study the capacity sizing problem in an environment in which there is also parameter uncertainty for mean arrival rate, deriving new prescriptions for such settings and articulating how to operate depending on whether one is in an uncertainty-dominated or a variability-dominated regime. Our work is complementary to this literature in the sense that we also analyze the performance of the system as measured by the probability of delay. In our model, however, the presence of spatial frictions affects dynamics and introduces state-dependencies, leading to fundamental changes in how capacity should be scaled to achieve QED performance. For an in depth discussion about limiting regimes (ranging from the conventional heavy traffic regime to the Halfin-Whitt regime and passing through the slowdown regime) and their implications for diffusion approximations in non-spatial environments we refer the reader to Ward (2012) and Atar (2012).

State-dependent service rate. The general spatial system we aim to understand is complex and generally intractable. To gain insight we consider a simpler Markovian version of it that can be regarded as an $M/M_Q/n$ system. Our work is thus related to the broad literature on Markovian system and birth and death processes, and in particular to the works that study service systems with state-dependent processing rates; for some examples we refer the reader to Mandelbaum and Pats (1995), Mandelbaum et al. (1998) and Powell and Schultz (2004). Chan et al. (2014) study an Erlang-R service system in which the service rate can be sped up whenever congestion is above a certain threshold . Using a fluid analysis they show that, depending on system parameters, speeding up service can lead to both desirable and undesirable system congestion levels. In related work, Dong et al. (2015) study a service system in which agents are sensitive to individual future work load and reduce their service rate as the system's workload increases. They show that depending on load sensitivity, the system's slowdowns can take it from moderate to substantial deterioration.

Our work can be considered as a combination of both speedups and slowdowns, and the exact form of these in our context is driven by *spatial economies of scale*. As the number customers in our system increases beyond n, the density of waiting customers increases and, therefore, the next idling servers can spend less time picking up customers, i.e., service rate speeds up. Similarly, when the number of customer is increasing but below n, the density of idle cars decreases and, therefore, arriving customers may experience larger pickup times, i.e., service rate is slowed down. These effects are a result of the physical nature of our system. Related to the above papers, and in particular Dong et al. (2015), our system features some form of bi-stability in an underlying tightly related deterministic model. In contrast, however, the equilibria emerge on different scales in our setting and asymptotically in the stochastic system, these can survive jointly.

Stochastic vehicle routing. Another related stream of related work is that of dynamic routing problems. Routing is a highly complex class of problems and measuring the performance of routing algorithms is challenging. Bertsimas and van Ryzin (1991, 1993) show that the scaling of queues in space is fundamentally different to the one when space is ignored. In particular, Bertsimas and van Ryzin (1991, 1993) obtain a lower bound for the minimum expected total time in the system under any dispatching policy given by $\Theta\left(\lambda/(n^2(1-\rho)^2)\right) + \Theta(1)$ in heavy traffic, as the offered load converges to 1. This is a remarkable result that provides a lower bound for all dispatching policies and in turn sets a target for the optimal expected time in the system which can be used as a guideline to measure the performance of policies. Interestingly, the size of the system scales with $1/(1-\rho)^2$ and not with $1/(1-\rho)$ as happens in the standard $M/M/n$ system. Thus the fact that we are taking into account space fundamentally changes how the system scales with traffic intensity.

Ride-hailing. In the young but quickly growing literature on ride-hailing systems, customers arrive in a spatial region and a platform matches them to drivers who, in turn, take the customers to their desired destinations. For the important problem of spatial incentives in ride-hailing system we refer the reader to Banerjee et al. (2015), Bimpikis et al. (2016), Castillo et al. (2017), Afèche et al. (2018) and Besbes et al. (2018).

Closer to our work are the studies that investigate the problem of matching to optimize certain performance metrics. Using a fluid approach in a closed queueing network, Braverman et al. (2016) study how to route empty cars in order to maximize network utility. In related work, Ozkan and Ward (2016) use a fluid approach to derive policies that maximize the number of matches. Banerjee et al. (2018) study matching in a closed queueing network, and show that for a Scaled MaxWeight policy, the proportion of dropped demand in steady state decays exponentially fast as the number of servers in the system grows large. In a circular city framework, Feng et al. (2017) analyze the waiting time performance of different matching mechanisms. The focus of this paper, in contrast, is to

understand how to think about capacity planning in spatial environments. Rather than optimizing over the space of dispatching policies, we anchor our analysis around a near-optimal dispatching policy.

Closest to our setting is Castillo et al. (2017). There, the authors also analyze inefficiencies stemming from additional workload in a spatial system, and study the possible use of surge pricing to alleviate these. Our study focuses on a different question, that of capacity planning. The two papers utilize different dispatch policies. Our framework can be used to analyze the type of dispatching considered there, in which the additional capacity needed would be of order the offered load. In contrast, in our case, we focus on a class of provably near-optimal dispatch rules based on the vehicle routing literature mentioned above, which, as we establish, enables one to only need a safety capacity of the order the offered load to the power of 2/3.

3. Spatial Queueing Model

We introduce a stochastic model for spatial capacity planning within a bounded region of a plane. Our model is an $M/M_Q/n$ queueing system (in Kendall's notation M_Q stands for state-dependent service time) that captures the fundamental aspects of a spatial system that experiences arrivals and dispatches servers to attend to those arrivals.

3.1. Model

Motivation. We consider two models in our paper. The first is what we call the *general system*, where spatial elements such as origin-destination pairs of customers are explicitly modeled. The second is a *Markovian system*, which is a queueing system that approximates the general system. In the Markovian system, the spatial frictions are captured in reduced form via a state-dependent service time. All of the mathematical results in the paper establish properties of the Markovian system that can be regarded as qualitative prescriptions for the general system. Indeed, in Section 6, we use simulation to demonstrate that the Markovian system approximates the behavior of the general system quite well.

We are interested in gaining insights on the following general system. There is a central platform, customers and servers that interact in a bounded connected subset $\mathcal{C}$ of $\mathbb{R}^2$ (the city). Customers arrive according to a Poisson process in the city at uniformly distributed locations in the city. Each customer wishes to travel from the point they arrive to some other point also drawn uniformly at random among all locations in $\mathcal{C}$. Customers are patient and will remain in the system until served.

There is a fixed number of servers in the system, and each one can serve one customer at a time at a constant velocity. A server first repositions to the arrival location of a customer, and

then she transports that customer to his destination. Upon arrival to his final destination, the customer leaves the system and the server becomes idle and waits until the platform relocates her. The repositioning of servers occurs according to some state-dependent dispatching algorithm and is controlled by the platform.

Any given customer experiences a total time in the system that is composed of three components: *waiting time*, *pickup time* and *en-route time*:

$$\text{Time in the system} = \text{Waiting} + \text{Pickup} + \text{En-route}. \tag{1}$$

The waiting time corresponds to the time a customer spends in the system before he is assigned a server to pick him up. The pickup time represents the time it takes for the server to relocate from where she currently is to the customer's origin location. The en-route time is the time it takes to transport a customer from his origin to his destination.

The system described at a high level above is complex and intractable to analyze in its full generality, given the stochasticity of the system, the high-dimensional state-space, and the complexity of the space of possible dispatching policies.

Queueing model. In this paper we study what we call the Markovian system, which is a simpler queueing model that still captures the spatial features of the general system. In setting up our model, we deliberately forego the complex interactions among agents that make the general system intractable, and focus on the overall physical dynamics that dictate the processing performance of the system. We further discuss our modeling assumptions in Section 3.2.

We focus on a model in which customers arrive to the system according to a Poisson process with rate λ, and stay until served. There is a total of n identical servers that provide service to one customer at a time in a first come, first serve fashion. We assume that the time between the assignment of a server to a customer and the end of the service is exponentially distributed with state-dependent rate $\mu(\cdot)$. Upon arrival, if a customer finds a server idle, he is immediately assigned a server; otherwise, he waits in line. This leads to an $M/M_Q/n$ queueing system. We use $Q(t)$ to denote the total number of customers in the system at time t, which includes both customers waiting and in service.

The distinctive feature of the system we analyze and what makes it depart from a traditional multi-server queue is that *servers must be repositioned to serve customers*. As a result, the total time a server spends on a single customer corresponds to pickup time plus en-route time as opposed to just en-route time—the analogue to service time in a traditional queueing system. In turn, in order to capture the overall processing performance of the general system, the key is to select an appropriate function $\mu(\cdot)$ that isolates spatial frictions through the combination of both pickup and en-route times as highlighted in Eq. (1).

Any sensible choice of the service rate must be such that its inverse, $1/\mu(\cdot)$, has two components: one reflecting pickup times and the other en-route times. En-route times are simple. They correspond to the distance between two random locations in $\mathcal{C}$ (properly scaled by the velocity) and do not depend on the state of the system. If we let $\bar{s}_t$ to denote the expected time to move between two random points in $\mathcal{C}$ (for some nominal velocity), then it follows that one of the components of $1/\mu(\cdot)$ will be equal to $\bar{s}_t$. The remaining component has to relate to pickup times. These are more involved as they depend on how, based on the state of the system, the platform decides to do the assignment of servers to customers—the dispatching algorithm. To overcome this difficulty it is convenient to look at the physics of the spatial system under a particular dispatching algorithm. In the present study, we anchor our analysis around the asymptotic behavior of one notable dispatching algorithm: *nearest-neighbors* dispatch (`NN`). This algorithm is simple, intuitively appealing, and it is also near-optimal.[2] If there are more servers than customers, `NN` assigns the next arriving customer to its closest available servers. If there are less servers than customers, `NN` assigns the next idling server to its closest waiting customer.

The asymptotic behavior of `NN`, which we discuss in Section 3.2, leads to a particular form of the expected service time which, in turn, motivates the following expression for the state-depend rate of our queueing system when its state is q

$$\frac{1}{\mu(q)} \triangleq \frac{\bar{s}_p}{\sqrt{|q-n| \vee 1}} + \bar{s}_t, \quad q \geq 0, \tag{2}$$

for two given positive constants $\bar{s}_p$ and $\bar{s}_t$, where $\bar{s}_p$ represents the average pickup time when there is one server available and one passenger request. The form in Eq. (2) captures spatial frictions in the following way. Consider the queueing system. If $Q(t) \ll n$, then $|Q(t) - n|$ is large and many servers are available, and thus, $1/\mu(Q(t))$ is close to $\bar{s}_t$, the expected en-route time. The pickup time should be negligible given the high density of free servers in space. Similarly, if $Q(t) \gg n$, then many customers are waiting and a match with a low pickup time could be found given the high density of customers in space. Indeed, we also have that $1/\mu(Q(t))$ is close to $\bar{s}_t$ in this scenario. The important point is that whenever there is a critical idle/waiting mass at either side of the market, the physical nature of the system allows it to process customers efficiently. When $Q(t) \approx n$, we expect the match between server and customer to lead to a significantly higher pickup time. In our model, a customer's total expected service time will be close to $\bar{s}_p + \bar{s}_t$ when $Q(t) \approx n$. For notational simplicity, we assume $\bar{s}_t = \bar{s}_p$ throughout the next few sections, and denote this quantity simply by $\bar{s}$. When we simulate the system in Section 6, we allow $\bar{s}_t$ and $\bar{s}_p$ to take distinct values.

[2] Among the policies that minimize customers' expected total system time, `NN` achieves near optimal performance (see e.g., Bertsimas and van Ryzin (1991)) .

Performance Metrics. The main objective of this paper is to understand the implications of spatial frictions on performance metrics of the service system. In particular, we analyze these in an asymptotic regime in which the number of servers and the arrival rate grow large. We analyze the system in heavy traffic and consider a sequence of $M/M_Q/n$ queues indexed by n, with arrival rate λ_n such that

$$\rho_n < 1, \quad \lim_{n\to\infty} \lambda_n = \infty, \quad \lim_{n\to\infty} (1-\rho_n) n^\alpha = \beta, \quad \text{for some } \beta \in \mathbb{R}_+,\ \alpha \in (0,1), \tag{3}$$

where ρ_n equals $\bar{s} \cdot \lambda_n / n$. Thus, ρ_n approaches 1 from below at rate $1/n^\alpha$. Under these different scalings (as α varies), our goal is to study key performance metrics associated with the system. We let $\{Q_n(t)\}_{t\geq 0}$ denote the number of customers in the n-system. The dynamics of $Q_n(t)$ can be written as follows. Let $A = \{A(t) : t \geq 0\}$ and $S = \{S(t) : t \geq 0\}$ be two independent unit rate Poisson processes. The path-wise construction of Q_n is

$$Q_n(t) = Q_n(0) + A(\lambda t) - S\left(\int_0^t \mu_n(Q_n(u)) \cdot \min(n, Q_n(u)) du\right), \quad Q_n(0) = Q_0. \tag{4}$$

The term Q_0 corresponds to the initial state of the system, the second term captures the cumulative arrivals up to time t, and the third term refers to the cumulative departures up to t. In the latter, $\mu_n(Q_n(t)) \cdot \min(n, Q_n(t))$ corresponds to the service rate of the system, with $\mu_n(Q_n(t))$ representing the service rate per server at time t and $\min(n, Q_n(t))$ the number of non-idle servers at time t.

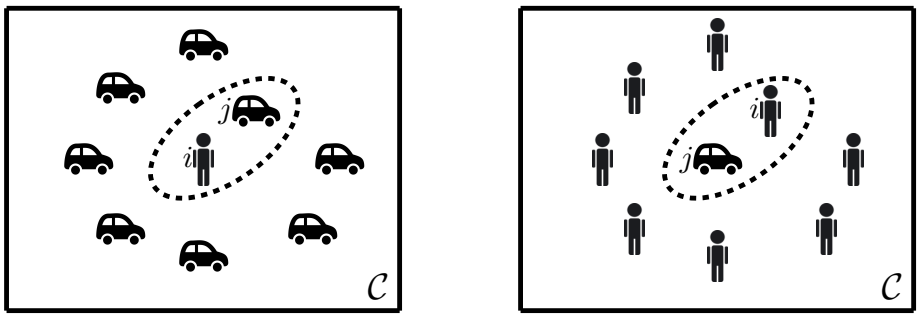
We use $Q_n(\infty)$ to denote a random variable representing the number of customers in the system in steady-state. One key central metric we are interested in quantifying is the steady-state limiting delay probability

$$P_\infty(W) \triangleq \lim_{n\to\infty} \mathbf{P}[Q_n(\infty) \geq n],$$

in order to assess the system performance. As in classical multi-server queues (see, e.g., Halfin and Whitt (1981)), if $P_\infty(W) = 1$, the system is said to be operating in the *efficiency-driven* (ED) regime, if $P_\infty(W) = 0$ the system is said to be operating in the *quality-driven* (QD) regime, and if $P_\infty(W) \in (0,1)$, the system is said to be in the *quality- and efficiency-driven* (QED) regime. In the coming sections, we characterize how $P_\infty(W)$ changes as the values of α and β change. In turn, we will also analyze implications on various other metrics such as, e.g., total system cost.

3.2. Discussion of the Modeling Assumptions

We now provide an asymptotic grounding for Eq. (2), based on the `NN` dispatching algorithm that is studied in the vehicle routing literature (Bertsimas and van Ryzin (1991)). Recall that for this policy, when there are more servers than customers, the closest idle server is assigned to a new arrival (see Figure 2 (a)). In the case when there are more customers than servers, as soon as a server becomes idle, we assign her to the closest customer (see Figure 2 (b)).



(a) **(b)**

Figure 2 **Nearest neighbor policy (NN). In (a) we have $Q(t) < n$, in (b) we have $Q(t) > n$.**

The connection between $\mu(\cdot)$ and NN comes from the following argument. Consider a general system operating under NN. Suppose that at time t there is a total of $Q(t)$ customers, and that server j was matched to customer i. Depending on the state of the system, the assignment could have happened in two different ways. If $Q(t) < n$, server j must be the closest idle server to customer i among $n - Q(t)$ idle servers (see Figure 2 (a)). If $Q(t) \geq n$, customer i must be the closest waiting customer to server j among $Q(t) - n$ waiting customers (see Figure 2 (b)). In either case customer i's pickup time can be computed by comparing the distance of the closest of $|Q(t) - n| \vee 1$ random variables uniformly distributed in $\mathcal{C}$ to a single point. We can then use the following standard result from probability to obtain an asymptotic approximation for a customer's expected pickup time under NN.

LEMMA 1. *Let $X_1, X_2, \ldots$ be a sequence of independent uniformly distributed random points in $\mathcal{C}$. Then, the expected minimum distance to any x_0 in the interior of $\mathcal{C}$ satisfies*

$$\mathbf{E}\left[\min_{i=1,\ldots,k} \|X_i - x_0\|\right] = \Theta\left(\frac{1}{\sqrt{k}}\right), \quad \text{as } k \uparrow \infty.$$

Conditioning on $Q(t)$ and ignoring any dependencies among the involved random variables, Lemma 1 suggests the following approximation for a customer's expected pickup time

$$\mathbf{E}[\text{Pickup}|Q(t)] \approx \frac{\bar{s}_p}{\sqrt{|Q(t) - n| \vee 1}},$$

for some positive $\bar{s}_p$. The first term in Eq. (2) incorporates this approximation.

We note that the particular approximation we use in $\mu(\cdot)$ discussed above is not the only simplifying assumption we use in the Markovian system. We also assume that server travel times, including both pickup and en-route times, are exponentially distributed. We argue in Section 6 using simulation that our approximations are reasonable, in the sense that the Markovian system approximates well the behavior of the general system.

First-dispatch. Another dispatching protocol that has received attention in the literature is *first-dispatch* (FD). Under FD, an arriving customer is assigned as soon as possible to the closest idle

server. Consider again Figure 2. In the situation depicted on the left panel (a), NN and FD operate according to the same rules. However, in the situation represented by the right panel (b) of Figure 2, the two dispatch rules operate quite differently. In this case, the FD dispatching algorithm assigns the next idling server to the longest waiting customer. As pointed out by Castillo et al. (2017), the FD dispatch rule can lead the system to a bad equilibrium they call the *Wild Goose Chase* in which servers spend long times picking up customers. Our framework can be used to analyze the systems' performance under the FD dispatch policy. Using Lemma 1 we can derive the following expression for an approximate service rate under FD:

$$\frac{1}{\mu_{\texttt{FD}}(q)} = \frac{\bar{s}_p}{\sqrt{(n-q)^+ \vee 1}} \; + \; \bar{s}_t, \quad q \geq 0.$$

Unlike the NN policy, the FD policy does not make use of *spatial economies of scale* when the system is heavily loaded with customers ($q > n$); instead, it serves customers on a first come first serve basis. This gives rise to the *Wild Goose Chase* phenomenon. Under this inefficient dispatching protocol, the number of servers required to escape ED performance equals the offered load plus a buffer term that is of the same order of the offered load, as opposed to a buffer of the order of the offered load to the power of 2/3 under NN. The NN dispatching protocol avoids this bad equilibrium outcome by exploiting spatial economies of scale even when the system is heavily loaded with customers.

4. Dynamics of a Related Deterministic System

Before we study the stochastic limiting properties of the Markovian system in Section 5, we analyze the properties of a deterministic version of it that will provide natural candidate focal points for the former system and initial insights on its behavior. In particular, we focus on a natural deterministic counterpart of Eq. (4).

Deterministic dynamics. Consider the dynamics of $\tilde{Q}_n(\cdot)$ described by

$$\tilde{Q}_n(t) = \tilde{Q}_n(0) + \lambda_n t - \int_0^t \mu_n\left(\tilde{Q}_n(u)\right) \cdot \min\left(n, \tilde{Q}_n(u)\right) du, \quad \tilde{Q}_n(0) = \tilde{Q}_0,$$

where $\tilde{Q}_0$ is a non-negative constant. This dynamical system has a simple interpretation. A fluid of customers joins the system at rate λ_n and departs at state-dependent rate $\mu_n(\tilde{Q}_n(t)) \cdot \min(n, \tilde{Q}_n(t))$. This dynamical system is a deterministic version of the one presented in Eq. (4). From the equation above, we can write $\tilde{Q}_n$ as the solution of the ordinary differential equation

$$\frac{d\tilde{Q}_n(t)}{dt} \; = \; f_n(\tilde{Q}_n(t)), \quad \tilde{Q}_n(0) = \tilde{Q}_0, \tag{5}$$

where

$$f_n(q) \triangleq \lambda_n - \mu_n(q) \cdot \min\{n, q\}.$$

Since $\mu_n(\cdot)$ is a Lipschitz continuous function, so is $f_n(\cdot)$. Therefore, by the Picard-Lindelof theorem, the ODE in Eq. (5) has a unique solution, which we denote by $\Phi(q_0, t)$ for a given $\tilde{Q}_n(0) = q_0$. In what follows, we study the equilibrium points of this solution.

Definition 1 (Equilibria). We say that a point q^* is an *equilibrium* point of the dynamic system presented in Eq. (5) if

$$\Phi(q^*, t) = q^*, \quad \text{for all } t \geq 0.$$

An equilibrium point q^* is such that if the systems starts at q^*, then the systems remains at q^* for all $t \geq 0$. Observe that we can compute an equilibrium by solving $f_n(q^*) = 0$. In general, a dynamical system can have multiple equilibria but these may have different properties. We classify the equilibria according to the following definition.

Definition 2 (Stability of Equilibria). An equilibrium q^* of Eq. (5) is said to be *stable* if for any $\epsilon > 0$, there exists $\delta > 0$ such that if $|q - q^*| < \delta$, then $|\Phi(q, t) - q^*| < \epsilon$ for all $t \geq 0$. Otherwise, q^* is *unstable*. If q^* is stable and there exists $\delta > 0$ such that if $|q - q^*| < \delta$, then $\lim_{t\to\infty} \Phi(q, t) = q^*$, we say that q^* is *locally asymptotically stable*. If $\lim_{t\to\infty} \Phi(q, t) = q^*$ for any $q \geq 0$, we say that q^* is *globally asymptotically stable*.

Informally, an equilibrium q^* is *stable* if whenever the system is slightly perturbed from q^*, it remains near q^*. An equilibrium q^* is *unstable* if small perturbations of the system around q^* take the system away from q^*. If for any starting point q, the dynamic $\Phi(q, t)$ converges to q^* then q^* is *globally asymptotically stable*. If the latter is true but only in a neighborhood of q^* then q^* is *locally asymptotically stable*. Next we study the equilibria of the dynamical system from Eq. (5).

Equilibria characterization. Recall that the equilibrium points of Eq. (5) can be found by solving $f_n(q^*) = 0$. The next theorem provides a complete description of the solutions to this equation for n large.

Theorem 1 **(Equilibrium Points)**. *Suppose* $\lim_{n\to\infty}(1 - \rho_n)n^\alpha = \beta$ *and* $\rho_n \uparrow 1$, *and let* $\beta_1^* = 3/(4^{1/3})$.

(i) Then, there exists n_0 *such that for all* $n \geq n_0$, *the system from Eq.* (5) *admits an equilibrium given by*

$$\overline{q}_n = n + \frac{\rho_n^2}{(1 - \rho_n)^2}.$$

Furthermore, this equilibrium is unique and globally asymptotically stable if $\alpha > 1/3$ *or if* $\alpha = 1/3$ *and* $\beta < \beta_1^*$.

(ii) Suppose $\alpha < 1/3$ *or* $\alpha = 1/3$ *and* $\beta > \beta_1^*$. *Then, there exists* n_0 *such that for all* $n \geq n_0$, *the system from Eq.* (5) *admits three equilibria given by*

$$\overline{q}_n = n + \frac{\rho_n^2}{(1 - \rho_n)^2}, \tag{6}$$

$$\underline{q}_n = n - n \cdot (1 - \rho_n) \cdot r_{0,n}(\rho_n), \tag{7}$$

$$\tilde{q}_n = n - n \cdot (1 - \rho_n) \cdot r_{1,n}(\rho_n), \tag{8}$$

where

$$r_{i,n}(\rho_n) = \frac{4}{3} \cdot \cos\left(\frac{1}{3}\arccos\left(-\sqrt{\frac{27\rho_n^2}{4n\cdot(1-\rho_n)^3}}\right) - \frac{2\pi i}{3}\right)^2, \quad i \in \{0,1\}.$$

Furthermore, $\overline{q}_n$ and $\underline{q}_n$ are locally asymptotically stable and $\tilde{q}_n$ is an unstable equilibrium.

The result establishes that there are two fundamentally different regimes where the system from Eq. (5) can operate. When the system is heavily loaded, in the sense that $\alpha > 1/3$ or $\alpha = 1/3$ and $\beta < \beta_1^*$, then the queue length converges to a point $\overline{q}_n > n$ as t grows to ∞, independently of the initial condition. Furthermore the exact characterization of $\overline{q}_n$ provides additional insights. We have

$$\overline{q}_n = n + \frac{\rho_n^2}{(1-\rho_n)^2} \approx n + \frac{1}{\beta^2}\, n^{2\alpha}.$$

Hence, in such a system, asymptotically, there are always order $n^{2\alpha}$ customers waiting in the system to be served.

As the load decreases (α decreases) and when the system is such that $\alpha < 1/3$ or $\alpha = 1/3$ and $\beta > \beta_1^*$, then the behavior of the system is more subtle. There are two locally stable equilibria and one unstable equilibrium. Now the same equilibrium $\overline{q}_n$ still exists and is locally stable, but a new locally stable equilibrium emerges, $\underline{q}_n$. It is possible to show that this new equilibrium is such that[3]

$$\underline{q}_n \approx n - c\, n^{1-\alpha},$$

for an appropriate constant c. In other words, in such an equilibrium, there are always idle servers, and there is order $n^{1-\alpha}$ such idle servers. Hence, there are two locally stable equilibria, one with all servers busy and customers waiting ($\overline{q}_n$) and one with idle servers and no customers waiting ($\underline{q}_n$).

Proof sketch and intuition. The proof of Theorem 1 relies on analyzing both equilibrium points and their stability properties. To establish the equilibria, we determine the zero crossings of $f_n(\cdot)$. With some slight rewriting,

$$f_n(q) = \lambda_n - \mu_n(q)\cdot\min\{n,q\} = \lambda_n\left[1 - \left(\frac{1}{\sqrt{|q-n|\vee 1}} + 1\right)^{-1}\cdot\frac{\min\{n,q\}}{\lambda_n\bar{s}}\right] = \lambda_n\left[1 - \frac{g_{2,n}(q)}{g_{1,n}(q)}\right],$$

with $\quad g_{1,n}(q) = 1 + \dfrac{1}{\sqrt{|n-q|\vee 1}}, \qquad g_{2,n}(q) = \dfrac{\min(n,q)}{\lambda_n\bar{s}}.$

The function $g_{1,n}(q)$ is proportional to the amount of work a system with n servers needs to do per customer when there are q customers in the system. Analogously, $g_{2,n}(q)$ is proportional to the amount of work the system with n servers is capable of doing per customer when there are

[3] This can be seeing by analyzing the Taylor expansion of the term $r_{0,n}(\rho_n)$.

q customers in the system. Hence, determining the sign of $f_n(q)$ amounts to comparing the sizes of $g_{1,n}(q)$ and $g_{2,n}(q)$. When the former is larger than the latter, we have $f_n(q) > 0$ and the queue size grows. When the inverse is true, $f_n(q) < 0$, the queue size shrinks. When they are equal, we obtain an equilibrium point by solving for q. Figure 3 depicts the two functions for the two different regimes.

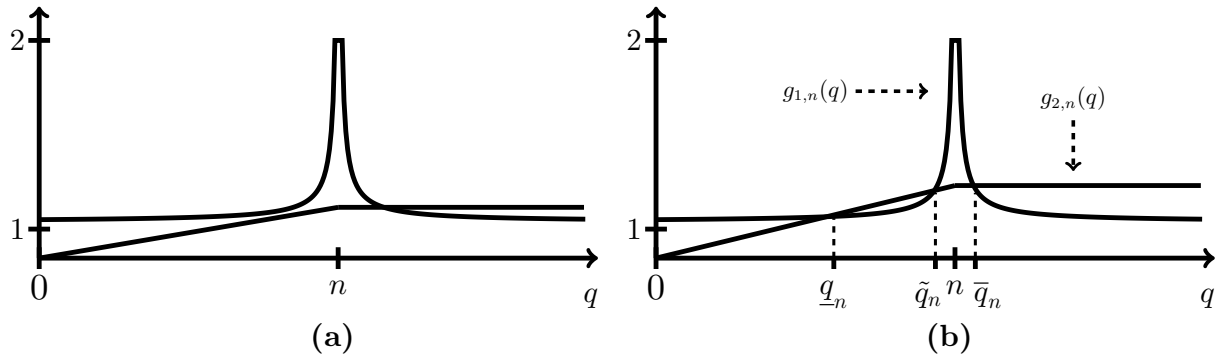


Figure 3 Equilibria points for system from Eq. (5). Plots (a) and (b) correspond to regimes (i) and (ii) from Theorem 1, respectively. The points where the functions $g_{1,n}(q)$ and $g_{2,n}(q)$ cross correspond to equilibria points.

As for stability, the queue length tends to grow when $g_{1,n}(q) > g_{2,n}(q)$ since the amount of work the system needs to perform per customer is greater than its ability to do work per customer. Similarly, $g_{1,n}(q) < g_{2,n}(q)$ implies the system can handle the current workload and that the queue size is decreasing. Therefore, the two equilibrium points in regime (ii) where $g_{1,n}(q) > g_{2,n}(q)$ to their left and $g_{1,n}(q) < g_{2,n}(q)$ to their right, $\underline{q}_n$ and $\overline{q}_n$, are stable, while $\tilde{q}_n$ is not.

An important observation is about what drives the differences between the regimes. From the heavy traffic scaling (see Eq. (3)) we have that $g_{2,n}(q) \approx q/(n - \beta \cdot n^{1-\alpha})$ for all $q < n$. It follows that for $q < n$ the slope of $g_{2,n}(q)$ is determined by both α and β. The theorem establishes that when α is large enough the slope of $g_{2,n}(q)$ is not steep enough to cross $g_{1,n}(q)$ and, therefore, the only possible equilibrium is $\overline{q}_n$ (See Figure 3 (a)). Similarly, if α is small enough then $g_{2,n}(q)$ is steep enough to cross $g_{1,n}(q)$; thus, the two extra equilibria $\underline{q}_n$ and $\tilde{q}_n$ emerge (See Figure 3 (b)). The transition point occurs when α equals 1/3. In this case, depending on the choice of β, the two extra equilibria may or may not exist. As β increases, the slope of $g_{2,n}(q)$ increases until it reaches a point from which on $g_{2,n}(q)$ is steep enough so that the two equilibria to the left of n materialize.

Interpretation in terms of the queueing system. In terms of the queuing model, when the number of customers is much larger than n, service times become shorter. In turn, the system processes customers more efficiently, which brings the total number of customers down. In addition, when the number of customers is close to n, service times are not as short as in the previous situation. This implies that the system is not as effective in processing customer, bringing the total

number of customers up. That is, the queueing system (and also the general system) has a self-regulating property that is captured by the deterministic system through the equilibrium $\overline{q}_n$. When the number of customers is low (when $q < \underline{q}_n$), despite the fact that each customer experiences a "short" pickup time, there are just not enough customers in the system so that the arrival rate dominates departure rate, which increases the number of customers in the system. For a medium number of customers (when $q \in (\underline{q}_n, \tilde{q}_n)$), there are enough idle servers so that we are processing customer efficiently, but also there are enough customers in the system so that arrivals can be dominated by departures. This brings the number of customers in the system down. For a large number of customers ($q \in (\tilde{q}_n, n)$), there are not enough idle servers. Therefore, the service time of customers becomes large and, as a consequence, so does the number of customers in the system. That is, for states below n, the queueing system also has a self-regulating property that is captured by the deterministic dynamics through the equilibrium $\underline{q}_n$. Therefore, one might expect $\underline{q}_n$ and $\overline{q}_n$ to play focal roles in the queueing system, which they indeed do when we analyze the stochastic version of the system in Section 5.

5. Limiting Regimes

In this section, we first investigate the properties of the Markovian system in steady state, where the equilibria derived in the previous section for the deterministic system from Eq. (5) will play a central role. We then analyze the system in the asymptotic regime from Eq. (3), parametrized by α and β. In turn, our results lead to a parametrization of the system's regimes: QD, ED and QED. We also discuss some managerial implications of the results.

5.1. Steady-State Analysis

Before we provide our main results, observe that for a given scale n, the process $Q_n(t)$ is a birth and death process with birth rate λ_n and state-dependent death rate $\mu_n(Q_n(u)) \cdot \min(n, Q_n(u))$. Letting $\pi_n(k)$ be the steady-state probability that the n-system is in state k, the detailed balance equations yield

$$\pi_n(k) \cdot \frac{f_n(k)}{\lambda_n} = \pi_n(k) - \pi_n(k-1), \quad k \geq 1. \tag{9}$$

We first characterize the shape of the steady-state distribution $\pi_n(\cdot)$ for systems with large scale.

PROPOSITION 1 **(Steady-state Probability Distribution)**. *Suppose that* $\lim_{n\to\infty}(1 - \rho_n)n^\alpha = \beta$, $\rho_n \uparrow 1$, *and let* $\beta_1^* = 3/(4^{1/3})$. *Then the following holds.*

(i) If $\alpha > 1/3$ *or if* $\alpha = 1/3$ *and* $\beta < \beta_1^*$, *then for* n *sufficiently large, the steady distribution* $\pi_n(\cdot)$ *is unimodular with a mode at* $\lfloor \overline{q}_n \rfloor$.

(ii) If $\alpha < 1/3$ or if $\alpha = 1/3$ and $\beta > \beta_1^*$, then for n sufficiently large, the steady distribution $\pi_n(\cdot)$ admits two modes, one at $\lfloor \underline{q}_n \rfloor$ and one at $\lfloor \overline{q}_n \rfloor$.

This result leverages Eq. (9) and the intuition obtained from Figure 3 to link the equilibria of the deterministic system from Eq. (5) with the modes of $\pi_n(k)$. From Eq. (9), we note that the monotonicity of $\pi_n(\cdot)$ can be determined by looking at the sign of $f_n(\cdot)$. In turn, Proposition 1 establishes that $\pi_n(\cdot)$ has at most two modes and that those modes are close to the equilibrium points. There is always one at $\lfloor \overline{q}_n \rfloor$, and, depending on the scaling parameters, there may or may not be another one at $\lfloor \underline{q}_n \rfloor$. We represent the two possibilities in Figure 4.

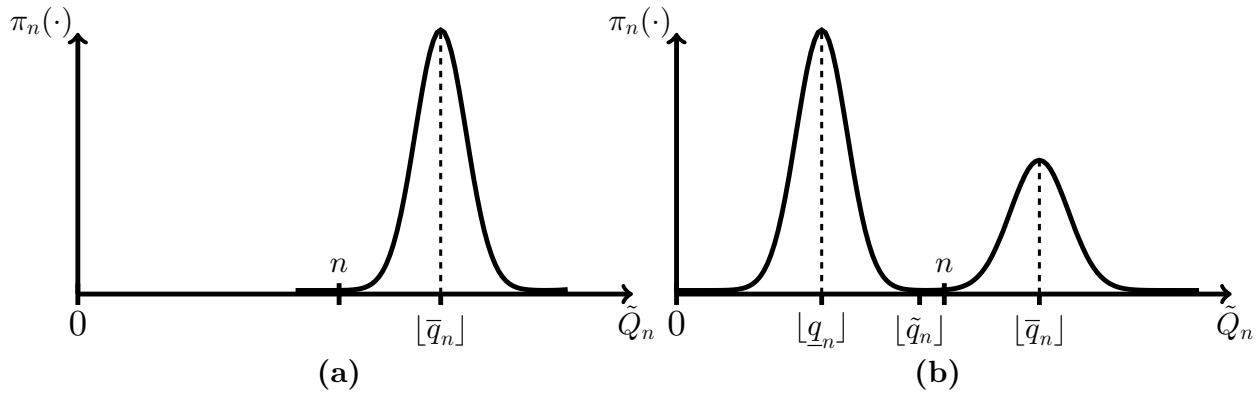


Figure 4 Steady-state probability $\pi_n(\cdot)$. In (a), which corresponds to regime (i) in Proposition 1, the state distribution is unimodal with a peak at $\lfloor \overline{q}_n \rfloor$. In (b), which corresponds to regime (ii) in Proposition 1, the state distribution is bimodal with peaks at $\lfloor \underline{q}_n \rfloor$ and $\lfloor \overline{q}_n \rfloor$.

Whenever $\alpha > 1/3$, $\pi_n(\cdot)$ is unimodal and it peaks once to the right of n, see Figure 4(a). If $\alpha < 1/3$, $\pi_n(\cdot)$ is bimodal and it also peaks to the left of n, see Figure 4(b). If $\alpha = 1/3$ these two cases are possible depending on the parameter β. This is in line with the intuition we obtained from the deterministic analysis in Section 4.

In steady-state, one expects that the system spends most of the time around the modes of the distribution. However, when assessing the performance of the system in terms of probability of having to wait for a server to be assigned, one needs to analyze the steady-state distribution beyond its modes to evaluate how mass is distributed. We do so next.

5.2. Service Regimes

We start our analysis of service regimes by analyzing the quality-driven (QD) and effiency-driven (ED) regimes.

5.2.1. QD and ED regimes. We first establish sufficient conditions for the ED and QD regimes to emerge.

THEOREM 2 **(Limiting Regimes)**. *Fix $\alpha \in (0,1)$ and $\beta > 0$. Suppose that $\lim_{n\to\infty} n^\alpha(1-\rho_n) = \beta$. Then, there exists $\beta_2^* > \beta_1^*$ such that*

(i) (ED Regime) if $\alpha \in (1/3,1)$ or if $\alpha = 1/3$ and $\beta < \beta_2^*$, then

$$P_\infty(W) = 1,$$

(ii) (QD Regime) if $\alpha \in (0,1/3)$ or if $\alpha = 1/3$ and $\beta > \beta_2^*$, then

$$P_\infty(W) = 0.$$

Theorem 2 provides a crisp characterization of the domains in which the ED and QD regimes emerge. If $\alpha \in (1/3,1)$ or if $\alpha = 1/3$ and $\beta < \beta_2^*$, then recall from Proposition 1 that the steady-state probability of the number of customers in the system admits only one mode at $\lfloor \overline{q}_n \rfloor$, which is higher than n, the number of servers. Part *(i)* of Theorem 2 establishes that the mass is concentrated to the right of n and hence servers are almost always either en route to customers or transporting customers and almost never idle. In turn, customers, will have to wait with probability close to 1 before being assigned a server.

If $\alpha \in (0,1/3)$ or if $\alpha = 1/3$ and $\beta > \beta_2^*$, then the the steady-state probability of the number of customers in the system admits two modes (cf. Proposition 1 part *(ii)*), one at $\lfloor \overline{q}_n \rfloor$ which is higher than n and one at $\lfloor \underline{q}_n \rfloor$ which is lower than n. Part *(ii)* of Theorem 2 establishes that the mass is concentrated to the left of n and hence there is almost always a fraction of servers that idle and customers almost never wait before being assigned a server. In other words, the mode to the right of n plays little role in this parameter regime.

Discussion of Capacity Planning. To further appreciate the result, recall that since $n^\alpha(1-\rho_n) \to \beta$ we have

$$\frac{n - \lambda_n \bar{s}}{(\lambda_n \bar{s})^{1-\alpha}} \to \beta, \quad \text{that is,} \quad n \approx \lambda_n \bar{s} \; + \; \beta \cdot (\lambda_n \bar{s})^{1-\alpha}. \tag{10}$$

The term $\lambda_n \bar{s}$ corresponds to the standard offered load of the system as defined for standard $M/M/n$ multi-server systems. In heavy traffic, this quantity determines how the capacity of the system should be scaled with the arrival rate of customers. First, there is a nominal term, which is simply $\lambda_n \bar{s}$, that accounts for the expected amount of work requested by customers. The second term $\beta \cdot (\lambda_n \bar{s})^{1-\alpha}$ is a buffer term that accounts for stochastic variations of the system. In a classical $M/M/n$ setting, when $\alpha < 1/2$, the system is in the QD regime, when $\alpha > 1/2$, the system is in the ED regime, and when $\alpha = 1/2$ the system is in the QED regime. In contrast, in our setting when the buffer term is $\beta \cdot (\lambda_n \bar{s})^{1/2}$, the system is in the ED regime no matter the choice of β. Since our

model captures spatial frictions, this result highlights that in a setting where servers need to reach customers before the start of effective service, the capacity needed to achieve QED performance is fundamentally different than in a standard setting. Moreover, spatial frictions create the need for more servers than in a standard setting for the system to operate in the QD regime. Indeed, in our model the buffer term must be $\beta \cdot (\lambda_n \bar{s})^m$ with $m \geq 2/3$. The transition between ED and QD occurs when the buffer term is $\beta \cdot (\lambda_n \bar{s})^{2/3}$, that is, the QED regime can only happen with a scaling of 2/3 which is orders of magnitude larger than the traditional SRS rule of thumb.

Proof sketch of Theorem 2. The proof of Theorem 2 consists on bounding above the terms $\mathbf{P}[Q_n(\infty) < n]$ and $\mathbf{P}[Q_n(\infty) \geq n]$, respectively, and then using asymptotic relations between the mode probabilities as established in the following result.

PROPOSITION 2. *Fix $\alpha \in (0,1)$ and $\beta > 0$. Suppose that $\lim_{n\to\infty} n^\alpha(1-\rho_n) = \beta$ then*

(i)

$$\lim_{n\to\infty} \frac{1}{n^\alpha} \log\Big(\frac{\pi_n(\lfloor \overline{q}_n \rfloor)}{\pi_n(n)}\Big) = \frac{1}{\beta},$$

(ii) if $\alpha < 1/3$ then

$$\lim_{n\to\infty} \frac{1}{n^{1-2\alpha}} \log\Big(\frac{\pi_n(\lfloor \overline{q}_n \rfloor)}{\pi_n(\lfloor \underline{q}_n \rfloor)}\Big) = -\frac{\beta^2}{2},$$

(iii) if $\alpha = 1/3$ then there exists a function $g(\cdot)$ such that

$$\lim_{n\to\infty} \frac{1}{n^{1/3}} \log\Big(\frac{\pi_n(\lfloor \overline{q}_n \rfloor)}{\pi_n(\lfloor \underline{q}_n \rfloor)}\Big) = g(\beta).$$

And there exists $\beta_2^* > \beta_1^*$ such that $g(\beta_2^*) = 0$ and if $\beta_1^* < \beta < \beta_2^*$ then $g(\beta) > 0$, whereas if $\beta > \beta_2^*$ then $g(\beta) < 0$.

Proposition 2 shows how the peak of the modes of $\pi_n(\cdot)$ compare to each other as n grows large. When $\alpha > 1/3$, for large n, there is only one peak given by $\lfloor \overline{q}_n \rfloor$. From part *(i)*, its steady-state probability satisfies

$$\pi_n(\lfloor \overline{q}_n \rfloor) \approx \pi_n(n) \cdot e^{n^\alpha/\beta},$$

that is, $\pi_n(\lfloor \overline{q}_n \rfloor)$ is exponentially larger than $\pi_n(n)$. Since $\pi_n(\cdot)$ is increasing to the left of $\lfloor \overline{q}_n \rfloor$ (see Proposition 1), this suggests that, in the limit, the number of customers in the system will be above n with high probabilty. In other words, the system will be in the ED regime.

For the case when $\alpha < 1/3$, Proposition 1 states that $\pi_n(\cdot)$ is bimodal and, therefore, there could be mass around both peaks. However, part *(ii)* of the proposition establishes that $\pi_n(\lfloor \underline{q}_n \rfloor)$ is exponentially larger than $\pi_n(\lfloor \overline{q}_n \rfloor)$,

$$\pi_n(\lfloor \underline{q}_n \rfloor) \approx \pi_n(\lfloor \overline{q}_n \rfloor) \cdot e^{\frac{1}{2}\beta^2 n^{1-2\alpha}}.$$

This suggests that when $\alpha > 1/3$, the tail of $\pi_n(\cdot)$ to the right of n vanishes as n becomes large. In turn, the number of customers in the system should be below n with high probability. In other words, while the distribution $\pi_n(\cdot)$, has two modes, only one mode "matters" and we expect the system to be in the QD regime.

The threshold case is $\alpha = 1/3$. In this case whether $\pi_n(\lfloor \overline{q}_n \rfloor)$ dominates $\pi_n(\lfloor \underline{q}_n \rfloor)$ (or vice-versa) is governed by β. When $\beta < \beta_1^*$, from Proposition 1, we know that $\lfloor \overline{q}_n \rfloor$ is the only mode and, therefore, $\pi_n(\lfloor \overline{q}_n \rfloor)$ dominates. If $\beta \in (\beta_1^*, \beta_2^*)$ then $\lfloor \underline{q}_n \rfloor$ is also a mode; however, part *(iii)* of the proposition establishes that $\pi_n(\lfloor \overline{q}_n \rfloor)$ is exponentially larger than $\pi_n(\lfloor \underline{q}_n \rfloor)$. That is, in this case $\lfloor \underline{q}_n \rfloor$ transitions into becoming a mode, but the mass it contributes is not large enough and it vanishes as n increases. Therefore, for $\beta < \beta_2^*$, the system will be in the ED regime. In contrast, when $\beta > \beta_2^*$, the roles of $\pi_n(\lfloor \overline{q}_n \rfloor)$ and $\pi_n(\lfloor \underline{q}_n \rfloor)$ reverse. This indicates that for $\beta > \beta_2^*$, the system will be in the QD regime.

5.2.2. QED regime Theorem 2 implies that the QED regime, in which the asymptotic probability that customers have to wait for a server to be assigned is such that $P_\infty(W) \in (0,1)$, may only occur if $\alpha = 1/3$ and $\beta = \beta_2^*$ as for all other values, the system is either in the ED or QD regimes. It is already apparent that the QED regime is much more subtle in our Markovian system than in classical $M/M/n$ systems as both the buffer order of magnitude (determined by α) *and* the constant in front of the buffer size (determined by β) need to be pinned down. The transition from QD to ED regimes does not occur through the constants in front of the buffer order of magnitude, leaving the question open of whether the QED regime exists at all in our Markovian system and, if so, how may it be reached. The next result establishes that there exists a QED regime and provides a characterization of it.

THEOREM 3 **(QED Regime)**. *Let $p_H \in (0,1)$. There exists a sequence $\{\gamma_n : n \geq 1\}$ with $\gamma_n \to 0$ as $n \uparrow \infty$ and a function $p_L(p_H) \in (0,1)$, such that if $n^{1/3}(1-\rho_n) = \beta_2^* + \gamma_n$ then*

$$p_L(p_H) \leq \liminf_{n\to\infty} \mathbf{P}[Q_n(\infty) \geq n] \leq \limsup_{n\to\infty} \mathbf{P}[Q_n(\infty) \geq n] \leq p_H,$$

with $p_L(\cdot)$ strictly increasing in p_H and such that $\lim_{p_H \to 1} p_L(p_H) = 1$ and $\lim_{p_H \to 0} p_L(p_H) = 0$.

This result establishes a regime such that for n large enough the probability of waiting to be assigned a server is in $(0,1)$. In turn, the probability of not waiting also belongs to $(0,1)$. That is, the system is in the QED regime. We have not pinned down an exact expression for these probabilities but, instead, we have provided a range. As one varies $p_H \in (0,1)$, one can achieve the extreme regimes. If $p_H \approx 1$ then from the theorem we can deduce that $\mathbf{P}[Q_n(\infty) \geq n] \approx 1$; if $p_H \approx 0$ then we can deduce that $\mathbf{P}[Q_n(\infty) \geq n] \approx 0$.

Capacity Planning for the QED Regime. From a practical perspective, Theorem 3 provides two important insights. First, it shows that QED performance is achieved at a different scaling than in traditional multi-server systems. Typically, in those system a SRS rule can balance the trade-off between waiting times and service efficiency. In a spatial setting this is no longer enough because servers must reach their customers before starting service. Our results suggest that the right scaling is 2/3 instead of 1/2. Second, notice that since $n^{1/3}(1-\rho_n)-\gamma_n \to \beta_2^*$ we have

$$\frac{n-\lambda_n\bar{s}}{(\lambda_n\bar{s})^{1-\alpha}}-\gamma_n \to \beta_2^*, \quad \text{that is,} \quad n \approx \lambda_n\bar{s} \;+\; \beta_2^*\cdot(\lambda_n\bar{s})^{2/3} +\; \gamma_n\cdot(\lambda_n\bar{s})^{2/3}. \tag{11}$$

From this equation we observe that, in addition to the traditional buffer term of the form $\beta\cdot(\lambda_n\bar{s})^m$, our result establishes that an extra lower order term is needed for QED performance. In particular, in our Markovian system, it is necessary to add the term $\gamma_n\cdot(\lambda_n\bar{s})^{2/3}$. Because $\gamma_n\to 0$ this term is of lower order than the second term in Eq. (11). Hence, the QED regime requires a very fine balance involving second order terms compared to the buffer size in this spatial setting, in stark contrast with the classical $M/M/n$ setting.

Proof sketch of Theorem 3. A necessary condition to achieve the QED regime is that the peaks of $\pi_n(\cdot)$ be in a constant proportion; otherwise, one would dominate the other and the system would be in the QD or ED regime. According to Proposition 2 part *(iii)*, this can only happen when $\alpha=1/3$ and $\beta=\beta_2^*$. In this case

$$\lim_{n\to\infty}\frac{1}{n^{1/3}}\log\Big(\frac{\pi_n(\lfloor\overline{q}_n\rfloor)}{\pi_n(\lfloor\underline{q}_n\rfloor)}\Big)=0,$$

that this, the $\log(\cdot)$ term is $o(n^{1/3})$. In turn, the ratio $\pi_n(\lfloor\overline{q}_n\rfloor)/\pi_n(\lfloor\underline{q}_n\rfloor)$ does not necessarily converge to a constant. To have it so, one would have to look at lower order terms for $\log(\pi_n(\lfloor\overline{q}_n\rfloor)/\pi_n(\lfloor\underline{q}_n\rfloor))$ and try to disentangle the exact rate at which $n^{1/3}(1-\rho_n)$ has to approach β_2^* so that the $\log(\cdot)$ converges to a constant. Instead of pursuing this, in the next result we show the existence of a sequence converging to zero, $\{\gamma_n^c : n\geq 1\}$, such that if $n^{1/3}(1-\rho_n)$ approaches β_2^* as $\beta_2^*+\gamma_n^c$, the peaks of $\pi_n(\cdot)$ will be in a constant proportion.

Proposition 3. *Fix $c\in\mathbb{R}$. Then, there exists a sequence $\{\gamma_n^c : n\geq 1\}$ with $\gamma_n^c\to 0$ such that if $n^{1/3}(1-\rho_n)=\beta_2^*+\gamma_n^c$, then*

$$\lim_{n\to\infty}\log\Big(\frac{\pi_n(\lfloor\overline{q}_n\rfloor)}{\pi_n(\lfloor\underline{q}_n\rfloor)}\Big)=c.$$

In the proof of the proposition we provide a detailed explanation of how to construct the sequence $\{\gamma_n^c : n\geq 1\}$. In turn, the proposition is not just an existence result, but it also provides the exact sequence that enables us to maintain the peaks in a constant proportion. It also establishes that, for any constant $c\in\mathbb{R}$, if $n^{1/3}(1-\rho_n)$ approaches β_2^* at an appropriate rate then

$$\pi_n(\lfloor\overline{q}_n\rfloor)\approx\pi_n(\lfloor\underline{q}_n\rfloor)\cdot e^c.$$

In particular, as we vary c we can achieve any desired proportion. For example, if $c<0$ then $\pi_n(\cdot)$ might look as depicted in Figure 4(b).

Even though there is a way to scale the system such that that the peaks are in constant proportion, this does not guarantee that the probability of being around each of them will be positive at the same time. It is possible, for example, that the dispersion of $\pi_n(\cdot)$ around $\lfloor \overline{q}_n \rfloor$ diminishes to zero while the proportion with the other peak remains constant. Therefore, we need to assess how the peaks compare to the mass around them. The next lemma provides a characterization of this.

LEMMA 2. *Fix $\alpha \in (0,1)$ and $\beta > 0$. Suppose that $\lim_{n\to\infty} n^{\alpha}(1-\rho_n) = \beta$, then*

(i)

$$\frac{\mathbf{P}[Q_n(\infty) \geq n]}{\pi_n(\lfloor \overline{q}_n \rfloor)} = \Theta(n^{\frac{3}{2}\alpha}).$$

(ii) if $\alpha \in (0,1/3)$, or $\alpha = 1/3$ and $\beta > \beta_1^*$, then

$$\frac{\mathbf{P}[Q_n(\infty) < n]}{\pi_n(\lfloor \underline{q}_n \rfloor)} = \Theta(\sqrt{n}).$$

This result establishes that the ratio of the mass to the right of n, to the peak in that region is $\Theta(n^{\frac{3}{2}\alpha})$. That is, with respect to $\pi_n(\lfloor \overline{q}_n \rfloor)$ the mass to the right of n is not negligible and, in fact, is approximately $n^{\frac{3}{2}\alpha}$ larger than $\pi_n(\lfloor \overline{q}_n \rfloor)$. Similarly, with respect to $\pi_n(\lfloor \underline{q}_n \rfloor)$, the mass to the left of n is non-trivial and, in fact, is approximately $\sqrt{n}$ larger than $\pi_n(\lfloor \underline{q}_n \rfloor)$.

Observe that in part *(i)* of the lemma, the order of the ratio depends on α. When $\alpha < 1/3$ then this ratio is not as big as the one for $\pi_n(\lfloor \underline{q}_n \rfloor)$ (which is $\Theta(\sqrt{n})$). This coincides with Theorem 2 in that for these values of α the mass to the left of n dominates the mass to the its right. Similarly, when $\alpha > 1/3$, the mass to the right of n dominates. For $\alpha = 1/3$, both ratios are of the same order. In turn, we have

$$\frac{\mathbf{P}[Q_n(\infty) \geq n]}{\mathbf{P}[Q_n(\infty) < n]} = \Theta\left(\frac{\pi_n(\lfloor \overline{q}_n \rfloor)}{\pi_n(\lfloor \underline{q}_n \rfloor)}\right).$$

Therefore, if the ratio of the peaks is constant, then the total mass to the left *and* to the right of n can be both (asymptotically) positive and separated away from zero. That is, both sides can be "balanced" whenever the peaks are in constant proportion. We can thus combine the results in Proposition 3 and Lemma 2 to find lower and upper bounds for $\mathbf{P}[Q_n(\infty) \geq n]$. In the proof of Lemma 2 we find exact expressions to control for the ratios as n increases, which we then leverage to provide explicit bounds for $\mathbf{P}[Q_n(\infty) \geq n]$ that can be mapped to probability values, p_H and $p_L(p_H)$, which satisfy the properties of Theorem 3.

5.3. Orders of Magnitudes of Queues and Wait Times

The results so far provide an understanding of the different regimes the system can operate in as a function of its load. Next, we quantify queue sizes and waiting times in our system as a function the scaling parameter α. The discussion in this section underlines the differences of a spatial server system with a traditional queueing system.

Let L^s and W^s denote respectively the steady-state expected queue length (excluding customers in service) and expected wait time. Similarly, let L^c and W^c denote the corresponding quantities in the classical $M/M/n$ system. From standard queueing theory, we have that

$$L^c = \frac{\rho_n}{(1-\rho_n)} \cdot C(n, \lambda_n/\bar{s}_t),$$

where $C(n, \lambda_n/\bar{s}_t)$ satisfies the Erlang's C formula, and represents the probability of waiting (see, e.g., Allen (2014)). Assuming that $n^\alpha(1-\rho_n) \to \beta$ we have that

$$C(n, \lambda_n/\bar{s}_t) \to \begin{cases} 1 & \text{if } \alpha > 1/2, \\ \text{constant} & \text{if } \alpha = 1/2, \\ 0 & \text{if } \alpha < 1/2. \end{cases}$$

In turn, using standard arguments, one can show that for $\alpha < 1/2$, we have that L^c is $o(1)$. Meanwhile, for $\alpha \geq 1/2$, L^c is $\Theta(n^\alpha)$. This implies that for $\alpha < 1/2$, W^c is $o(1)$, while for $\alpha \geq 1/2$, W^c is $\Theta(n^{\alpha-1})$. In particular, in the Halfin-Whitt regime ($\alpha = 1/2$), we have that L^c is $\Theta(\sqrt{n})$ and W^c is $\Theta(1/\sqrt{n})$. Next, we compare these classic results with the results obtained from our Markovian system.

We first provide a rigorous statement about the order of magnitude of the size of our Markovian system around the equilibria, in the sense of deriving the subset of the real line where the queue lengths fluctuations are constrained to, assuming n is sufficiently large. We use this result to provide approximate expressions for L^s and W^s.

PROPOSITION 4. *Suppose* $\lim_{n\to\infty} n^\alpha(1-\rho_n) = \beta$. *Then,*

(i) If $\alpha \in (1/3, 1)$ *or if* $\alpha = 1/3$ *and* $\beta < \beta_2^*$ *then there exists* $C > 0$ *such that*

$$\lim_{n\to\infty} \mathbf{P}\Big[-C \leq \frac{Q_n(\infty) - \lfloor \overline{q}_n \rfloor}{\sqrt{\log(n)} \cdot n^{\frac{3}{2}\alpha}} \leq C \Big] = 1.$$

(ii) If $\alpha \in (0, 1/3)$ *or if* $\alpha = 1/3$ *and* $\beta > \beta_2^*$ *then there exists* $C > 0$ *such that*

$$\lim_{n\to\infty} \mathbf{P}\Big[-C \leq \frac{Q_n(\infty) - \lfloor \underline{q}_n \rfloor}{\sqrt{\log(n)} \cdot \sqrt{n}} \leq C \Big] = 1.$$

Let's consider first part *(i)* of the proposition. In this case we can use Eq. (6) to deduce that

$$L^s \approx r^2 \frac{\rho_n^2}{(1-\rho_n)^2} \ \pm \ C \cdot n^{\frac{3}{2}\alpha} \cdot \sqrt{\log(n)} = \Theta(n^{2\alpha}),$$

Little's law delivers

$$W^s \approx r^2 \frac{\rho^2}{\lambda_n(1-\rho)^2} \pm \frac{C}{\lambda_n} \cdot n^{\frac{3}{2}\alpha} \cdot \sqrt{\log(n)} = \Theta(n^{2\alpha-1}).$$

There are several interesting observations. First, for $\alpha = 1/2$, the queue size is approximately $\Theta(n)$ and the wait time is approximately $\Theta(1)$. Note the contrast to a classical $M/M/n$ system, where $L^c = \Theta(\sqrt{n})$ and $W^c = \Theta(1/\sqrt{n})$. This makes precise how much more work we are adding to the system by including pickups. It also highlights that for $\alpha = 1/2$, the Markovian system is in the ED regime, with its long queues. Second, note that $\alpha = 1/2$ is the largest value for which W^s does not explode. In contrast, in the $M/M/n$ system, for any $\alpha \in (1/2, 1)$, the expected waiting time approaches zero.

If we focus on pickup times, we can gain further intuition about how the QED regime works in our system. Let P^s denote the expected pickup time. Then, from part *(i)* of the proposition,

$$P^s \approx \frac{\bar{s}}{\sqrt{|Q_n(\infty) - n| \vee 1}} \approx \Theta(1/n^\alpha).$$

For $\alpha = 1/3$, pickup times are of order $1/n^\alpha$ and W^s is of order $n^{2\alpha-1}$. This showcases the interplay between wait times and pickup times. When the load of the system increases (as measured by α), wait times increase because of the greater number of customers in the system, while pickup times decrease due to the increased spatial density of customers. If one attempts to minimize expected customer system times, we therefore need to balance W^s and P^s. For the regime where $\alpha \geq 1/3$, this occurs at $\alpha = 1/3$.

For the regime from part *(ii)* of the proposition, we have that $L^s \approx 0$ and $W^s \approx 0$. Moreover, we can use the fact that $\underline{q}_n \approx n - \Theta(n^{1-\alpha})$ to deduce that the expected number of idle server is $\Theta(n^{1-\alpha})$ and $P^s \approx \Theta(1/(n^{\frac{1-\alpha}{2}}))$. As we increase the load in the system (as measured by α), we reduce the number of idle servers. However, at the same time, pickup times increase due to the decreased spatial density of servers.

5.4. A Social Planner's Perspective

An alternative approach to determining the proper safety staffing level is to start from a social planner's objective, and then find the staffing level that optimizes it. A natural social planner's objective is one that incurs a cost per server of building capacity plus a waiting (and pick-up time) cost per customer. We now show that this objective function also leads us to the conclusion that a safety staffing that is equal to the offered load to the power of 2/3 is optimal.

Let us consider a service provider that pays c_s per unit of capacity and customers that incur a waiting cost of c_w per unit of waiting. That is, a social planner would like to select the level of capacity n that solves the following optimization problem

$$\min_n \quad c_s \cdot n + \lambda \cdot c_w \cdot \mathbf{E}[P_n + W_n]. \tag{12}$$

The first term in Eq. (12) corresponds to the cost of having n servers in the system. The second, to the cost experienced by customers while they wait to be assigned a server, W_n, and to be picked up, P_n.

Notice that from Eq. (3) we can write n as $\lambda \cdot \bar{s} + \beta(\lambda \cdot \bar{s})^{1-\alpha}$. Now, depending on our choice of α we can have one of two cases. When $\alpha \geq 1/3$, the average pick up times are of order $\Theta((\lambda \cdot \bar{s})^{-\alpha})$ while average waiting times are of order $\Theta((\lambda \cdot \bar{s})^{2\alpha-1})$. Replacing this in Eq. (12) delivers the following expression for the objective

$$c_s \cdot (\lambda \cdot \bar{s} + \beta(\lambda \cdot \bar{s})^{1-\alpha}) + c_w \cdot ((\lambda \cdot \bar{s})^{1-\alpha} + (\lambda \cdot \bar{s})^{2\alpha}).$$

Among all values $\alpha > 1/3$ the term that dominates in the expression above is the total waiting times, that is, $(\lambda \cdot \bar{s})^{2\alpha}$. This is increasing in α. Hence, $\alpha = 1/3$ leads to lower (asymptotic) costs compared to all values of $\alpha > 1/3$.

For the case α $le1/3$, let $\underline{\pi}_\lambda$ be the steady state probability the number of customers being below n and let $\overline{\pi}_\lambda$ be $1 - \underline{\pi}_\lambda$. Similar to the case when $\alpha > 1/3$, we can rewrite the objective in Eq. (12) to obtain

$$c_s \cdot (\lambda \cdot \bar{s} + \beta(\lambda \cdot \bar{s})^{1-\alpha}) + c_w \cdot \Big(\{\underline{\pi}_\lambda \cdot (\lambda \cdot \bar{s})^{\frac{1+\alpha}{2}} + \overline{\pi}_\lambda \cdot (\lambda \cdot \bar{s})^{1-\alpha}\} + \{\underline{\pi}_\lambda \cdot 0 + \overline{\pi}_\lambda (\lambda \cdot \bar{s})^{2\alpha}\} \Big).$$

When $\alpha < 1/3$ the term that dominates is of order $(\lambda \cdot \bar{s})^{1-\alpha}$. This term is decreasing in α. In this case, $\alpha = 1/3$ leads to lower (asymptotic) costs compared to all values of $\alpha \leq 1/3$.

In conclusion, in a large system, the system total social cost measured by capacity cost and waiting cost will be minimized by selecting the number of servers n according to $\lambda \cdot \bar{s} + \beta(\lambda \cdot \bar{s})^{2/3}$, where β should be tuned.

6. Numerical Experiments and General Simulation

In this section, we aim at (i) illustrating the results in the Markovian system (§6.1), and also to (ii) compare the behavior obtained in the Markovian system to that of the actual physical system that motivated the Markovian system (§6.2).

Simulation setup. We consider a square city $\mathcal{C} = [0,2] \times [0,2]$ and assume $v = 1$, implying that $\bar{s}_t \cdot v \approx 1.0428$. The time horizon will be $T = 4{,}000$. We simulate the general system introduced in Section 3 and the Markovian system under several different conditions, starting from $Q_n(0) = 0$, in order to capture the ED, QD and QED regimes. We scale the number of servers in the system according to

$$n = \lceil \lambda \bar{s}_t \ + \ \beta \cdot (\lambda \bar{s}_t)^{1-\alpha} \rceil. \tag{13}$$

For $\alpha \in \{1/4, 1/2\}$, we consider $\beta = 2.1$. For $\alpha = 1/3$, we vary $\beta \in \{2.1, 2.4, 2.7\}$.

6.1. Markovian System

We begin by numerically illustrating our theoretical results for the Markovian system. We consider the rate

$$\frac{1}{\mu(q)} = \frac{\bar{s}_p}{\sqrt{|q-n| \vee 1}} \; + \; \bar{s}_t, \quad q \geq 0, \tag{14}$$

with $\bar{s}_p = \bar{s}_t = 1.0428$, that is, the coefficient in front of the pickup times coincides with the expected travel time between two points. Recall from §3.1 that these two parameters need not to be the same because $\bar{s}_p$ comes from an asymptotic approximation. In the next section we consider more realistic values for $\bar{s}_p$ that we estimate from simulation.

In Figures 5-6, we depict sample paths of the the number of customers in the system minus the number of servers for the various parameters and superimpose a corresponding histogram (taken from the path between periods 500 and 4,000). Furthermore, the two modes $\lfloor \overline{q}_n \rfloor$ and $\lfloor \underline{q}_n \rfloor$ (when they exist) minus n are depicted.

In Figure 5(a), $\alpha = 0.25$ and we depict the system for three different scales. In line with Theorem 2, one observes that the system spends almost all its time around $\lfloor \underline{q}_n \rfloor$ and as the scale increases, the probability of wait approaches zero. The system is in the QD regime.

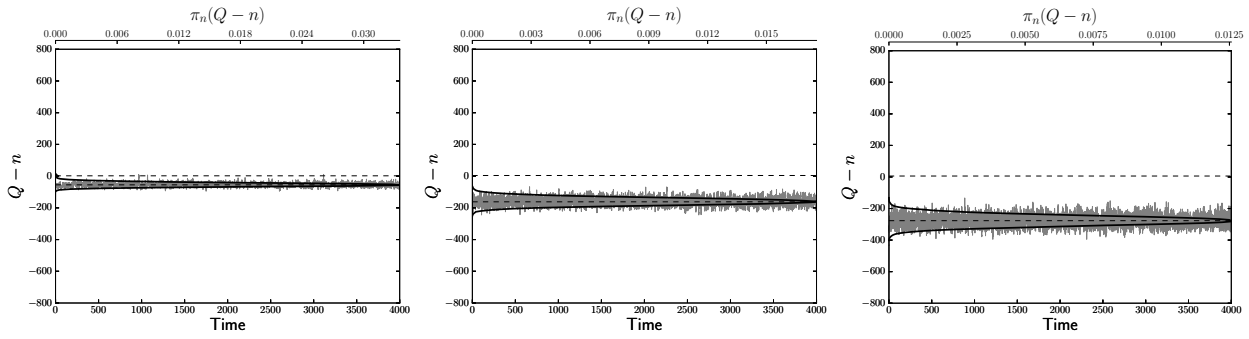
In Figure 5(b), $\alpha = 0.5$ and we depict the system for three different scales. Note that in this case, there is only one mode, $\lfloor \overline{q}_n \rfloor$. In line with Theorem 2, one observes that the system spends almost all its time around $\lfloor \overline{q}_n \rfloor$ and as the scale increases, the probability of wait approaches 1. The system is in the ED regime.

In Figure 6, $\alpha = 1/3$ and we depict the system for three values of β. This is the only setting where, asymptotically and depending on β, the system can oscillate between the two equilibria and asymptotically, a positive fraction of the customers (separated from 0 and 1) will wait before being assigned a server. Indeed, we observe that for small values of β, the system operates most often with $Q > n$, as in the ED regime. As β increases (center plot), the fraction of time the system spends in states such that $Q < n$ increases, in which case, the system is in the QED regime. When β increases further, the system enters the QD regime.

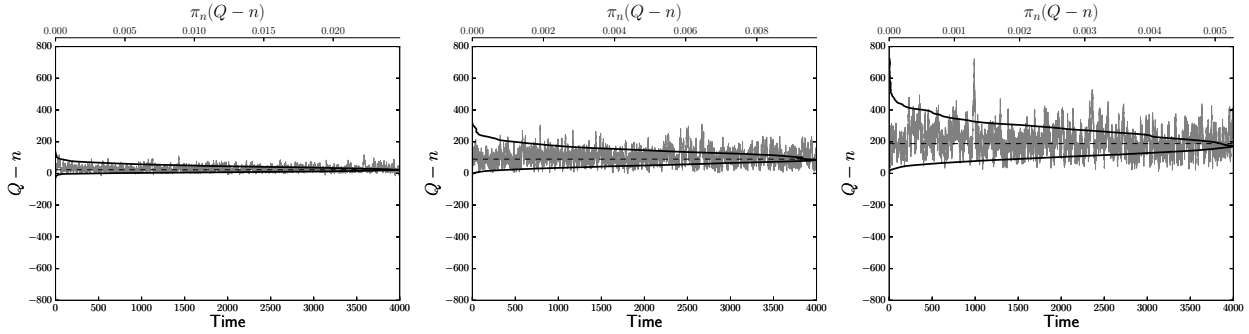
6.2. Comparing the General and Markovian Systems

Next we simulate the general system and compare it the Markovian system. Our purpose in this section is two-fold. First, we illustrate the system's behavior under the different scalings. In particular, we test whether for $\alpha < 1/3$ and $\alpha > 1/3$ the general system oscillates around the equilibria to the left and right of n, respectively. For $\alpha = 1/3$, we also test how by varying β the general system can, as predicted by the Markovian system, oscillate around both equilibria.

Second, we provide numerical evidence for the quality of the Markovian system as an approximation to the general system. To ensure an appropriate comparison, we proceed as follows:



(a) $\alpha = 0.25$



(b) $\alpha = 0.5$

Figure 5 **Simulation of the Markovian system. We consider $\beta = 2.1$ and from left to right $\lambda \in \{100, 400, 800\}$. The bottom $x-$axis corresponds to the simulation time, while the top $x-$axis corresponds to probabilities. In the figure we observe both a sample path and $\pi_n(\cdot)$. The dashed lines correspond to the modes $\lfloor \underline{q}_n \rfloor$ and $\lfloor \overline{q}_n \rfloor$ as given by Theorem 1.**

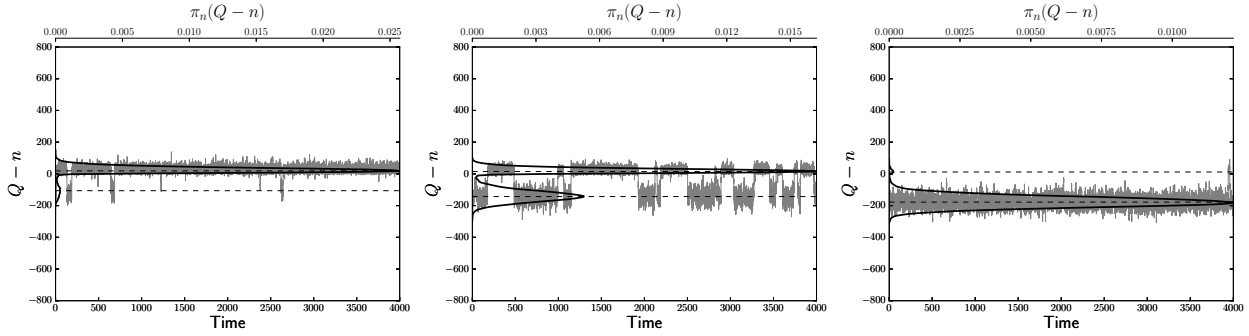


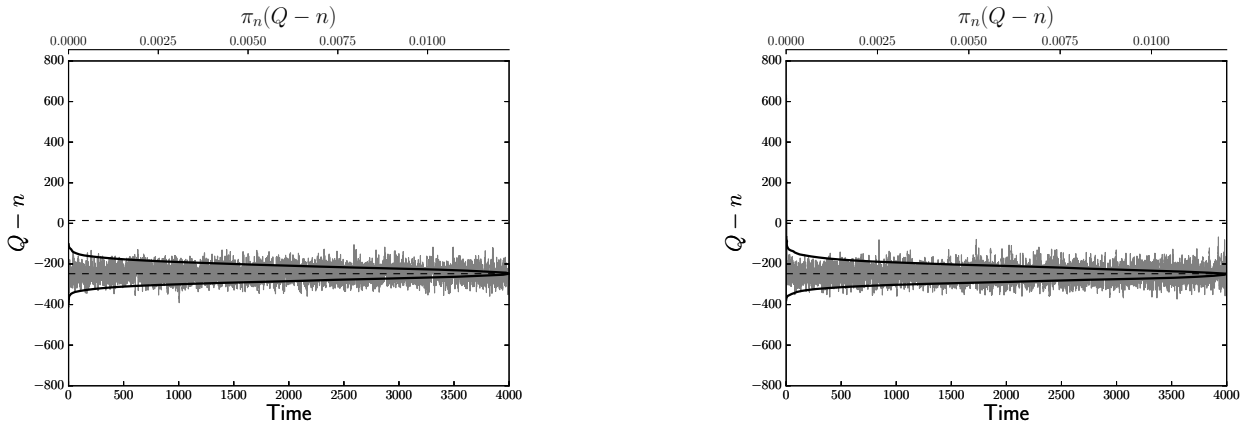
Figure 6 **Simulation of the Markovian system. We consider $\alpha = 1/3$ and $\lambda = 800$ and from left to right $\beta \in \{2.1, 2.4, 2.7\}$. The bottom $x-$axis corresponds to the simulation time, while the top $x-$axis corresponds to probabilities. In the figure we observe both a sample path and $\pi_n(\cdot)$. The dashed lines correspond to the modes $\lfloor \underline{q}_n \rfloor$ and $\lfloor \overline{q}_n \rfloor$ as given by Theorem 1.**

- We fix λ, α and β, and use Eq. (13) to obtain the number of servers.
- We simulate the general system for the computed value of n.
- We estimate $\bar{s}_p$, see Eq. (14). Then we simulate the Markovian system with rate given by Eq. (14), and compute the theoretical modes/equilibria.

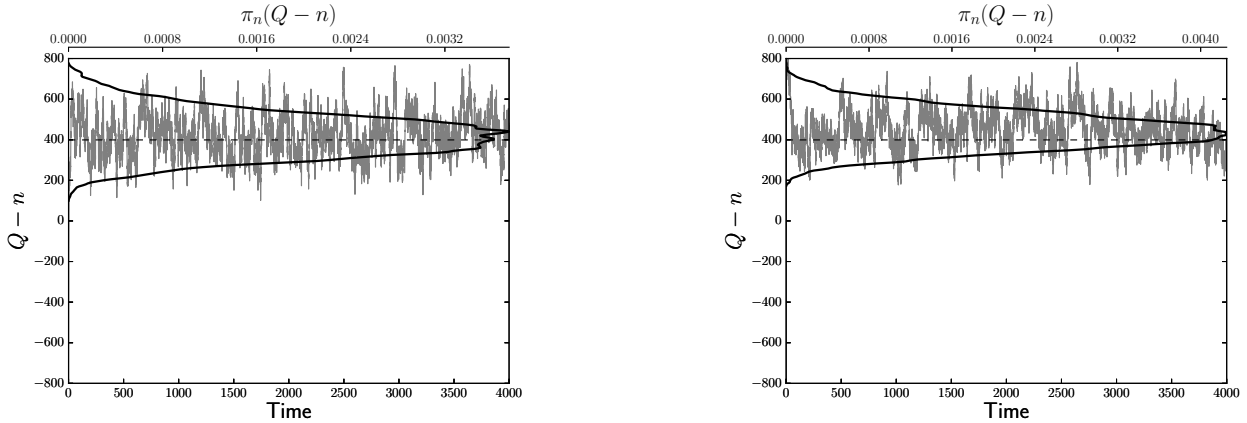
- We compare the system behavior for both the Markovian and general systems.

In Figures 7-8, we depict sample paths of the queue lengths in the general system (right column) and compare it to the Markovian system (left column). For the sake of exposition we fix $\lambda = 800$ throughout, but all the simulations are consistent for large values of λ.

We observe that for low α ($\alpha = 0.25$, Figure 7(a)), the general system queue admits a behavior very similar to the proposed Markovian approximation. In particular, the general system also admits a mode exactly around $\lfloor \underline{q}_n \rfloor$ (as predicted by the theory for the Markovian system) and this behavior is consistent across different scales.



(a) $\alpha = 0.25$, $n = 1161$, $\bar{s}_p = 1.531$, $\lfloor \underline{q}_n \rfloor - n = -248$, $\lfloor \overline{q}_n \rfloor - n = 14$



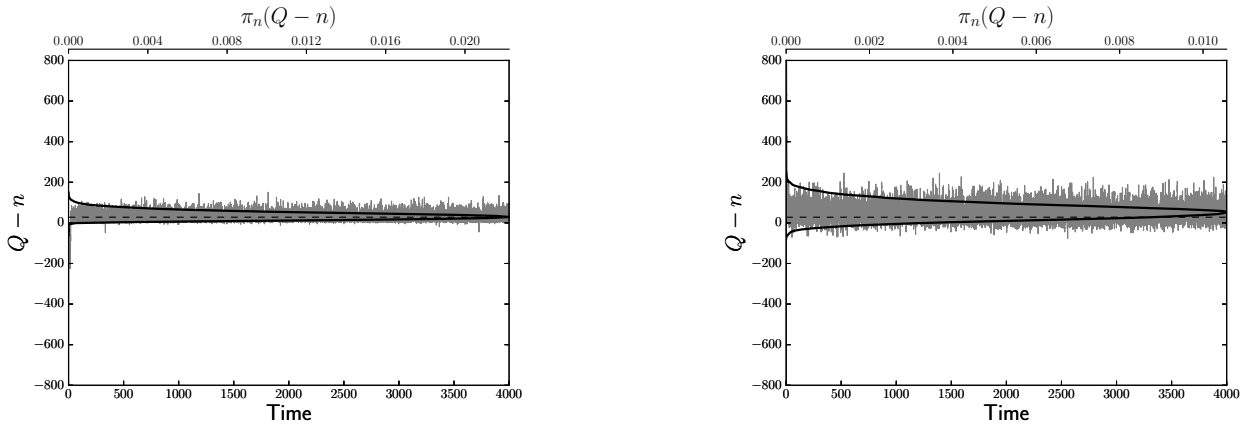
(b) $\alpha = 0.5$, $n = 895$, $\bar{s}_p = 1.518$, $\lfloor \overline{q}_n \rfloor - n = 399$

Figure 7 Simulation for Markovian (left) and General (right) systems. We consider $\beta = 2.1$. The bottom $x-$axis corresponds to the simulation time, while the top $x-$axis corresponds to probabilities. In the figure we observe both a sample path and $\pi_n(\cdot)$. The dashed lines correspond to the modes $\lfloor \underline{q}_n \rfloor$ and $\lfloor \overline{q}_n \rfloor$ as given by Theorem 1.

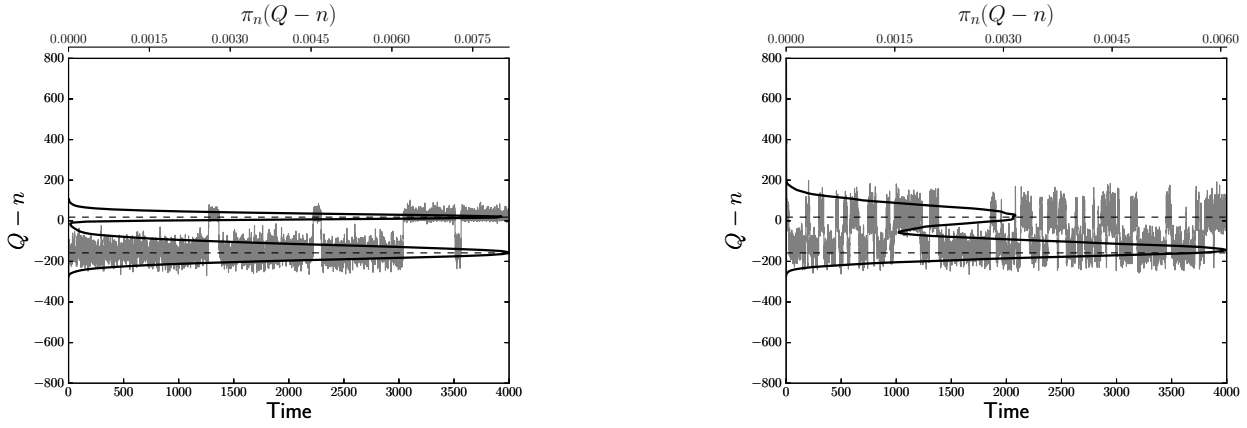
For high α ($\alpha = 0.5$, Figure 7(b)), the general system queue admits again a behavior very similar to the proposed Markovian approximation. Again, the general system also admits a mode exactly around $\lfloor \overline{q}_n \rfloor$ (as predicted by the theory for the Markovian system).

For the critical value of α ($\alpha = 1/3$, Figures 7(a) and 8(b)), the general system queue admits again a behavior very similar to the proposed Markovian approximation. For low values of β (Figure 8(a)), both systems operate in the ED regime. As β increases (Figure 8(b)), both systems move into the QED regime, as the queue oscillates between the two equilibria.

Across values of α and β and across scales, this simulation highlights the usefulness of the Markovian system in capturing some of the key features and predicting some of the behavior of the general system.



(a) $\alpha = 1/3$, $\beta = 2.1$, $n = 1021$, $\bar{s}_p = 1.256$, $\lfloor \overline{q}_n \rfloor - n = 28$



(b) $\alpha = 1/3$, $\beta = 2.7$, $n = 1074$, $\bar{s}_p = 1.286$, $\lfloor \underline{q}_n \rfloor - n = -158$, $\lfloor \overline{q}_n \rfloor - n = 18$

Figure 8 Simulation for Markovian (left) and General (right) systems. We consider $\alpha = 1/3$. The bottom $x-$axis corresponds to the simulation time, while the top $x-$axis corresponds to probabilities. In the figure we observe both a sample path and $\pi_n(\cdot)$. The dashed lines correspond to the modes $\lfloor \underline{q}_n \rfloor$ and $\lfloor \overline{q}_n \rfloor$ as given by Theorem 1.

7. Conclusion

In the present paper, we have proposed a framework for studying how spatial frictions affect capacity planning. In particular, we propose a reduced-form Markovian system that captures spatial economies of scale, leading to a crisp characterization of the trade-offs at play in such environments.

We have established a mapping from load to types of regimes in heavy traffic. In particular, recalling Eq.(3), we have focused on regimes parametrized by α and β, where

$$\lim_{n\to\infty}(1-\rho_n)n^{\alpha}=\beta, \qquad \text{for some } \beta\in\mathbb{R}_+,\ \alpha\in(0,1).$$

Figure 9 summarizes some of the main findings. The ED regime emerges whenever $\alpha > 1/3$ and

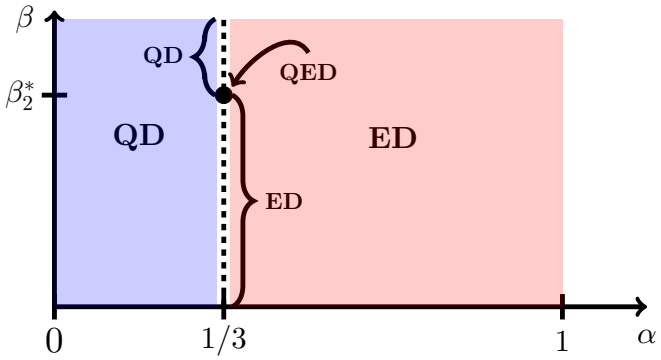


Figure 9 **Regimes for different values of α and β.**

the QD regime emerges whenever $\alpha < 1/3$. When $\alpha = 1/3$, the three regimes QD, ED and QED can emerge and the latter can only emerge for one critical value of β, which we label β_2^*. We have further demonstrated through simulations that the Markovian approximation provides a reliable guideline for the behavior of a general system.

This paper opens up various avenues of potential research, from both methodological and modeling perspectives. Analyzing the case when customers are impatient and might abandon the system if not served after some time is a natural extension. On the one hand, abandonment decreases the workload of the system as fewer customer have to be processed; on the other hand, it increases the system's workload as having fewer customers implies that *spatial economies of scale* become less advantageous. The important question in this case is whether, in order to achieve QED performance, abandonment necessitates just a change in β or more fundamental change in α. Another interesting extension is to study how the results in this study can be generalized to cases where origin-destination demand patterns generate imbalances in the system. In this case, the additional workload stemming from pickups might be even larger. How would this impact capacity sizing? An additional important practical question is to consider time-varying demand patterns that might require alternatives to steady-state analysis.

From a methodological perspective, an interesting extension would be to establish some of form of convergence of the processes in the general system to those in the Markovian approximation. More generally, there is potential to generalize the main result of this paper to any near optimal dispatching protocol by directly studying the spatial system. A simple back-of-the-envelope calculation serves to enlighten the latter claim. From Bertsimas and van Ryzin (1993) we can deduce that the expected number of customers in the system in steady state is bound below by

$$\frac{n}{2} - n \cdot (1 - \rho_n) + C \cdot \frac{\rho_n^2}{(1 - \rho_n)^2}.$$

The second term in this expression represents the number of idle server in the system, $n - n\rho_n$; while the third term maps to the number of customers waiting or being picked up. These two terms are opposing forces that push the system to have less and more customers, respectively. Using the heavy traffic scaling from Eq. (3), we can deduce that the second term scales as $\Theta(n^{1-\alpha})$, while the third term as $\Theta(n^{2\alpha})$. Observe that these scalings relate to those of the equilibria in Theorem 1. Intuitively, quality and efficiency should balance when the two opposing forces balance each other. This occurs when $1 - \alpha$ equals 2α or, equivalently, when α equals $1/3$, as our results prescribe. Note that this derivation does not rely on a specific dispatching protocol, but only on one that is optimal or "near optimal" compared to the lower bound. Deriving the $2/3$-scaling result at a full level of generality is an exciting direction for future work.

References

Afèche P, Liu Z, Maglaras C (2018) Ride-hailing networks with strategic drivers: The impact of platform control capabilities on performance. *Working paper, Columbia University* .

Aksin Z, Armony M, Mehrotra V (2007) The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management* 16(6):665–688.

Allen AO (2014) *Probability, statistics, and queueing theory* (Academic Press).

Atar R (2012) A diffusion regime with nondegenerate slowdown. *Operations Research* 60(2):490–500.

Banerjee S, Kanoria Y, Qian P (2018) State dependent control of closed queueing networks. *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, 2–4 (ACM).

Banerjee S, Riquelme C, Johari R (2015) Pricing in ride-share platforms: A queueing-theoretic approach. Working paper, Stanford University.

Bassamboo A, Randhawa RS, Zeevi A (2010) Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science* 56(10):1668–1686.

Bertsimas DJ, van Ryzin G (1991) A stochastic and dynamic vehicle routing problem in the euclidean plane. *Operations Research* 39(4):601–615.

Bertsimas DJ, van Ryzin G (1993) Stochastic and dynamic vehicle routing in the euclidean plane with multiple capacitated vehicles. *Operations Research* 41(1):60–76.

Besbes O, Castro F, Lobel I (2018) Surge pricing and its spatial supply response. Working paper, Columbia University.

Bimpikis K, Candogan O, Saban D (2016) Spatial pricing in ride-sharing networks. *Operations Research (forthcoming)* .

Braverman A, Dai JG, Liu X, Ying L (2016) Empty-car routing in ridesharing systems. Working paper, Northwestern University.

Castillo JC, Knoepfle D, Weyl G (2017) Surge pricing solves the wild goose chase. *Proceedings of the 2017 ACM Conference on Economics and Computation*, 241–242 (ACM).

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

Dong J, Feldman P, Yom-Tov GB (2015) Service systems with slowdowns: Potential failures and proposed solutions. *Operations Research* 63(2):305–324.

Feng G, Kong G, Wang Z (2017) We are on the way: Analysis of on-demand ride-hailing systems. Working paper, University of Minnesota.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* 5(2):79–141.

Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing & Service Operations Management* 4(3):208–227.

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Operations research* 29(3):567–588.

Mandelbaum A, Pats G (1995) State-dependent queues: approximations and applications. *Stochastic networks* 71:239–282.

Mandelbaum A, Pats G, et al. (1998) State-dependent stochastic networks. part i. approximations and applications with continuous diffusion limits. *The Annals of Applied Probability* 8(2):569–646.

Ozkan E, Ward AR (2016) Dynamic matching for real-time ridesharing. Working paper, University of Chicago.

Powell SG, Schultz KL (2004) Throughput in serial lines with state-dependent behavior. *Management Science* 50(8):1095–1105.

Reed J (2009) The G/GI/N queue in the halfin–whitt regime. *The Annals of Applied Probability* 19(6):2211–2269.

Shelbey S (1975) CRC standard mathematical tables .

Ward AR (2012) Asymptotic analysis of queueing systems with reneging: A survey of results for fifo, single class models. *Surveys in Operations Research and Management Science* 17(1):1–14.

Whitt W (2004) Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* 50(10):1449–1461.

Whitt W (2007) What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics (NRL)* 54(5):476–484.

Online Appendix for: Spatial Capacity Planning

A. Proofs for Section 3.2

Proof of Lemma 1. Let x_0 be in the interior of $\mathcal{C}$, a bounded subset of $\mathbb{R}^2$ with area denoted by $|\mathcal{C}|$. It is enough to prove that the following limit exists

$$\lim_{k\to\infty} \sqrt{k} \cdot \mathbf{E}\left[\min_{i=1,\ldots,k} \|X_i - x_0\|\right].$$

Let $Z_k \triangleq \min_{i=1,\ldots,k} \|X_i - x_0\|$. First, note that since x_0 is in the interior of the bounded region we can always find a ball $B(x_0,\epsilon)$ that is contain in $\mathcal{C}$ (below we take ϵ small enough). From this and the fact that the points X_i are drawn uniformly at random in $\mathcal{C}$, we have the following lower and upper bounds for any $i = 1,\ldots,k$

$$\frac{\pi\cdot(z\wedge\epsilon)^2}{|\mathcal{C}|} = \mathbf{P}[\|X_i - x_0\| \le z\wedge\epsilon] \le \mathbf{P}[\|X_i - x_0\| \le z] \le \frac{\pi\cdot z^2}{|\mathcal{C}|}.$$

Second, from these bounds and the fact that the points X_i are IID we deduce

$$\left(1 - \frac{\pi\cdot z^2}{|\mathcal{C}|}\right)^k \vee 0 \le \mathbf{P}[Z_k > z] \le \left(1 - \frac{\pi\cdot(z\wedge\epsilon)^2}{|\mathcal{C}|}\right)^k \vee 0.$$

This yields the following bound for $\mathbf{E}[Z_k]$

$$\int_0^{\sqrt{|\mathcal{C}|/\pi}} \left(1 - \frac{\pi\cdot z^2}{|\mathcal{C}|}\right)^k dz \le \mathbf{E}[Z_k] \le \int_0^{\epsilon} \left(1 - \frac{\pi\cdot z^2}{|\mathcal{C}|}\right)^k dz + \int_{\epsilon}^{R_{\mathcal{C}}} \left(1 - \frac{\pi\cdot \epsilon^2}{|\mathcal{C}|}\right)^k dz,$$

where $R_{\mathcal{C}} = \max_{x,y\in\mathcal{C}} \|x - y\|$ and we are assuming that $\epsilon < R_{\mathcal{C}}$. Note that

$$\lim_{k\to\infty} \sqrt{k}\cdot \int_{\epsilon}^{R_{\mathcal{C}}} \left(1 - \frac{\pi\cdot \epsilon^2}{|\mathcal{C}|}\right)^k dz = \lim_{k\to\infty} \sqrt{k}\cdot \left(1 - \frac{\pi\cdot \epsilon^2}{|\mathcal{C}|}\right)^k \cdot (R_{\mathcal{C}} - \epsilon) = 0,$$

where we are using that ϵ is small enough such that $\pi\cdot\epsilon^2/|\mathcal{C}| < 1$. Therefore, we have that

$$\sqrt{k}\cdot \int_0^{\epsilon} \left(1 - \frac{\pi\cdot z^2}{|\mathcal{C}|}\right)^k dz \le \sqrt{k}\cdot \mathbf{E}[Z_k] \le \sqrt{k}\cdot \int_0^{\epsilon} \left(1 - \frac{\pi\cdot z^2}{|\mathcal{C}|}\right)^k dz + \sqrt{k}\cdot \int_{\epsilon}^{R_{\mathcal{C}}} \left(1 - \frac{\pi\cdot \epsilon^2}{|\mathcal{C}|}\right)^k dz,$$

where the last term on the RHS above converges to zero. To complete the proof note that

$$\lim_{k\to\infty} \sqrt{k}\cdot \int_0^{\epsilon} \left(1 - \frac{\pi\cdot z^2}{|\mathcal{C}|}\right)^k dz = \sqrt{\frac{|\mathcal{C}|}{\pi}} \cdot \lim_{k\to\infty} \sqrt{k}\cdot \int_0^{\epsilon\cdot\sqrt{\pi/|\mathcal{C}|}} \left(1 - z^2\right)^k dz \approx 0.886 \cdot \sqrt{\frac{|\mathcal{C}|}{\pi}},$$

where in the last step we use that for any $0 < \delta < 1$ the limit as $k \uparrow \infty$ of $\sqrt{k}\int_0^{\delta}(1-z^2)^k dz$ is approximately 0.886. $\square$

B. Proofs for Section 4

Proof of Theorem 1. We make use of Proposition B-1 which we state and prove after the proof of this theorem. We prove each statements in the theorem.

(i) First we show that $\overline{q}_n$ as given in the statement is always an stable equilibrium. We have that $\overline{q}_n = n + z_n^2$ with $z_n = \rho_n/(1-\rho_n)$. Any equilibrium solves $f_n(q) = 0$, thus we just need to verify that

$$1 + \frac{1}{z_n} = \frac{n}{\lambda_n \bar{s}} = \frac{1}{\rho_n},$$

which is clearly satisfied. To verify stability we proceed using the Lyapunov method. Let $V(q) = |q - \overline{q}_n|$, then $\dot{V}(q) = \operatorname{sgn}(q - \overline{q}_n) \cdot f_n(q)$. We need to verify that $\dot{V}(q) < 0$ for $q \neq \overline{q}_n$ (for n large enough). By Proposition B-1 part *(i)*, if $q \in (\overline{q}_n, \overline{q}_n + \delta]$ we have that $\dot{V}(q) = f_n(q) < 0$, and if $\in [\overline{q}_n - \delta, \overline{q}_n)$ $\dot{V}(q) = -f_n(q) < 0$ for $\delta > 0$ small enough. Hence, $\overline{q}_n$ is a locally asymptotically stable equilibrium.

If $\alpha > 1/3$ or if $\alpha = 1/3$ and $\beta < \beta_1^*$ by Proposition B-1 we have that $f_n(q) > 0$ for all $q \in [0, \overline{q}_n)$. Therefore the same Lyapunov analysis as before leads to the conclusion that $\overline{q}_n$ is a globally asymptotically stable equilibrium.

(ii) Both equilibria $\underline{q}_n$ and $\tilde{q}_n$ can be found by equating $g_{1,n}(q)$ and $g_{2,n}(q)$. This turns out to be equivalent to solving the equation

$$(n - q) + \frac{n \cdot \rho_n}{\sqrt{n - q}} = n \cdot (1 - \rho_n). \tag{B-1}$$

For the current values of α and β, Proposition B-1 part *(iii)*, we know the latter equation has two solutions: $\tilde{q}_n$ and $\underline{q}_n$. Let's start with $\tilde{q}_n$. From Proposition B-1 we know that in a vicinity to the left of $\tilde{q}_n$ we have $f_n(q) < 0$, that is, in a vicinity to the left of $\tilde{q}_n$ we have $d\tilde{Q}_n(t)/dt < 0$ and, therefore, the systems moves away from $\tilde{q}_n$. Similarly, in a vicinity to the right of $\tilde{q}_n$ we have $f_n(q) > 0$ and, therefore, the system moves away from $\tilde{q}_n$. This shows that this equilibrium is unstable.

For $\underline{q}_n$ we can use the same Lyapunov analysis as before, together with Proposition B-1, to show that it is a locally asymptotically stable equilibrium.

To conclude we need to provide a closed form characterization the two equilibria. We transform the equation that defines them, Eq. (B-1), in to a cubic equation. Consider the change of variables $w = \sqrt{n - q}$, then the equation becomes

$$w^3 - n \cdot (1 - \rho_n) \cdot w + n \cdot \rho_n = 0. \tag{B-2}$$

The solution to this equation can be found in Shelbey (1975). When the term $-4n^3 \cdot (1-\rho_n)^3 + 27n^2 \cdot \rho_n^2$ is non-positive the three possible solutions to (B-2) are real and given by

$$w_i = 2\sqrt{\frac{n \cdot (1 - \rho_n)}{3}} \cdot \cos\Big(\frac{1}{3} \arccos\Big(-\sqrt{\frac{27\rho_n^2}{4n \cdot (1-\rho_n)^3}}\Big) - \frac{2\pi i}{3}\Big), \quad i = 0, 1, 2.$$

In order to verify that $-4n^3 \cdot (1-\rho_n)^3 + 27n^2 \cdot \rho_n^2 \leq 0$, note that this is equivalent to $27\rho_n^2 \leq 4n^{1-3\alpha} \cdot (n^\alpha(1-\rho_n))^3$. For large n, this last inequality holds for $\alpha < 1/3$. The same is true for $\alpha = 1/3$ and $\beta > \beta_1^*$. Therefore, the solutions w_k are all real. Furthermore, it is possible to verify that they are ordered, $w_0 \geq w_1 \geq w_2$, and that w_2 satisfies

$$w_2 = -2\sqrt{\frac{n\cdot(1-\rho_n)}{3}} \cdot \cos\Big(\frac{1}{3}\arccos\Big(\sqrt{\frac{27\rho_n^2}{4n\cdot(1-\rho_n)^3}}\Big)\Big) < 0,$$

and $w_1 \geq 0$ for large n. Since we are using the change of variables $w = \sqrt{n-q}$, we can disregard w_2 as a solution and take w_0 and w_1 to compute the solutions of our original equation. Because $\underline{q}_n \leq \tilde{q}_n$ we obtain that $\underline{q}_n = n - w_0^2$ and $\tilde{q}_n = n - w_1^2$.

□

PROPOSITION B-1. *Suppose* $\lim_{n\to\infty}(1-\rho_n)n^\alpha = \beta$ *and that* $\rho_n \uparrow 1$. *Let* $\beta_1^* = 3/4^{1/3}$ *then*

(i) there exists n_0 *such that for all* $n \geq n_0$ *there exists* $\overline{q}_n > n$ *for which*

$$f_n(q)\begin{cases} = 0 & \text{if } q = \overline{q}_n \\ < 0 & \text{if } q > \overline{q}_n \\ > 0 & \text{if } q \in [n, \overline{q}_n). \end{cases}$$

(ii) if $\alpha > 1/3$, *or if* $\alpha = 1/3$ *and* $\beta < \beta_1^*$, *there exists* n_0 *such that for all* $n \geq n_0$ *we have* $f_n(q) > 0$ *for all* $q \in [0, \overline{q}_n)$.

(iii) if $\alpha < 1/3$, *or if* $\alpha = 1/3$ *and* $\beta > \beta_1^*$ *then there exists* n_0 *such that for all* $n \geq n_0$ *there exist* $\underline{q}_n$ *and* $\tilde{q}_n$ *with* $0 \leq \underline{q}_n < n - (\frac{n\cdot\rho_n}{2})^{2/3} < \tilde{q}_n < n-1$ *such that*

$$f_n(q)\begin{cases} = 0 & \text{if } q \in \{\underline{q}_n, \tilde{q}_n\} \\ < 0 & \text{if } q \in (\underline{q}_n, \tilde{q}_n) \\ > 0 & \text{if } q \in [0, \underline{q}_n) \cup (\tilde{q}_n, \overline{q}_n). \end{cases}$$

Proof of Proposition B-1. First note that from the definition of f_n we have

$$f_n(q) = \lambda_n - \frac{1}{\frac{\bar{s}}{\sqrt{|q-n|\vee 1}} + \bar{s}} \cdot \min(n, q). \tag{B-3}$$

Next prove each part of the statement separately.

(i) Consider $q \geq n+1$ then $f_n(q) = 0$ if and only if

$$\Big(1 + \frac{1}{\sqrt{q-n}}\Big) = \frac{1}{\rho_n}.$$

The left hand side is a decreasing function of q with maximum value equal to 2 for $q \geq n+1$. Also, since $\rho_n < 1$ we have that $1/\rho_n > 1$. If n is large enough so that $1/\rho_n < 2$, we can always find a solution $\overline{q}_n > n$ such that $f_n(\overline{q}_n) = 0$. Moreover, $f_n(\overline{q}_n) < 0$ for $q > \overline{q}_n$, and $f_n(\overline{q}_n) > 0$ for $q \in [0, \overline{q}_n)$.

(ii) First suppose that $q \in [n, \overline{q}_n)$, from what we did in the proof of *(i)* we can conclude that $f_n(q) > 0$ for n large enough. For $q \in [n-1, n)$, $f_n(q) > 0$ if and only if $2 > q/(n\rho_n)$. Since $\rho_n \uparrow 1$ and q is at most n this last inequality holds for all n large enough.

Next, suppose that $q < n-1$. Note that $f_n(q) > 0$ if and only if

$$\left(1 + \frac{1}{\sqrt{n-q}}\right) > \frac{q}{n\rho_n}.$$

We can rewrite the previous equation in the following equivalent form

$$\underbrace{x_n + \frac{n \cdot \rho_n}{\sqrt{x_n}}}_{g_n(x_n)} > n \cdot (1 - \rho_n),$$

where $x_n = n - q$. Hence, $f_n(q) > 0$ if and only if $g_n(x_n) > n \cdot (1-\rho)$. Note that

$$\frac{dg_n(x)}{dx} = 1 - \frac{n \cdot \rho}{2x^{3/2}}, \quad \text{and} \quad \frac{d^2 g_n(x)}{dx^2} = \frac{3n \cdot \rho}{4x^{5/2}}.$$

Hence, $g_n(x)$ is a convex function with minimum at $x_n^* = (\frac{n \cdot \rho_n}{2})^{2/3}$. Thus, whenever $g_n(x_n^*) > n \cdot (1-\rho_n)$ we have that $f_n(q) > 0$. Observe that

$$g_n(x_n^*) > n \cdot (1-\rho_n) \Leftrightarrow (n \cdot \rho_n)^{2/3} \underbrace{(\frac{1}{2^{2/3}} + 2^{1/3})}_{\beta_1^*} > n \cdot (1-\rho_n) \Leftrightarrow \rho_n^{2/3} \beta_1^* > n^{1/3-\alpha} \cdot (1-\rho_n) n^{\alpha}.$$

If $\alpha > 1/3$ then, because $(1-\rho_n) \cdot n^{1/3} \to \beta$, the last inequality above holds for all n sufficiently large. If $\alpha = 1/3$ the last inequality above becomes $\rho_n^{2/3} \cdot \beta_1^* > (1-\rho_n) \cdot n^{1/3}$, and if $\beta < \beta_1^*$, since $(1-\rho_n) \cdot n^{1/3} \to \beta$ and $\rho_n \uparrow 1$, we would have $g_n(x_n^*) > n \cdot (1-\rho_n)$ for all n sufficiently large. Therefore in both cases we have that $f_n(q) > 0$ for all $q < n-1$.

(iii) Similarly, we can argue that if $\alpha < 1/3$, or if $\alpha = 1/3$ and $\beta > \beta_1^*$ then $g_n(x_n^*) < n \cdot (1-\rho_n)$ for n sufficiently large. When $g_n(x_n^*) < n \cdot (1-\rho_n)$ the function $g_n(x)$ (recall this is a convex function) crosses $n \cdot (1-\rho_n)$ at two points: $\underline{x}_{1,n}$ and $\overline{x}_{1,n}$, with $1 < \underline{x}_{1,n} < x_n^* < \overline{x}_{1,n} \leq n$. Defining $\underline{q}_n = n - \overline{x}_{1,n}$ and $\tilde{q}_n = n - \underline{x}_{1,n}$ we conclude the result.

□

C. Proofs for Section 5

Proof of Proposition 1. We make use of Eq. (9) and Proposition B-1.

(i) Note that from Proposition B-1 part *(i)* we have that $f_n(k) \geq 0$ for all $k \in [n, \overline{q}_n]$, since $\lfloor \overline{q}_n \rfloor \leq \overline{q}_n$ from Eq. (9) we deduce that $\pi_n(k)$ is increasing for all $k \in [n, \lfloor \overline{q}_n \rfloor] \cap \mathbb{N}$. Moreover, because $f_n(k) < 0$ for $k > \overline{q}_n$ and $\overline{q}_n < \lfloor \overline{q}_n \rfloor + 1$ from Eq. (9) we have that $\pi_n(k)$ decreases for all $k \in (\lfloor \overline{q}_n \rfloor, \infty) \cap \mathbb{N}$. Finally, using a similar argument and Proposition B-1 part *(ii)*, we deduce that $\pi_n(k)$ is increasing for all $k \in [0, n] \cap \mathbb{N}$.

(ii) Note that from Proposition B-1 part *(iii)* we have that $f_n(k) \geq 0$ for all $k \in [0, \underline{q}_n]$, $f_n(k) < 0$ for all $k \in (\underline{q}_n, \tilde{q}_n)$, and $f_n(k) \geq 0$ for all $k \in [\tilde{q}_n, \overline{q}_n)$. Eq. (9) then implies that $\pi_n(k)$ increases for $k \in [0, \lfloor \underline{q}_n \rfloor] \cap \mathbb{N}$, it decreases for $k \in (\lfloor \underline{q}_n \rfloor, \lfloor \tilde{q}_n \rfloor] \cap \mathbb{N}$, and it increases for $k \in (\lfloor \tilde{q}_n \rfloor, \lfloor \overline{q}_n \rfloor) \cap \mathbb{N}$.

□

Proof of Theorem 2. This result relies on Proposition 2 which is stated in the main text in the Proof sketch of Theorem 2 discussion. We provide a proof for Proposition 2 after the present proof.

We prove each statement in the theorem separately.

(i) We analyze different cases. First we consider $\alpha \in (1/3, 1)$. In this case from Proposition 1 part *(ii)* we now that $\pi_n(k) \leq \pi_n(n)$ for all k, for all n large enough. Moreover, from Proposition 2 part *(i)* we have that for $\epsilon \in (0, 1/\beta)$ for all n large enough the following inequality holds

$$\frac{\pi_n(n)}{\pi_n(\lfloor \overline{q}_n \rfloor)} \leq \exp\Big(- n^{\alpha}(\frac{1}{\beta} - \epsilon)\Big).$$

Therefore,

$$\mathbf{P}[Q_n(\infty) < n] = \sum_{k=0}^{n-1} \pi_n(k) \leq n \cdot \pi_n(n) = n \cdot \frac{\pi_n(n)}{\pi_n(\lfloor \overline{q}_n \rfloor)} \cdot \pi_n(\lfloor \overline{q}_n \rfloor) \leq n \cdot \exp\Big(- n^{\alpha}(\frac{1}{\beta} - \epsilon)\Big) \cdot \pi_n(\lfloor \overline{q}_n \rfloor) \to 0.$$

Next, consider $\alpha = 1/3$ and $\beta < \beta_2^*$. Let $\pi_n(k|\beta)$ be the steady-state probability when λ_n is such that $(1 - \rho_n)n^{1/3} = \beta$. For notational clarity we use $\lambda_n(\beta), \overline{q}_n(\beta)$ and $\underline{q}_n(\beta)$ instead of $\lambda_n, \overline{q}_n$ and $\underline{q}_n$. It is possible to show that for $\beta < \beta'$ and n large enough we must have that

$$\frac{\pi_n(k|\beta)}{\pi_n(\lfloor \overline{q}_n(\beta') \rfloor | \beta)} \leq \frac{\pi_n(k|\beta')}{\pi_n(\lfloor \overline{q}_n(\beta') \rfloor | \beta')}, \quad \forall k \leq n - 1. \tag{C-1}$$

Before we show Eq. (C-1), we will use to conclude this part of the proof. Fix $\beta < \beta_2^*$ then we can find $\beta' \in (\max\{\beta_1^*, \beta\}, \beta_2^*)$ for which Eq. (C-1) holds and, therefore, from Proposition 2 we can take $\epsilon \in (0, g(\beta'))$ such that for n large enough we have

$$\mathbf{P}[Q_n(\infty) < n] = \sum_{k=0}^{n-1} \pi_n(k|\beta)$$

$$
\begin{aligned}
&\leq \sum_{k=0}^{n-1} \frac{\pi_n(k|\beta)}{\pi_n(\lfloor \overline{q}_n(\beta')\rfloor|\beta)} \\
&\leq \sum_{k=0}^{n-1} \frac{\pi_n(k|\beta')}{\pi_n(\lfloor \overline{q}_n(\beta')\rfloor|\beta')} \\
&\leq n \cdot \frac{\pi_n(\lfloor \underline{q}_n(\beta')\rfloor|\beta')}{\pi_n(\lfloor \overline{q}_n(\beta')\rfloor|\beta')} \\
&= n \cdot \exp\Big(- n^{1/3}(g(\beta') - \epsilon)\Big) \to 0,
\end{aligned}
$$

as $n \to \infty$. Next, we verify Eq. (C-1). Note that for $k < \lfloor \overline{q}_n(\beta')\rfloor$

$$
\frac{\pi_n(\lfloor \overline{q}_n(\beta')\rfloor|\beta)}{\pi_n(k|\beta)} = \prod_{m=k+1}^{\lfloor \overline{q}_n(\beta')\rfloor} \frac{\lambda_n(\beta)\bar{s}}{\min\{m,n\}} \cdot \Big(1 + \frac{1}{\sqrt{|n-m| \vee 1}}\Big)
$$

and

$$
\frac{\pi_n(\lfloor \overline{q}_n(\beta')\rfloor|\beta')}{\pi_n(k|\beta')} = \prod_{m=k+1}^{\lfloor \overline{q}_n(\beta')\rfloor} \frac{\lambda_n(\beta')\bar{s}}{\min\{m,n\}} \cdot \Big(1 + \frac{1}{\sqrt{|n-m| \vee 1}}\Big).
$$

Hence, Eq. (C-1) is satisfied if and only if

$$
\lambda_n(\beta')^{\lfloor \overline{q}_n(\beta')\rfloor - k} \leq \lambda_n(\beta)^{\lfloor \overline{q}_n(\beta')\rfloor - k} \Leftrightarrow \lambda_n(\beta') \leq \lambda_n(\beta) \Leftrightarrow n^{1/3}\Big(1 - \frac{\lambda_n(\beta')\bar{s}}{n}\Big) \geq n^{1/3}\Big(1 - \frac{\lambda_n(\beta)\bar{s}}{n}\Big),
$$

since both expression in the last inequality above converge to β' and β (respectively) and $\beta' > \beta$, we can always find n large enough so that the inequality is true. This shows Eq. (C-1).

(ii) Consider first $\alpha \in (0, 1/3)$. Write

$$
\mathbf{P}[Q_n(\infty) \geq n] = \sum_{k=n}^{\lfloor \overline{q}_n\rfloor} \pi_n(k) + \sum_{k=\lfloor \overline{q}_n\rfloor+1}^{\infty} \pi_n(k). \tag{C-2}
$$

We next bound both terms and then show they converge to zero. The first term in Eq. (C-2) is bounded above

$$
\sum_{k=n}^{\lfloor \overline{q}_n\rfloor} \pi_n(k) \leq \pi_n(\lfloor \overline{q}_n\rfloor) \cdot (\lfloor \overline{q}_n\rfloor - n + 1) = \pi_n(\lfloor \overline{q}_n\rfloor) \cdot (\lfloor \overline{q}_n\rfloor - \overline{q}_n + \overline{q}_n - n + 1) \leq \pi_n(\lfloor \overline{q}_n\rfloor) \cdot (\frac{\rho_n^2}{(1-\rho_n)^2} + 1),
$$

where in the last inequality we used that $\lfloor \overline{q}_n\rfloor \leq \overline{q}_n$, and Theorem 1 part *(i)* to obtain an expression for $\overline{q}_n$. In order to bound the second term in Eq. (C-2), first note that

$$
\frac{\pi_n(k)}{\pi_n(\lfloor \overline{q}_n\rfloor)} = \prod_{\ell=\lfloor \overline{q}_n\rfloor+1}^{k} \rho_n \cdot \Big(1 + \frac{1}{\sqrt{\ell - n}}\Big), \quad \forall k > \lfloor \overline{q}_n\rfloor.
$$

Let

$$
a_n = \rho_n \cdot \Big(1 + \frac{1}{\sqrt{\lfloor \overline{q}_n\rfloor + 1 - n}}\Big),
$$

which satisfies $a_n < 1$ for all n. Indeed,

$$\rho_n \cdot \Big(1 + \frac{1}{\sqrt{\lfloor \overline{q}_n \rfloor + 1 - n}}\Big) < 1 \;\Leftrightarrow\; \frac{\rho_n^2}{(1-\rho_n)^2} < \lfloor \overline{q}_n \rfloor + 1 - n \;\Leftrightarrow\; \frac{\rho_n^2}{(1-\rho_n)^2} < 1 - (\overline{q}_n - \lfloor \overline{q}_n \rfloor) + \overline{q}_n - n,$$

from Theorem 1 part *(i)* the last inequality becomes $(\overline{q}_n - \lfloor \overline{q}_n \rfloor) < 1$, which is always is true. then

$$\begin{aligned}
\sum_{k=\lfloor \overline{q}_n \rfloor + 1}^{\infty} \pi_n(k) &= \pi_n(\lfloor \overline{q}_n \rfloor) \cdot \sum_{k=\lfloor \overline{q}_n \rfloor + 1}^{\infty} \prod_{\ell=\lfloor \overline{q}_n \rfloor + 1}^{k} \rho_n \cdot \Big(1 + \frac{1}{\sqrt{\ell - n}}\Big) \\
&\overset{(a)}{\leq} \pi_n(\lfloor \overline{q}_n \rfloor) \cdot \sum_{k=\lfloor \overline{q}_n \rfloor + 1}^{\infty} \prod_{\ell=\lfloor \overline{q}_n \rfloor + 1}^{k} a_n \\
&= \pi_n(\lfloor \overline{q}_n \rfloor) \cdot \sum_{k=\lfloor \overline{q}_n \rfloor + 1}^{\infty} a_n^{k - \lfloor \overline{q}_n \rfloor} \\
&= \pi_n(\lfloor \overline{q}_n \rfloor) \cdot a_n^{-\lfloor \overline{q}_n \rfloor} \cdot \frac{a_n^{\lfloor \overline{q}_n \rfloor + 1}}{1 - a_n} \\
&< \pi_n(\lfloor \overline{q}_n \rfloor) \cdot \frac{1}{1 - a_n},
\end{aligned}$$

where (a) holds because the term $1 + 1/\sqrt{\ell - n}$ is decreasing in ℓ. Putting the upper bounds for Eq. (C-2) together yields

$$\mathbf{P}[Q_n(\infty) \geq n] \leq \pi_n(\lfloor \overline{q}_n \rfloor) \cdot \Big(\frac{\rho_n^2}{(1-\rho_n)^2} + 1 + \frac{1}{1 - a_n}\Big).$$

Observe that the term in brackets is $O(n^\gamma)$ for some $\gamma > 0$. Also, we can always consider $\epsilon > 0$ such that $\beta^2/2 > \epsilon$ and then we can use Theorem 2 to find n_0 such that for all $n \geq n_0$

$$\pi_n(\lfloor \overline{q}_n \rfloor) \leq \pi_n(\lfloor \underline{q}_n \rfloor) \cdot \exp\Big(-(\frac{\beta^2}{2} - \epsilon) \cdot n^{1-2\alpha}\Big).$$

Since $\pi_n(\lfloor \underline{q}_n \rfloor) \leq 1$ and $1 - 2\alpha > 0$ we conclude that

$$\mathbf{P}[Q_n(\infty) \geq n] \leq \exp\Big(-(\frac{\beta^2}{2} - \epsilon) \cdot n^{1-2\alpha}\Big) \cdot O(n^\gamma) \longrightarrow 0, \quad \text{as } n \to \infty.$$

Note that for $\alpha = 1/3$ and $\beta > \beta_2^*$ the same argument holds, we only need to chose $\epsilon > 0$ such that $|g(\beta)| > \epsilon$. This is always possible since for $\beta > \beta_2^*$ Theorem 2 establishes that $g(\beta) < 0$. This concludes the proof.

□

Proof of Proposition 2. We prove each part separately. First, note that

$$\frac{\pi_n(\lfloor \overline{q}_n \rfloor)}{\pi_n(m)} = \prod_{k=m+1}^{\lfloor \overline{q}_n \rfloor} \frac{\lambda_n}{\mu_n(k) \cdot \min\{k, n\}} = \prod_{k=m+1}^{\lfloor \overline{q}_n \rfloor} \frac{\lambda_n \bar{s}}{\min\{k, n\}} \cdot \Big(1 + \frac{1}{\sqrt{|n-k| \vee 1}}\Big),$$

for any $m < \lfloor \overline{q}_n \rfloor$. Then

$$\log\Big(\frac{\pi_n(\lfloor \overline{q}_n \rfloor)}{\pi_n(m)}\Big) = (\lfloor \overline{q}_n \rfloor - m)\log(\rho_n) + \sum_{k=m+1}^{\lfloor \overline{q}_n \rfloor} \log\Big[\frac{n}{\min\{k, n\}} \cdot \Big(1 + \frac{1}{\sqrt{|n-k| \vee 1}}\Big)\Big] \tag{C-3}$$

(i) For $m=n$: Let $x_n=\lfloor \overline{q}_n\rfloor - n$, then equation (C-3) becomes

$$\begin{aligned}\log\Big(\frac{\pi_n(\lfloor \overline{q}_n\rfloor)}{\pi_n(n)}\Big) &= x_n\log(\rho_n)+\sum_{k=n+1}^{\lfloor \overline{q}_n\rfloor}\log\Big[1+\frac{1}{\sqrt{k-n}}\Big]\\ &= x_n\cdot\log(\rho_n)+\int_1^{x_n}\log\Big[1+\frac{1}{\sqrt{x}}\Big]dx+O(1)\\ &= x_n\cdot\log(\rho_n)+\Big[\sqrt{x}+x\log(1+\frac{1}{\sqrt{x}})-\log(1+\sqrt{x})\Big]\Big|_1^{x_n}+O(1)\\ &= \sqrt{x_n}-\log(1+\sqrt{x_n})+x_n\cdot\Big(\log(\rho_n)+\log\Big(1+\frac{1}{\sqrt{x_n}}\Big)\Big)+O(1).\end{aligned}$$

In the expression above we can use that $x_n\to\infty$, $x_n=\lfloor \overline{q}_n\rfloor-\overline{q}_n+\frac{\rho^2}{(1-\rho)^2}$ and Taylor expansions to conclude that

$$\sqrt{x_n}=\frac{\rho_n}{(1-\rho_n)}+o(1),\quad \text{and that}\quad x_n\cdot\Big(\log(\rho_n)+\log\Big(1+\frac{1}{\sqrt{x_n}}\Big)\Big)=-\frac{\rho^2}{(1-\rho)^2}+\sqrt{x_n}+O(1)=O(1).$$

Since $(1-\rho_n)n^\alpha\to\beta$ we have

$$\lim_{n\to\infty}\frac{1}{n^\alpha}\log\Big(\frac{\pi_n(\lfloor \overline{q}_n\rfloor)}{\pi_n(n)}\Big)=\lim_{n\to\infty}\frac{1}{n^\alpha}\frac{\rho_n}{(1-\rho_n)}=\frac{1}{\beta}.$$

(ii) We assume that $\alpha<1/3$ and we take $m=\lfloor \underline{q}_n\rfloor$. Note that since $\alpha<1/3$ we have

$$\frac{27\rho_n^2}{4n\cdot(1-\rho_n)^3}\to 0,\quad as\quad n\to\infty.$$

Then, we can use Theorem 1 and do a Taylor expansion to deduce that

$$r_{0,n}(\rho_n)=1-\frac{2}{3\sqrt{3}}\sqrt{x}-\frac{2}{27}x-\frac{5}{81\sqrt{3}}x^{3/2}+O(x^2)\Big|_{x=\frac{27\rho_n^2}{4n\cdot(1-\rho_n)^3}}.$$

Hence, since $\alpha<1/3$ we deduce that

$$n-\underline{q}_n=n\cdot(1-\rho_n)+O(n^{(1+\alpha)/2}). \tag{C-4}$$

In order to prove the result for this part of the proposition we need to analyze the term

$$\begin{aligned}\log\Big(\frac{\pi_n(\lfloor \overline{q}_n\rfloor)}{\pi_n(\lfloor \underline{q}_n\rfloor)}\Big) &= (\lfloor \overline{q}_n\rfloor-\lfloor \underline{q}_n\rfloor)\log(\rho_n)+\sum_{k=\lfloor \underline{q}_n\rfloor+1}^{n-1}\log\Big[\frac{n}{k}\cdot\Big(1+\frac{1}{\sqrt{n-k}}\Big)\Big]+\sum_{k=1}^{\lfloor \overline{q}_n\rfloor-n}\log\Big[1+\frac{1}{\sqrt{k}}\Big]+\log(2)\\ &= \underbrace{(n-\lfloor \underline{q}_n\rfloor)\log(\rho_n)}_{A}+\underbrace{\sum_{k=\lfloor \underline{q}_n\rfloor+1}^{n-1}\log\Big[\frac{n}{k}\cdot\Big(1+\frac{1}{\sqrt{n-k}}\Big)\Big]}_{B}\\ &\quad+\underbrace{(\lfloor \overline{q}_n\rfloor-n)\log(\rho_n)+\sum_{k=1}^{\lfloor \overline{q}_n\rfloor-n}\log\Big[1+\frac{1}{\sqrt{k}}\Big]}_{C}+\log(2).\end{aligned}$$

Let's look at each one of the terms A, B and C. For A, using Eq. (C-5), we have that

$$(n-\lfloor \underline{q}_n \rfloor)\log(\rho_n) = n\cdot(1-\rho_n)\log(\rho_n) + O(n^{(1-\alpha)/2}) = -n\cdot(1-\rho_n)^2 + O(n^{1-3\alpha}) + O(n^{(1-\alpha)/2}),$$

and because $\alpha < 1/3$, we have that $A/n^{1-2\alpha} \to -\beta^2$. So we only need to case analyze B and C. From the proof of part *(i)* we have

$$C = \frac{\rho_n}{(1-\rho_n)} + \log(1-\rho_n) + O(1) = o(n^{1-2\alpha}),$$

where the last equality comes from $\alpha < 1/3$. For B,

$$\begin{aligned}
B &= \int_{\lfloor \underline{q}_n \rfloor}^{n-1} \log\Big[\frac{n}{x}\cdot\Big(1+\frac{1}{\sqrt{n-x}}\Big)\Big]dx + o(n^{1-2\alpha}) \\
&= \Big[x\log(\frac{n}{x}) + x - \sqrt{n-x} - (n-x)\log(1+\frac{1}{\sqrt{n-x}}) + \log(1+\sqrt{n-x})\Big]\Big|_{\lfloor \underline{q}_n \rfloor}^{n-1} + o(n^{1-2\alpha}) \\
&= n - 1 - \Big[\lfloor \underline{q}_n \rfloor \log(\frac{n}{\lfloor \underline{q}_n \rfloor}) + \lfloor \underline{q}_n \rfloor - \sqrt{n-\lfloor \underline{q}_n \rfloor} - (n-\lfloor \underline{q}_n \rfloor)\log(1+\frac{1}{\sqrt{n-\lfloor \underline{q}_n \rfloor}}) \\
&\quad + \log(1+\sqrt{n-\lfloor \underline{q}_n \rfloor})\Big] + o(n^{1-2\alpha}) \\
&= n - \lfloor \underline{q}_n \rfloor \log(\frac{n}{\lfloor \underline{q}_n \rfloor}) - \lfloor \underline{q}_n \rfloor + o(n^{1-2\alpha}) \\
&= n - \lfloor \underline{q}_n \rfloor - \lfloor \underline{q}_n \rfloor \cdot \Big(\frac{(n-\lfloor \underline{q}_n \rfloor)}{\lfloor \underline{q}_n \rfloor} - \frac{(n-\lfloor \underline{q}_n \rfloor)^2}{2\lfloor \underline{q}_n \rfloor^2}\Big) + o(n^{1-2\alpha}) \\
&= \frac{(n-\lfloor \underline{q}_n \rfloor)^2}{2\lfloor \underline{q}_n \rfloor} + o(n^{1-2\alpha}),
\end{aligned}$$

using that $\alpha < 1/3$ it follows that this last expression, when scaled by $1/n^{1-2\alpha}$, converges to $\beta^2/2$. Therefore,

$$\frac{1}{n^{1-2\alpha}}\log\Big(\frac{\pi_n(\lfloor \overline{q}_n \rfloor)}{\pi_n(\lfloor \underline{q}_n \rfloor)}\Big) \to -\beta^2 + \frac{\beta^2}{2} + 0 = -\frac{\beta^2}{2}, \quad \text{as} \quad n \to \infty,$$

as required.

(iii) We assume that $\alpha = 1/3$ and we take $m = \lfloor \underline{q}_n \rfloor$. Note that

$$r_{0,n}(\rho_n) \to \frac{4}{3}\cdot\cos\Big(\frac{1}{3}\arccos\Big(-\sqrt{\Big(\frac{\beta_1^*}{\beta}\Big)^3}\Big)\Big)^2 \triangleq r(\beta), \quad as \quad n\to\infty. \tag{C-5}$$

Observe that since we are considering $\beta \geq \beta_1^*$ the $\arccos(\cdot)$ term is well defined and, therefore, so is $r(\beta)$. We need to analyze the following expression

$$\begin{aligned}
\log\Big(\frac{\pi_n(\lfloor \overline{q}_n \rfloor)}{\pi_n(\lfloor \underline{q}_n \rfloor)}\Big) &= \underbrace{(n-\lfloor \underline{q}_n \rfloor)\log(\rho_n)}_{A} + \underbrace{\sum_{k=\lfloor \underline{q}_n \rfloor+1}^{n-1} \log\Big[\frac{n}{k}\cdot\Big(1+\frac{1}{\sqrt{n-k}}\Big)\Big]}_{B} \\
&\quad + \underbrace{(\lfloor \overline{q}_n \rfloor - n)\log(\rho_n) + \sum_{k=1}^{\lfloor \overline{q}_n \rfloor - n} \log\Big[1+\frac{1}{\sqrt{k}}\Big]}_{C} + \log(2).
\end{aligned}$$

Let's look at each one of the terms A, B and C. For A, using Theorem 1 we have that

$$A = (n - \lfloor \underline{q}_n \rfloor) \log(\rho_n) = n \cdot (1-\rho_n) \cdot r_{0,n}(\rho_n) \log(\rho_n) + o(1) = -n(1-\rho_n)^2 r_{0,n}(\rho_n) + o(n^{1/3}).$$

Similarly to part *(ii)* above, for C we deduce

$$C = \frac{\rho_n}{(1-\rho_n)} + \log(1-\rho_n) + O(1) = \frac{\rho_n}{(1-\rho_n)} + o(n^{1/3}).$$

Finally, for B (similarly to part *(ii)* above)

$$\begin{aligned} B &= n - \left[\lfloor \underline{q}_n \rfloor \log(\frac{n}{\lfloor \underline{q}_n \rfloor}) + \lfloor \underline{q}_n \rfloor - \sqrt{n - \lfloor \underline{q}_n \rfloor} - (n - \lfloor \underline{q}_n \rfloor) \log(1 + \frac{1}{\sqrt{n - \lfloor \underline{q}_n \rfloor}}) + \log(1 + \sqrt{n - \lfloor \underline{q}_n \rfloor}) \right] + o(n^{1/3}) \\ &= n - \lfloor \underline{q}_n \rfloor - \lfloor \underline{q}_n \rfloor \log(\frac{n}{\lfloor \underline{q}_n \rfloor}) + 2\sqrt{n - \lfloor \underline{q}_n \rfloor} + o(n^{1/3}) \\ &= \frac{(n - \lfloor \underline{q}_n \rfloor)^2}{2\lfloor \underline{q}_n \rfloor} + 2\sqrt{n - \lfloor \underline{q}_n \rfloor} + o(n^{1/3}), \end{aligned}$$

and, therefore, using that $n^{1/3}(1-\rho_n) \to \beta$, $n - \lfloor \underline{q}_n \rfloor = n(1-\rho_n) r_{0,n}(\rho_n)$ and Eq. (C-5) we can compute the limit

$$\lim_{n\to\infty} \frac{B}{n^{1/3}} = \lim_{n\to\infty} \frac{1}{n^{1/3}} \cdot \frac{(n - \lfloor \underline{q}_n \rfloor)^2}{2\lfloor \underline{q}_n \rfloor} + 2\frac{\sqrt{n - \lfloor \underline{q}_n \rfloor}}{n^{1/3}} = \frac{\beta^2 r(\beta)^2}{2} + 2\sqrt{\beta r(\beta)},$$

where $r(\beta)$ is defined in Eq. (C-5). From this we can deduce that

$$\frac{1}{n^{1/3}} \log\left(\frac{\pi_n(\overline{q}_n)}{\pi_n(\underline{q}_n)}\right) \to -\beta^2 r(\beta) + \frac{\beta^2 r(\beta)^2}{2} + 2\sqrt{\beta r(\beta)} + \frac{1}{\beta} \triangleq g(\beta), \quad \text{as} \quad n \to \infty. \tag{C-6}$$

It is possible to verify that $g(\beta)$ satisfies $g(\beta_1^*) > 0$ and it is strictly decreasing for $\beta \geq \beta_1^*$, with $\lim_{\beta\to\infty} g(\beta) = -\infty$, see Figure 10. Therefore, there exists $\beta_2^* > \beta_1^*$ such that $g(\beta_2^*) = 0$. Thus we have verified that $g(\beta)$ is such that if $\beta_1^* < \beta < \beta_2^*$ then $g(\beta) > 0$, whereas if $\beta > \beta_2^*$ then $g(\beta) < 0$.

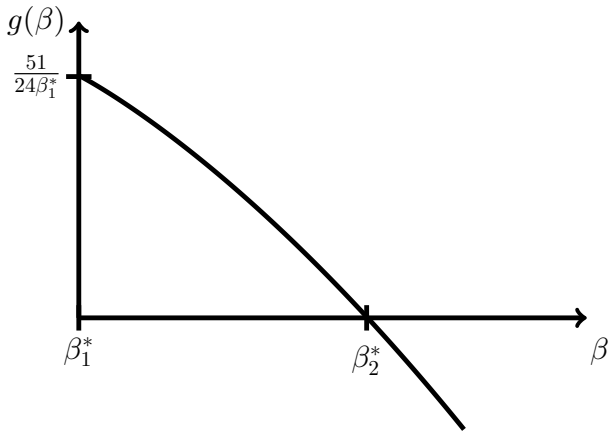


Figure 10 **Function $g(\beta)$ as defined in Eq. (C-6), $g(\beta)$ is strictly decreasing and it crosses zero at β_2^*.**

□

Proof of Theorem 3. We make use of the lemmata C-1 and C-2 which we first state and then prove after the proof of this theorem. We also make use of Proposition 3 which is stated in the main text and proven in this appendix.

In order to simplify notation let $p_n^+ = \mathbf{P}[Q_n(\infty) \geq n]$. Let $\beta = \beta_2^*$ and $\alpha = 1/3$ then from Lemma C-1 and Lemma C-2 there exists n_1 such that

$$\underbrace{\frac{\frac{1-\exp\left(-\frac{C^2\beta^3}{2}\right)}{\frac{C^2\beta^3}{2}}}{1+\frac{\exp\left(-\frac{C^2}{2}\left(1-\frac{1}{2(\beta\cdot r(\beta))^{3/2}}\right)\right)}{\frac{C^2}{2}\left(1-\frac{1}{2(\beta\cdot r(\beta))^{3/2}}\right)}}}_{A(C)}\cdot\frac{\pi_n(\lfloor\overline{q}_n\rfloor)}{\pi_n(\lfloor\underline{q}_n\rfloor)} \leq \frac{p_n^+}{(1-p_n^+)} \leq \underbrace{\frac{1+1\cdot\frac{\exp\left(-\frac{C^2\beta^3}{4}\right)}{\frac{C^2\beta^3}{4}}}{\frac{1-\exp\left(-C^2\cdot\left(1-\frac{1}{2(\beta\cdot r(\beta))^{3/2}}\right)\right)}{C^2\cdot\left(1-\frac{1}{2(\beta\cdot r(\beta))^{3/2}}\right)}}}_{B(C)}\cdot\frac{\pi_n(\lfloor\overline{q}_n\rfloor)}{\pi_n(\lfloor\underline{q}_n\rfloor)}, \quad \forall n \geq n_1.$$

Next, fix $\epsilon > 0$ then by Proposition 3 we have that there exists n_2 such that

$$\exp(-\epsilon + c) \leq \frac{\pi_n(\lfloor\overline{q}_n\rfloor)}{\pi_n(\lfloor\underline{q}_n\rfloor)} \leq \exp(\epsilon + c), \quad \forall n \geq n_2.$$

Therefore, for all $n \geq \max\{n_1, n_2\}$

$$A(C)\cdot\exp(-\epsilon + c) \leq \frac{p_n^+}{(1-p_n^+)} \leq B(C)\cdot\exp(\epsilon + c),$$

or, alternatively, (letting $\epsilon \to 0$)

$$\frac{A(C)}{e^{-c}+A(C)} \leq \liminf_{n\to\infty} p_n^+ \leq \limsup_{n\to\infty} p_n^+ \leq \frac{B(C)}{e^{-c}+B(C)}.$$

Now we want to find the tightest upper and lower bound. To do this it is enough to maximize the LHS and minimize the RHS above as a function of C. Since all the parameters are known ($\beta_2^* \approx 2.6030$ and $r(\beta_2^*) \approx 0.7192$) we can obtain numerical values,

$$\max_{C>0}\left\{\frac{A(C)}{e^{-c}+A(C)}\right\} \approx \frac{0.0524}{e^{-c}+0.0524}, \quad \text{and} \quad \min_{C>0}\left\{\frac{B(C)}{e^{-c}+B(C)}\right\} \approx \frac{1.3173}{e^{-c}+1.3173}.$$

So if we fix $p_H \in (0,1)$ then there exists $c^* \in \mathbb{R}$ such that

$$\frac{1.3173}{e^{-c^*}+1.3173} = p_H,$$

and c^* increases with p_H. Therefore if we let

$$p_L(p_H) = \frac{0.0524}{e^{-c^*}+0.0524},$$

we have that $p_L(p_H) \in (0,1)$ increases with p_H. In particular, $\lim_{p_H\to 1} p_L(p_H) = 1$ and $\lim_{p_H\to 0} p_L(p_H) = 0$, as desired.

□

LEMMA C-1. *Fix $\alpha \in (0,1/3)$ and $\beta > 0$, or $\alpha = 1/3$ and $\beta > \beta_1^*$. Suppose that $\lim_{n\to\infty} n^\alpha(1-\rho_n) = \beta$ and let $C > 0$ be a constant then*

$$2 \cdot \frac{1 - \exp\Big(-C^2 \cdot \Big(1 - \frac{\mathbf{1}_{\{\alpha=1/3\}}}{2(\beta \cdot r(\beta))^{3/2}}\Big)\Big)}{C^2 \cdot \Big(1 - \frac{\mathbf{1}_{\{\alpha=1/3\}}}{2(\beta \cdot r(\beta))^{3/2}}\Big)} \leq \liminf_{n\to\infty} \frac{1}{C\sqrt{n}} \cdot \frac{\mathbf{P}[Q_n(\infty) < n]}{\pi_n(\lfloor \underline{q}_n \rfloor)},$$

and

$$\limsup_{n\to\infty} \frac{1}{C\sqrt{n}} \cdot \frac{\mathbf{P}[Q_n(\infty) < n]}{\pi_n(\lfloor \underline{q}_n \rfloor)} \leq 2 \ + \ 2 \cdot \frac{\exp\Big(-\frac{C^2}{2}\Big(1 - \frac{\mathbf{1}_{\{\alpha=1/3\}}}{2(\beta \cdot r(\beta))^{3/2}}\Big)\Big)}{\frac{C^2}{2}\Big(1 - \frac{\mathbf{1}_{\{\alpha=1/3\}}}{2(\beta \cdot r(\beta))^{3/2}}\Big)},$$

where $r(\beta) = \lim_{n\to\infty} r_{0,n}(\rho_n)$.

LEMMA C-2. *Fix $\alpha \in (0,1)$ and $\beta > 0$. Suppose that $\lim_{n\to\infty} n^\alpha(1-\rho_n) = \beta$ and let $C > 0$ be a constant then*

$$2 \cdot \frac{1 - \exp\Big(-\frac{C^2\beta^3}{2}\Big)}{\frac{C^2\beta^3}{2}} \leq \liminf_{n\to\infty} \frac{1}{Cn^{\frac{3}{2}\alpha}} \cdot \frac{\mathbf{P}[Q_n(\infty) \geq n]}{\pi_n(\lfloor \overline{q}_n \rfloor)},$$

and

$$\limsup_{n\to\infty} \frac{1}{Cn^{\frac{3}{2}\alpha}} \cdot \frac{\mathbf{P}[Q_n(\infty) \geq n]}{\pi_n(\lfloor \overline{q}_n \rfloor)} \leq 2 \ + \ 2 \cdot \frac{\exp\Big(-\frac{C^2\beta^3}{4}\Big)}{\frac{C^2\beta^3}{4}}.$$

Proof of Lemma C-1. We start we the lower bound. Let $b_n = C\sqrt{n}$ and note that

$$\begin{aligned}
\frac{\mathbf{P}[Q_n(\infty) < n]}{\pi_n(\lfloor \underline{q}_n \rfloor) b_n} &= \frac{1}{b_n} \sum_{k=0}^{n-1} \frac{\pi_n(k)}{\pi_n(\lfloor \underline{q}_n \rfloor)} \\
&\geq \frac{1}{b_n} \sum_{k=\lfloor \underline{q}_n \rfloor - b_n}^{\lfloor \underline{q}_n \rfloor + b_n} \frac{\pi_n(k)}{\pi_n(\lfloor \underline{q}_n \rfloor)} \\
&= \frac{1}{b_n} \sum_{k=\underline{q}_n - b_n}^{\lfloor \underline{q}_n \rfloor} \prod_{\ell=k+1}^{\lfloor \underline{q}_n \rfloor} \frac{1}{\rho_n} \frac{\ell}{n} \frac{1}{\Big(1 + \frac{1}{\sqrt{n-\ell}}\Big)} + \frac{1}{b_n} \sum_{k=\lfloor \underline{q}_n \rfloor + 1}^{\lfloor \underline{q}_n \rfloor + b_n} \prod_{\ell=\lfloor \underline{q}_n \rfloor + 1}^{k} \rho_n \frac{n}{\ell} \Big(1 + \frac{1}{\sqrt{n-\ell}}\Big) \\
&\overset{(a)}{\geq} \frac{1}{b_n} \sum_{k=\lfloor \underline{q}_n \rfloor - b_n}^{\lfloor \underline{q}_n \rfloor} \Big(\underbrace{\frac{1}{\rho_n} \frac{\lfloor \underline{q}_n \rfloor - b_n}{n} \frac{1}{\Big(1 + \frac{1}{\sqrt{n - \lfloor \underline{q}_n \rfloor + b_n}}\Big)}}_{s_{1n}} \Big)^{\lfloor \underline{q}_n \rfloor - k} \\
&\quad + \frac{1}{b_n} \sum_{k=\lfloor \underline{q}_n \rfloor + 1}^{\lfloor \underline{q}_n \rfloor + b_n} \Big(\underbrace{\rho_n \frac{n}{\lfloor \underline{q}_n \rfloor + b_n} \Big(1 + \frac{1}{\sqrt{n - \lfloor \underline{q}_n \rfloor - b_n}}\Big)}_{s_{2n}} \Big)^{k - \lfloor \underline{q}_n \rfloor} \\
&= \frac{1}{b_n} \cdot \frac{1 - s_{1n}^{b_n+1}}{1 - s_{1n}} + \frac{1}{b_n} \cdot \frac{s_{2n} - s_{2n}^{b_n+1}}{1 - s_{2n}},
\end{aligned} \tag{C-7}$$

where (a) comes from the fact that the function

$$h_n(x) = \frac{1}{x} \cdot (1 + \frac{1}{\sqrt{n-x}}),$$

is decreasing in $[0, \underline{q}_n + b_n]$ for n large, we show this at the end of the proof. Next we show that both terms in Eq. (C-7) above converge to a constant. First note that from Theorem 1 we have that $\underline{q}_n = n - z_n^2$ where z_n^2 is given by $n \cdot (1-\rho_n) \cdot r_{0,n}(\rho_n)$. Note that $1 - r_{0,n}(\rho_n)$ is of order $O(n^{-(1-3\alpha)/2})$ if $\alpha < 1/3$ and $r_{0,n}(\rho_n)$ converges to a function of β, $r(\beta)$, for $\alpha = 1/3$

$$r(\beta) = \frac{4}{3} \cdot \cos\Big(\frac{1}{3}\arccos\Big(-\sqrt{\Big(\frac{\beta_1^*}{\beta}\Big)^3}\Big)\Big)^2.$$

For the rest of the proof we will use $\tilde{b}_n$ to denote $b_n + (\underline{q}_n - \lfloor \underline{q}_n \rfloor)$. Note that $|(\underline{q}_n - \lfloor \underline{q}_n \rfloor)| \le 1$. Let $\ell_n = (n - z_n^2 - \tilde{b}_n)/n$, for s_{1n} we have that

$$\begin{aligned}
s_{1n} &= \frac{1}{\rho_n}\ell_n \frac{1}{\Big(1 + \frac{1}{\sqrt{z_n^2 + \tilde{b}_n}}\Big)} \\
&= \frac{1}{\rho_n}\ell_n \Big(1 - \frac{1}{\sqrt{z_n^2 + \tilde{b}_n}} + O(\frac{1}{z_n^2 + \tilde{b}_n})\Big) \\
&= \frac{1}{\rho_n}\ell_n \Big(1 - \frac{1}{z_n}\frac{1}{\sqrt{1 + \frac{\tilde{b}_n}{z_n^2}}} + O(\frac{1}{z_n^2 + \tilde{b}_n})\Big) \\
&= \frac{1}{\rho_n}\ell_n \Big(1 - \frac{1}{z_n} + \frac{\tilde{b}_n}{2z_n^3} + O(\frac{\tilde{b}_n^2}{z_n^5}) + O(\frac{1}{z_n^2 + \tilde{b}_n})\Big) \\
&= \frac{\ell_n}{\rho_n} - \frac{\ell_n}{\rho_n z_n} + \frac{\ell_n \tilde{b}_n}{2\rho_n z_n^3} + O(\frac{\ell_n \tilde{b}_n^2}{\rho_n z_n^5}) + O(\frac{\ell_n}{\rho_n (z_n^2 + \tilde{b}_n)}),
\end{aligned}$$

the last two terms above times b_n converge to zero. Hence,

$$b_n \cdot (1 - s_{1n}) = b_n - \frac{b_n \ell_n}{\rho_n} + \frac{b_n \ell_n}{\rho_n z_n} - \frac{\ell_n b_n \tilde{b}_n}{2\rho_n z_n^3} + o(1).$$

The expression above converges to $C^2 \cdot \Big(1 - \frac{\mathbf{1}_{\{\alpha = 1/3\}}}{2(\beta \cdot r(\beta))^{3/2}}\Big)$. Indeed, the fourth term above is $O(n)/O(n^{\frac{3}{2}(1-\alpha)})$ which is $o(1)$ when $\alpha < 1/3$ and converges to $-C^2/(2\beta^{3/2} \cdot r(\beta)^{3/2})$ when $\alpha = 1/3$. The first three terms converge to C^2. Indeed, recall that $\underline{q}_n$ solves the equation

$$(n - \underline{q}_n) + \frac{n\rho_n}{\sqrt{n - \underline{q}_n}} = n(1 - \rho_n), \quad \text{or equivalently,} \quad z_n^2 + \frac{n\rho_n}{z_n} = n(1 - \rho_n). \tag{C-8}$$

Hence

$$\begin{aligned}
b_n \cdot (1 - s_{1n}) &= b_n - \frac{b_n \ell_n}{\rho_n} + \frac{b_n \ell_n}{\rho_n z_n} - \frac{\ell_n b_n \tilde{b}}{2\rho_n z_n^3} + o(1) \\
&= b_n - \Big(1 - \frac{z_n^2}{n} - \frac{\tilde{b}_n}{n}\Big) \cdot \frac{b_n}{\rho_n} + \Big(1 - \frac{z_n^2}{n} - \frac{\tilde{b}_n}{n}\Big) \cdot \frac{b_n}{\rho_n z_n} - \frac{\ell_n b_n \tilde{b}}{2\rho_n z_n^3} + o(1) \\
&= b_n - \Big(1 - \frac{z_n^2}{n}\Big) \cdot \frac{b_n}{\rho_n} + \frac{b_n}{\rho_n z_n} + \frac{b_n \tilde{b}_n}{\rho_n n} - \frac{\ell_n b_n \tilde{b}}{2\rho_n z_n^3} + o(1) \\
&\overset{\text{Eq.(C-8)}}{=} b_n - \Big(1 - \frac{z_n^2}{n}\Big) \cdot \frac{b_n}{\rho_n} + \frac{b_n}{\rho_n^2}\Big((1 - \rho_n) - \frac{z_n^2}{n}\Big) + \frac{b_n \tilde{b}_n}{\rho_n n} - \frac{\ell_n b_n \tilde{b}}{2\rho_n z_n^3} + o(1)
\end{aligned}$$

$$= \frac{b_n \tilde{b}_n}{\rho_n n} - \frac{\ell_n b_n \tilde{b}}{2\rho_n z_n^3} + o(1)$$
$$\to C^2 - \mathbf{1}_{\{\alpha=1/3\}} \cdot \frac{C^2}{2(\beta \cdot r(\beta))^{3/2}}.$$

Given this, we have

$$\frac{1}{b_n} \cdot \frac{1 - s_{1n}^{b_n+1}}{1 - s_{1n}} = \frac{1 - \exp\Big((b_n+1)\log(s_{1n})\Big)}{b_n(1-s_{1n})}$$
$$= \frac{1 - \exp\Big(-b_n(1-s_{1n}) + o(1)\Big)}{b_n(1-s_{1n})}$$
$$\to \frac{1 - \exp\Big(-C^2 \cdot \Big(1 - \frac{\mathbf{1}_{\{\alpha=1/3\}}}{2(\beta \cdot r(\beta))^{3/2}}\Big)\Big)}{C^2 \cdot \Big(1 - \frac{\mathbf{1}_{\{\alpha=1/3\}}}{2(\beta \cdot r(\beta))^{3/2}}\Big)},$$

note that the function $(\beta r(\beta))^{3/2}$ is strictly increasing and equal to $1/2$ at $\beta = \beta_1^*$. Because we are considering $\beta > \beta_1^*$, the last expression above is positive. Finally, since this limit is a lower bound we obtain the desired lower bound for the $\liminf$.

A similar argument shows that

$$\frac{1}{b_n} \cdot \frac{s_{2n} - s_{2n}^{b_n+1}}{1 - s_{2n}} \to \frac{1 - \exp\Big(-C^2 \cdot \Big(1 - \frac{\mathbf{1}_{\{\alpha=1/3\}}}{2(\beta \cdot r(\beta))^{3/2}}\Big)\Big)}{C^2 \cdot \Big(1 - \frac{\mathbf{1}_{\{\alpha=1/3\}}}{2(\beta \cdot r(\beta))^{3/2}}\Big)}$$

Next we move to the upper bound. We first note that

$$\sum_{k=\lfloor \underline{q}_n \rfloor - b_n}^{\lfloor \underline{q}_n \rfloor + b_n} \pi_n(k) \le \pi_n(\lfloor \underline{q}_n \rfloor) \cdot (2 \cdot b_n + 1).$$

Now we bound the terms in $[0, \lfloor \underline{q}_n \rfloor - b_n - 1]$ and $[\lfloor \underline{q}_n \rfloor + b_n + 1, n-1]$ separately.

$$\frac{1}{b_n \cdot \pi_n(\lfloor \underline{q}_n \rfloor)} \cdot \sum_{k=0}^{\lfloor \underline{q}_n \rfloor - b_n - 1} \pi_n(k) = \frac{1}{b_n} \cdot \sum_{k=0}^{\lfloor \underline{q}_n \rfloor - b_n - 1} \prod_{\ell=k+1}^{\lfloor \underline{q}_n \rfloor} \frac{1}{\rho_n} \cdot \frac{\ell}{n} \cdot \frac{1}{\Big(1 + \frac{1}{\sqrt{n-\ell}}\Big)}$$
$$\overset{(a)}{\le} \frac{1}{b_n} \cdot \sum_{k=0}^{\lfloor \underline{q}_n \rfloor - b_n - 1} \prod_{\ell=k+1}^{\lfloor \underline{q}_n \rfloor - 1} \frac{1}{\rho_n} \cdot \frac{\ell}{n} \cdot \frac{1}{\Big(1 + \frac{1}{\sqrt{n-\ell}}\Big)}$$
$$\overset{(b)}{\le} \frac{1}{b_n} \cdot \sum_{k=0}^{\lfloor \underline{q}_n \rfloor - b_n - 1} \left\{ \frac{1}{\lfloor \underline{q}_n \rfloor - k - 1} \cdot \sum_{\ell=k+1}^{\lfloor \underline{q}_n \rfloor - 1} \frac{1}{\rho_n} \cdot \frac{\ell}{n} \cdot \frac{1}{\Big(1 + \frac{1}{\sqrt{n-\ell}}\Big)} \right\}^{\lfloor \underline{q}_n \rfloor - k - 1}$$
$$\overset{(c)}{\le} \frac{1}{b_n} \cdot \sum_{k=0}^{\lfloor \underline{q}_n \rfloor - b_n - 1} \left\{ \underbrace{\frac{1}{b_n} \cdot \sum_{\ell=\lfloor \underline{q}_n \rfloor - b_n}^{\lfloor \underline{q}_n \rfloor - 1} \frac{1}{\rho_n} \cdot \frac{\ell}{n} \cdot \frac{1}{\Big(1 + \frac{1}{\sqrt{n-\ell}}\Big)}}_{s_{1n}} \right\}^{\lfloor \underline{q}_n \rfloor - k - 1}$$
$$= \frac{s_{1n}^{b_n} - s_{1n}^{\lfloor \underline{q}_n \rfloor - 1}}{b_n \cdot (1 - s_{1n})},$$

where in (a) we use that

$$\frac{1}{\rho_n}\cdot\frac{\lfloor \underline{q}_n\rfloor}{n}\cdot\frac{1}{\Big(1+\frac{1}{\sqrt{n-\lfloor \underline{q}_n\rfloor}}\Big)}\leq 1,$$

in (b) the inequality of arithmetic and geometric means, and in (c) the fact that $h_n(x)$ is decreasing for $x\leq \underline{q}_n+b_n$. In order to simplify notation let $\tilde{z}_n^2=n-\lfloor \underline{q}_n\rfloor$. Let us analyze s_{1n},

$$\begin{aligned}
s_{1n}&=\frac{1}{b_n}\cdot\sum_{\ell=\lfloor \underline{q}_n\rfloor-b_n}^{\lfloor \underline{q}_n\rfloor-1}\frac{1}{\rho_n}\cdot\frac{\ell}{n}\cdot\frac{1}{\Big(1+\frac{1}{\sqrt{n-\ell}}\Big)}\\
&\leq\frac{1}{b_n}\cdot\int_{\lfloor \underline{q}_n\rfloor-b_n}^{\lfloor \underline{q}_n\rfloor}\frac{1}{\rho_n}\cdot\frac{x}{n}\cdot\frac{1}{\Big(1+\frac{1}{\sqrt{n-x}}\Big)}dx\\
&=\frac{1}{b_n n\rho_n}\cdot\Bigg[\frac{1}{6}\Big(-3n^2+n(8\sqrt{n-x}+6)+4x\sqrt{n-x}-12\sqrt{n-x}+3x^2-6x\Big)\\
&\quad-2(n-1)\log(\sqrt{n-x}+1)\Bigg]\Bigg|_{\lfloor \underline{q}_n\rfloor-b_n}^{\lfloor \underline{q}_n\rfloor}\\
&=\frac{1}{b_n n\rho_n}\cdot\Bigg[\frac{12n-4\tilde{z}_n^2-12}{6}\Big(\tilde{z}_n-\sqrt{\tilde{z}_n^2+b_n}\Big)+\frac{4}{6}b_n\sqrt{\tilde{z}_n^2+b_n}+(n-\tilde{z}_n^2)b_n-\frac{b_n^2}{2}\\
&\quad-2(n-1)\log\Big(\frac{\tilde{z}_n+1}{\sqrt{\tilde{z}_n^2+b_n}+1}\Big)\Bigg],
\end{aligned}$$

If we denote this last expression $\tilde{s}_{1n}$ then for $b_n(1-\tilde{s}_{1n})$ we have that

$$\begin{aligned}
b_n(1-\tilde{s}_{1n})&=b_n-\frac{1}{n\rho_n}\Bigg[\frac{12n-4\tilde{z}_n^2-12}{6}\tilde{z}_n\Big(1-\sqrt{1+\frac{b_n}{\tilde{z}_n^2}}\Big)+(n-\tilde{z}_n^2)b_n-\frac{b_n^2}{2}\Bigg]+o(1)\\
&=b_n-\frac{1}{n\rho_n}\Bigg[\frac{12n-4\tilde{z}_n^2-12}{6}\Big(-\frac{b_n}{2\tilde{z}_n}+\frac{b_n^2}{8\tilde{z}_n^3}\Big)+(n-\tilde{z}_n^2)b_n-\frac{b_n^2}{2}\Bigg]+o(1)\\
&\overset{\text{Eq.(C-8)}}{=}b_n-\Big(1-\frac{\tilde{z}_n^2}{n}\Big)\frac{b_n}{\rho_n}+\Big(\frac{(1-\rho_n)}{\rho_n}-\frac{z_n^2}{n\rho_n}\Big)\frac{b_n}{\rho_n}\cdot\frac{z_n}{\tilde{z}_n}-\frac{b_n^2}{4\rho_n\tilde{z}_n^3}+\frac{b_n^2}{2\rho_n n}+o(1)\\
&=b_n\frac{(1-\rho_n)^2}{\rho_n^2}(1-r_{0,n})-\frac{b_n^2}{4\rho_n\tilde{z}_n^3}+\frac{b_n^2}{2\rho_n n}+o(1)\\
&=-\frac{b_n^2}{4\rho_n\tilde{z}_n^3}+\frac{b_n^2}{2\rho_n n}+o(1),
\end{aligned}$$

where in the last equality we used that when $\alpha<1/3$ then $(1-r_{0,n})=O(n^{-(1-3\alpha)/2})$. This last expression converges to $\frac{1}{2}(C^2-\mathbf{1}_{\{\alpha=1/3\}}\cdot\frac{C^2}{2(\beta\cdot r(\beta))^{3/2}})$, which is a positive quantity. Therefore, for n large enough we have $\tilde{s}_{1n}\leq 1$ and, thus

$$\frac{1}{b_n\cdot\pi_n(\lfloor \underline{q}_n\rfloor)}\cdot\sum_{k=0}^{\lfloor \underline{q}_n\rfloor-b_n-1}\pi_n(k)\leq\frac{s_{1n}^{b_n}-s_{1n}^{\lfloor \underline{q}_n\rfloor+1}}{b_n\cdot(1-s_{1n})}\leq\frac{\tilde{s}_{1n}^{b_n}}{b_n\cdot(1-\tilde{s}_{1n})}\to\frac{\exp\Big(-\frac{1}{2}\Big(C^2-\mathbf{1}_{\{\alpha=1/3\}}\cdot\frac{C^2}{2(\beta\cdot r(\beta))^{3/2}}\Big)\Big)}{\frac{1}{2}\Big(C^2-\mathbf{1}_{\{\alpha=1/3\}}\cdot\frac{C^2}{2(\beta\cdot r(\beta))^{3/2}}\Big)},$$

where is the second inequality we used that for n large enough $\tilde{s}_{1n}\leq 1$.

Next we move to the range $[\lfloor \underline{q}_n \rfloor + b_n + 1, n-1]$. First observe that

$$\sum_{k=\lfloor \tilde{q}_n \rfloor}^{n-1} \pi_n(k) \leq \pi_n(n) \cdot (n - \lfloor \tilde{q}_n \rfloor) \leq \frac{\pi_n(n)}{\pi_n(\lfloor \overline{q}_n \rfloor)} \cdot (n - \lfloor \tilde{q}_n \rfloor) \to 0,$$

where the limit follows from Proposition 2 part *i)*. Thus,

$$\begin{aligned}
\frac{1}{b_n \cdot \pi_n(\lfloor \underline{q}_n \rfloor)} \cdot \sum_{k=\lfloor \underline{q}_n \rfloor + b_n + 1}^{n-1} \pi_n(k) &= \frac{1}{b_n} \cdot \sum_{k=\lfloor \underline{q}_n \rfloor + b_n + 1}^{\lfloor \tilde{q}_n \rfloor} \prod_{\ell=\lfloor \underline{q}_n \rfloor + 1}^{k} \rho_n \cdot \frac{n}{\ell} \cdot \Big(1 + \frac{1}{\sqrt{n-\ell}}\Big) + o(1) \\
&\leq \frac{1}{b_n} \cdot \sum_{k=\lfloor \underline{q}_n \rfloor + b_n + 1}^{\lfloor \tilde{q}_n \rfloor} \left\{ \frac{1}{k - \lfloor \underline{q}_n \rfloor} \cdot \sum_{\ell=\lfloor \underline{q}_n \rfloor + 1}^{k} \rho_n \cdot \frac{n}{\ell} \cdot \Big(1 + \frac{1}{\sqrt{n-\ell}}\Big) \right\}^{k - \lfloor \underline{q}_n \rfloor} \\
&\leq \frac{1}{b_n} \cdot \sum_{k=\lfloor \underline{q}_n \rfloor + b_n + 1}^{\lfloor \tilde{q}_n \rfloor} \Bigg\{ \underbrace{\frac{1}{b_n} \cdot \sum_{\ell=\lfloor \underline{q}_n \rfloor + 1}^{\lfloor \underline{q}_n \rfloor + b_n} \rho_n \cdot \frac{n}{\ell} \cdot \Big(1 + \frac{1}{\sqrt{n-\ell}}\Big)}_{s_{2n}} \Bigg\}^{k - \lfloor \underline{q}_n \rfloor} \\
&\leq \frac{s_{2n}^{b_n+1}}{b_n \cdot (1 - s_{2n})}.
\end{aligned}$$

Let us analyze s_{2n},

$$\begin{aligned}
s_{2n} &= \frac{1}{b_n} \cdot \sum_{\ell=\lfloor \underline{q}_n \rfloor + 1}^{\lfloor \underline{q}_n \rfloor + b_n} \rho_n \cdot \frac{n}{\ell} \cdot \Big(1 + \frac{1}{\sqrt{n-\ell}}\Big) \\
&\leq \frac{1}{b_n} \cdot \int_{\lfloor \underline{q}_n \rfloor}^{\lfloor \underline{q}_n \rfloor + b_n} \rho_n \cdot \frac{n}{x} \cdot \Big(1 + \frac{1}{\sqrt{n-x}}\Big) dx \\
&= \frac{\rho_n \cdot n}{b_n} \cdot \Big[\log(x) + \frac{1}{\sqrt{n}} \Big(\log(\sqrt{n} - \sqrt{n-x}) - \log(\sqrt{n-x} + \sqrt{n}) \Big) \Big] \Big|_{\lfloor \underline{q}_n \rfloor}^{\lfloor \underline{q}_n \rfloor + b_n} \\
&= \frac{\rho_n \cdot n}{b_n} \cdot \Big[\log(1 + \frac{b_n}{\lfloor \underline{q}_n \rfloor}) + \frac{1}{\sqrt{n}} \Big(\log \Big[\frac{1 - \frac{\tilde{z}_n}{\sqrt{n}} \cdot \sqrt{1 - \frac{b_n}{\tilde{z}_n^2}}}{1 + \frac{\tilde{z}_n}{\sqrt{n}} \cdot \sqrt{1 - \frac{b_n}{\tilde{z}_n^2}}} \Big] - \log \Big[\frac{1 - \frac{\tilde{z}_n}{\sqrt{n}}}{1 + \frac{\tilde{z}_n}{\sqrt{n}}} \Big] \Big) \Big].
\end{aligned}$$

Denoting this last expression by $\tilde{s}_{2n}$ we have that

$$\begin{aligned}
b_n \cdot (1 - \tilde{s}_{2n}) &= b_n - \rho_n \cdot n \cdot \log(1 + \frac{b_n}{\lfloor \underline{q}_n \rfloor}) - \rho_n \cdot \sqrt{n} \cdot \Big(\log \Big[\frac{1 - \frac{\tilde{z}_n}{\sqrt{n}} \cdot \sqrt{1 - \frac{b_n}{\tilde{z}_n^2}}}{1 + \frac{\tilde{z}_n}{\sqrt{n}} \cdot \sqrt{1 - \frac{b_n}{\tilde{z}_n^2}}} \Big] - \log \Big[\frac{1 - \frac{\tilde{z}_n}{\sqrt{n}}}{1 + \frac{\tilde{z}_n}{\sqrt{n}}} \Big] \Big) \\
&= b_n - \rho_n \cdot n \cdot \Big(\frac{b_n}{\lfloor \underline{q}_n \rfloor} - \frac{b_n^2}{2 \lfloor \underline{q}_n \rfloor^2} \Big) - \rho_n \cdot \sqrt{n} \cdot \Big(\frac{b_n}{\sqrt{n} \tilde{z}_n} + \frac{b_n^2}{4 \sqrt{n} \tilde{z}_n^3} \Big) + o(1) \\
&= \rho_n \cdot n \cdot \frac{b_n^2}{2 \lfloor \underline{q}_n \rfloor^2} - \rho_n \cdot \frac{b_n^2}{4 \tilde{z}_n^3} + b_n \cdot \Big(1 - \frac{\rho_n \cdot n}{\lfloor \underline{q}_n \rfloor} - \frac{\rho_n}{\tilde{z}_n} \Big) + o(1) \\
&\overset{\text{Eq.(C-8)}}{=} \rho_n \cdot n \cdot \frac{b_n^2}{2 \lfloor \underline{q}_n \rfloor^2} - \rho_n \cdot \frac{b_n^2}{4 \tilde{z}_n^3} + b_n \cdot (1 - \rho_n)^2 \cdot (1 - r_n) \cdot r_n + o(1) \\
&= \rho_n \cdot n \cdot \frac{b_n^2}{2 \lfloor \underline{q}_n \rfloor^2} - \rho_n \cdot \frac{b_n^2}{4 \tilde{z}_n^3} + o(1)
\end{aligned}$$

where from the first to second equality we we did a Taylor expansion around zero of the functions $\log(1+x), \log((1-x)/(1+x))$ and $\sqrt{1-x}$, and collected the $o(1)$ terms. In the last equality we used that when $\alpha < 1/3$ then $(1-r_n) = O(n^{-(1-3\alpha)/2})$. As before we can argue that $\tilde{s}_{2n} \leq 1$ for n large. From this we have

$$\frac{1}{b_n \cdot \pi_n(\lfloor \underline{q}_n \rfloor)} \cdot \sum_{k=\lfloor \underline{q}_n \rfloor + b_n + 1}^{n-1} \pi_n(k) \leq \frac{s_{2n}^{b_n+1}}{b_n \cdot (1-s_{2n})} \leq \frac{\tilde{s}_{2n}^{b_n}}{b_n \cdot (1-\tilde{s}_{2n})} \to \frac{\exp\Big(-\frac{1}{2}\Big(C^2 - \mathbf{1}_{\{\alpha=1/3\}} \cdot \frac{C^2}{2(\beta \cdot r(\beta))^{3/2}}\Big)\Big)}{\frac{1}{2}\Big(C^2 - \mathbf{1}_{\{\alpha=1/3\}} \cdot \frac{C^2}{2(\beta \cdot r(\beta))^{3/2}}\Big)}.$$

Finally, since this limit is an upper bound we obtain the desired upper bound for the lim sup.

Remaining proofs. Let

$$h_n(x) = \frac{1}{x} \cdot (1 + \frac{1}{\sqrt{n-x}}),$$

we show is decreasing in $(0, \underline{q}_n + b_n]$ for n large. First,

$$\frac{dh_n}{dx}(x) = -\frac{1}{x^2} \cdot \Big(1 + \frac{1}{\sqrt{n-x}}\Big) + \frac{1}{2x}(n-x)^{-3/2},$$

so $h_n(x)$ is decreasing if and only if $x \leq 2((n-x)^{3/2} + n - x)$. Note that the LHS in the previous inequality is strictly increasing and the RHS is strictly decreasing. Also, at $x = 0$ the LHS is below the RHS, and for $x = n$ the converse is true. Therefore, there if for some y,

$$\frac{y}{2} \leq (n-y)^{3/2} \cdot \Big(1 + \frac{1}{\sqrt{n-y}}\Big) \tag{C-9}$$

then the same is true for all $x \leq y$. Consider $y = \underline{q}_n + b_n$ and let $\ell_n = 1 - \frac{b_n}{n - \underline{q}_n}$

$$\begin{aligned}
(n - \underline{q}_n - b_n)^{3/2} \cdot \Big(1 + \frac{1}{\sqrt{n - \underline{q}_n - b_n}}\Big) &= (n - \underline{q}_n)^{3/2} \ell_n^{3/2} \cdot \Big(1 + \frac{1}{\sqrt{\ell_n}} \frac{1}{\sqrt{n - \underline{q}_n}}\Big) \\
&\overset{\text{Eq.(C-8)}}{=} (n - \underline{q}_n)^{3/2} \ell_n^{3/2} \cdot \Big(1 + \frac{1}{\sqrt{\ell_n}} \cdot \frac{(\underline{q}_n - n\rho_n)}{n\rho_n}\Big) \\
&= (n - \underline{q}_n)^{3/2} \ell_n \cdot \Big(\frac{n\rho_n(\sqrt{\ell_n} - 1) + \underline{q}_n}{n\rho_n}\Big),
\end{aligned}$$

note that for n large enough $n\rho_n(\sqrt{\ell_n} - 1) + \underline{q}_n > 0$. Then Eq. (C-9) is satisfied if and only if

$$\frac{\rho_n}{2} \cdot \underbrace{\Big[\frac{\underline{q}_n + b_n}{\ell_n(n\rho_n(\sqrt{\ell_n} - 1) + \underline{q}_n)}\Big]}_{H_n} \leq \frac{(n - \underline{q}_n)^{3/2}}{n} = n^{(1-3\alpha)/2}(n^\alpha(1-\rho_n)r_{0,n}(\rho_n))^{3/2}, \tag{C-10}$$

where we used that $n - \underline{q}_n = n(1-\rho_n)r_{0,n}(\rho_n)$. Since $\ell_n \to 1$, $H_n \to 1$. If $\alpha < 1/3$ then for n large enough the previous inequality hold. If $\alpha = 1/3$ and $\beta > \beta_1^*$, the LHS in Eq (C-10) converges to $1/2$ and the RHS to $(\beta r(\beta))^{3/2}$. This last function is strictly increasing and equal to $1/2$ at $\beta = \beta_1^*$. This implies that for n large enough Eq. (C-10) is satisfied, completing the proof.

□

Proof of Lemma C-2. We start we the lower bound, let $b_n = Cn^{\frac{3\alpha}{2}}$ and note that

$$\begin{aligned}
\frac{\mathbf{P}[Q_n(\infty) \geq n]}{b_n \pi_n(\lfloor \overline{q}_n \rfloor)} &= \frac{1}{b_n} \sum_{k=n}^{\infty} \frac{\pi_n(k)}{\pi_n(\lfloor \overline{q}_n \rfloor)} \\
&\geq \frac{1}{b_n} \sum_{k=\lfloor \overline{q}_n \rfloor - b_n}^{\lfloor \overline{q}_n \rfloor + b_n} \frac{\pi_n(k)}{\pi_n(\lfloor \overline{q}_n \rfloor)} \\
&= \frac{1}{b_n} \sum_{k=\lfloor \overline{q}_n \rfloor - b_n}^{\lfloor \overline{q}_n \rfloor} \prod_{\ell=k+1}^{\lfloor \overline{q}_n \rfloor} \frac{1}{\rho_n} \frac{1}{\left(1 + \frac{1}{\sqrt{\ell - n}}\right)} + \frac{1}{b_n} \sum_{k=\lfloor \overline{q}_n \rfloor + 1}^{\lfloor \overline{q}_n \rfloor + b_n} \prod_{\ell=\lfloor \overline{q}_n \rfloor + 1}^{k} \rho_n \Big(1 + \frac{1}{\sqrt{\ell - n}}\Big) \\
&\geq \frac{1}{b_n} \sum_{k=\lfloor \overline{q}_n \rfloor - b_n}^{\lfloor \overline{q}_n \rfloor} \Bigg[\underbrace{\frac{1}{\rho_n} \frac{1}{\left(1 + \frac{1}{\sqrt{\lfloor \overline{q}_n \rfloor - b_n - n}}\right)}}_{s_{1n}} \Bigg]^{\lfloor \overline{q}_n \rfloor - k} + \frac{1}{b_n} \sum_{k=\lfloor \overline{q}_n \rfloor + 1}^{\lfloor \overline{q}_n \rfloor + b_n} \Bigg[\underbrace{\rho_n \Big(1 + \frac{1}{\sqrt{\lfloor \overline{q}_n \rfloor + b_n - n}}\Big)}_{s_{2n}} \Bigg]^{k - \lfloor \overline{q}_n \rfloor} \\
&= \frac{1}{b_n} \cdot \frac{1 - s_{1n}^{b_n+1}}{1 - s_{1n}} + \frac{1}{b_n} \cdot \frac{s_{2n} - s_{2n}^{b_n+1}}{1 - s_{2n}}.
\end{aligned}$$

Next we compute limits for $b_n(1 - s_{1n})$ and $b_n(1 - s_{2n})$. Before we begin note that $\overline{q}_n = n + z_n^2$ where $z_n^2 = \rho_n^2/(1-\rho_n)^2$ and let $\tilde{b}_n = b_n + (\overline{q}_n - \lfloor \overline{q}_n \rfloor)$ then

$$\begin{aligned}
b_n(1 - s_{1n}) &= \frac{b_n}{\tilde{b}_n} \Big[\tilde{b}_n - \tilde{b}_n \cdot \frac{1}{\rho_n} \frac{1}{\left(1 + \frac{1}{\sqrt{\lfloor \overline{q}_n \rfloor - b_n - n}}\right)} \Big] \\
&= \frac{b_n}{\tilde{b}_n} \Big[\tilde{b}_n - \tilde{b}_n \cdot \frac{1}{\rho_n} \Big(1 - \frac{1}{\sqrt{z_n^2 - \tilde{b}_n}}\Big) + o(1) \Big] \\
&= \frac{b_n}{\tilde{b}_n} \Big[\tilde{b}_n - \tilde{b}_n \cdot \frac{1}{\rho_n} \Big(1 - \frac{1}{z_n} \Big\{ 1 + \frac{\tilde{b}_n}{2z_n^2} \Big\} \Big) + o(1) \Big] \\
&= \frac{b_n}{\tilde{b}_n} \Big[\tilde{b}_n \frac{(1-\rho_n)^2}{\rho_n^2} + \frac{\tilde{b}_n^2}{2\rho_n^4} \cdot (1-\rho_n)^3 + o(1) \Big] \\
&\to \frac{C^2 \beta^3}{2}.
\end{aligned}$$

Thus,

$$\frac{1}{b_n} \cdot \frac{1 - s_{1n}^{b_n+1}}{1 - s_{1n}} \to \frac{1 - \exp\Big(-\frac{C^2\beta^3}{2}\Big)}{\frac{C^2\beta^3}{2}}.$$

For $b_n(1 - s_{2n})$ we have

$$\begin{aligned}
b_n(1 - s_{2n}) &= \frac{b_n}{\tilde{b}_n} \Big[\tilde{b}_n - \tilde{b}_n \cdot \rho_n \Big(1 + \frac{1}{\sqrt{z_n^2 + \tilde{b}_n}}\Big) \Big] \\
&= \frac{b_n}{\tilde{b}_n} \Big[\tilde{b}_n - \tilde{b}_n \cdot \rho_n \Big(1 + \frac{1}{z_n} \Big\{ 1 - \frac{\tilde{b}_n}{2z_n^2} \Big\} \Big) + o(1) \Big] \\
&= \frac{b_n}{\tilde{b}_n} \Big[\frac{\tilde{b}_n^2}{2\rho_n^2} \cdot (1-\rho_n)^3 + o(1) \Big] \\
&\to \frac{C^2 \beta^3}{2}.
\end{aligned}$$

Thus,

$$\frac{1}{b_n}\cdot\frac{s_{2n}-s_{2n}^{b_n+1}}{1-s_{2n}}\to\frac{1-\exp\left(-\frac{C^2\beta^3}{2}\right)}{\frac{C^2\beta^3}{2}}.$$

Finally, since this limit is a lower bound we obtain the desired lower bound for the lim inf.

For the upper bound note that

$$\frac{\mathbf{P}[Q_n(\infty)\geq n]}{b_n\pi_n(\lfloor\overline{q}_n\rfloor)}=\frac{1}{b_n}\sum_{k=n}^{\infty}\frac{\pi_n(k)}{\pi_n(\lfloor\overline{q}_n\rfloor)}\leq 2+\frac{1}{b_n}\sum_{k=n}^{\lfloor\overline{q}_n\rfloor-b_n}\frac{\pi_n(k)}{\pi_n(\lfloor\overline{q}_n\rfloor)}+\frac{1}{b_n}\sum_{k=\lfloor\overline{q}_n\rfloor+b_n+1}^{\infty}\frac{\pi_n(k)}{\pi_n(\lfloor\overline{q}_n\rfloor)},\qquad\text{(C-11)}$$

so we just need to upper bound both summation on the right hand side of Eq. (C-11) and take the limit. For the first summation we have

$$\begin{aligned}\frac{1}{b_n}\sum_{k=n}^{\lfloor\overline{q}_n\rfloor-b_n}\frac{\pi_n(k)}{\pi_n(\lfloor\overline{q}_n\rfloor)}&=\frac{1}{b_n}\sum_{k=n}^{\lfloor\overline{q}_n\rfloor-b_n}\prod_{\ell=k+1}^{\lfloor\overline{q}_n\rfloor}\frac{1}{\rho_n\cdot\left(1+\frac{1}{\sqrt{\ell-n}}\right)}\\&\overset{(a)}{\leq}\frac{1}{b_n}\sum_{k=n}^{\lfloor\overline{q}_n\rfloor-b_n}\Bigg[\underbrace{\frac{1}{b_n}\cdot\sum_{\ell=\lfloor\overline{q}_n\rfloor-b_n+1}^{\lfloor\overline{q}_n\rfloor-1}\frac{1}{\rho_n\cdot\left(1+\frac{1}{\sqrt{\ell-n}}\right)}}_{s_{1n}}\Bigg]^{\lfloor\overline{q}_n\rfloor-k}\\&\leq\frac{1}{b_n}\frac{s_{1n}^{b_n}}{1-s_{1n}},\end{aligned}$$

where in (a) we used the inequality of arithmetic and geometric means, and the fact that the function inside the summation is increasing. For s_{1n} we have

$$\begin{aligned}s_{1n}&=\frac{1}{\rho_n\cdot b_n}\cdot\sum_{\ell=\lfloor\overline{q}_n\rfloor-b_n+1}^{\lfloor\overline{q}_n\rfloor-1}\frac{1}{\left(1+\frac{1}{\sqrt{\ell-n}}\right)}\\&\leq\frac{1}{\rho_n\cdot b_n}\cdot\int_{\overline{q}_n-b_n}^{\overline{q}_n}\frac{1}{\left(1+\frac{1}{\sqrt{x-n}}\right)}dx\\&=\frac{1}{\rho_n\cdot b_n}\cdot\Big[-2\sqrt{x-n}+2\log(\sqrt{x-n}+1)+x-n\Big]\Big|_{\overline{q}_n-b_n}^{\overline{q}_n}\\&=\frac{1}{\rho_n\cdot b_n}\cdot\Big[-2z_n+2\log(z_n+1)+2\sqrt{z_n^2-b_n}-2\log(\sqrt{z_n^2-b_n}+1)+b_n\Big],\end{aligned}$$

then denoting the last expression above by $\tilde{s}_{1n}$ we have

$$\begin{aligned}b_n\cdot(1-\tilde{s}_{1n})&=b_n-\frac{1}{\rho_n}\cdot\Big[-2z_n+2\log(z_n+1)+2\sqrt{z_n^2-b_n}-2\log(\sqrt{z_n^2-b_n}+1)+b_n\Big]\\&=b_n+\frac{b_n}{\rho_n z_n}+\frac{b_n^2}{4\rho_n z_n^3}-\frac{b_n}{\rho_n}+o(1)\\&\to\frac{C^2\beta^3}{4}.\end{aligned}$$

Hence, since (for n large) $\tilde{s}_{1n}\leq 1$ we have

$$\frac{1}{b_n}\sum_{k=n}^{\overline{q}_n-b_n}\frac{\pi_n(k)}{\pi_n(\overline{q}_n)}\leq\frac{1}{b_n}\frac{s_{1n}^{b_n}}{1-s_{1n}}\leq\frac{1}{b_n}\frac{\tilde{s}_{1n}^{b_n}}{1-\tilde{s}_{1n}}\to\frac{\exp\left(-\frac{C^2\beta^3}{4}\right)}{\frac{C^2\beta^3}{4}}.$$

Now let us consider the second summation in Eq. (C-11),

$$\begin{aligned}\frac{1}{b_n}\sum_{k=\lfloor \overline{q}_n\rfloor+b_n+1}^{\infty}\frac{\pi_n(k)}{\pi_n(\lfloor \overline{q}_n\rfloor)} &= \frac{1}{b_n}\sum_{k=\lfloor \overline{q}_n\rfloor+b_n+1}^{\infty}\prod_{\ell=\lfloor \overline{q}_n\rfloor+1}^{k}\rho_n\Big(1+\frac{1}{\sqrt{\ell-n}}\Big)\\ &\le \frac{1}{b_n}\sum_{k=\lfloor \overline{q}_n\rfloor+b_n}^{\infty}\Bigg[\underbrace{\frac{1}{b_n}\cdot\sum_{\ell=\lfloor \overline{q}_n\rfloor+1}^{\lfloor \overline{q}_n\rfloor+b_n}\rho_n\cdot\Big(1+\frac{1}{\sqrt{\ell-n}}\Big)}_{s_{2n}}\Bigg]^{k-\lfloor \overline{q}_n\rfloor}\\ &= \frac{1}{b_n}\frac{s_{2n}^{b_n}}{1-s_{2n}},\end{aligned}$$

where we used the inequality of arithmetic and geometric means, the fact that the function inside the summation is decreasing, and that for app $\ell \ge \lfloor \overline{q}_n\rfloor+1$ the terms in the summation are strictly bounded above by 1. For s_{2n}, if we let $\tilde{z}_n^2 = \lfloor \overline{q}_n\rfloor - n$, we have

$$\begin{aligned}s_{2n} &\le \frac{\rho_n}{b_n}\cdot\int_{\lfloor \overline{q}_n\rfloor}^{\lfloor \overline{q}_n\rfloor+b_n}\Big(1+\frac{1}{\sqrt{x-n}}\Big)dx\\ &= \frac{\rho_n}{b_n}\cdot\Big[2\sqrt{x-n}+x\Big]\Big|_{\lfloor \overline{q}_n\rfloor}^{\lfloor \overline{q}_n\rfloor+b_n}\\ &= \frac{\rho_n}{b_n}\cdot\Big[2\sqrt{\tilde{z}_n^2+b_n}-2\tilde{z}_n+b_n\Big],\end{aligned}$$

denoting this last term by $\tilde{s}_{2n}$ we have

$$\begin{aligned}b_n\cdot(1-\tilde{s}_{2n}) &= b_n-\rho_n\cdot\Big[2\sqrt{\tilde{z}_n^2+b_n}-2\tilde{z}_n+b_n\Big]\\ &= b_n(1-\rho_n)-\rho_n\cdot\Big[\frac{b_n}{\tilde{z}_n}-\frac{b_n^2}{4\tilde{z}_n^3}\Big]+o(1)\\ &\to \frac{C^2\beta^3}{4}.\end{aligned}$$

Thus, since $\tilde{s}_{2n}\le 1$ (for n large) we have

$$\frac{1}{b_n}\sum_{k=\lfloor \overline{q}_n\rfloor+b_n+1}^{\infty}\frac{\pi_n(k)}{\pi_n(\lfloor \overline{q}_n\rfloor)}\le\frac{1}{b_n}\frac{s_{2n}^{b_n}}{1-s_{2n}}\le\frac{1}{b_n}\frac{\tilde{s}_{2n}^{b_n}}{1-\tilde{s}_{2n}}\to\frac{\exp\Big(-\frac{C^2\beta^3}{4}\Big)}{\frac{C^2\beta^3}{4}}.$$

Finally, since this limit is an upper bound we obtain the desired upper bound for the lim sup. □

Proof of Lemma 2. This result is a direct consequence of Lemmata C-1 and C-2 which were stated and proved right before the present proof. □

Proof of Proposition 3. Consider the following

$$\rho_n(y) = 1-\frac{\beta_2^*}{n^{1/3}}-\frac{y}{n^{1/3}},\quad y\in(-(\beta_2^*-\beta_1^*),(\beta_2^*-\beta_1^*)) = D.$$

Note that $n^{1/3}(1-\rho_n(y)) = \beta_2^*+y > \beta_1^*$ and $\rho_n(y)\uparrow 1$, hence we can always find n_1 such that for all $n\ge n_1$ the leftmost equilibrium $\underline{q}_n$ is well defined. Note that for $\rho_n(y)$ we have

$$\log\Big(\frac{\pi_n(\lfloor \overline{q}_n\rfloor)}{\pi_n(\lfloor \underline{q}_n\rfloor)}\Big) = (\lfloor \overline{q}_n\rfloor-\lfloor \underline{q}_n\rfloor)\log(\rho_n(y))+\sum_{k=\lfloor \underline{q}_n\rfloor+1}^{n-1}\log\Big[\frac{n}{k}\cdot\Big(1+\frac{1}{\sqrt{n-k}}\Big)\Big]+\sum_{k=1}^{\lfloor \overline{q}_n\rfloor-n}\log\Big[1+\frac{1}{\sqrt{k}}\Big]+\log(2). \tag{C-12}$$

Furthermore, observe that both $\underline{q}_n$ and $\overline{q}_n$ are continuous functions of y,

$$\underline{q}_n(y) = n - n(1-\rho_n(y)) \cdot r_{0,n}(\rho_n(y)) \quad \text{and} \quad \overline{q}_n(y) = n + \frac{\rho_n(y)^2}{(1-\rho_n(y))^2}.$$

Define,

$$f_n(y) \triangleq \log\Big(\frac{\pi_n(\lfloor \overline{q}_n(y) \rfloor)}{\pi_n(\lfloor \underline{q}_n(y) \rfloor)}\Big).$$

since we are using the floor function, $f_n(\cdot)$ might not be continuos. In the first step of this proof we show that the potential jumps of $f_n(\cdot)$ in D converge to zero (Step 1). Then we show that there exists a sequence γ_n^c such that $f_n(\gamma_n^c) \to c$ (Step 2) and $\gamma_n^c \to 0$ (Step 3).

Step 1. Fix $\epsilon > 0$. First, we prove that there exists $\tilde{n}$ such that for all $n \geq \tilde{n}$ we have that

$$\forall y \in D,\ \exists \delta > 0 \text{ such that } \forall \tilde{y}\colon |\tilde{y} - y| < \delta \Rightarrow |f_n(\tilde{y}) - f_n(y)| < \epsilon. \tag{C-13}$$

We choose $\tilde{n}$ such that for all $n \geq \tilde{n}$:

- $\sup_{z\in D} 2|\log(\rho_n(z))| \leq \epsilon/9$. This is possible because $\rho_n(z) \to 1$ uniformly in D.
- $$\sup_{z\in D} \Big|\log\Big[\frac{n}{\lfloor \underline{q}_n(z) \rfloor + 1}\Big]\Big| \leq \frac{\epsilon}{6}, \quad \text{and} \quad \sup_{z\in D} \Big|\log\Big(1 + \frac{1}{\sqrt{n - \lfloor \underline{q}_n(z) \rfloor - 1}}\Big)\Big| \leq \frac{\epsilon}{6}.$$

This is possible because for any $z \in D$, $n/(\lfloor \underline{q}_n(z) \rfloor + 1) \to 1$.

- $$\sup_{z\in D} \Big|\log\Big[1 + \frac{1}{\sqrt{\lfloor \overline{q}_n(z) \rfloor - n}}\Big]\Big| \leq \frac{\epsilon}{3}.$$

This is possible because for any $z \in D$, $(\lfloor \overline{q}_n(z) \rfloor - n) \uparrow \infty$.

Let $n \geq n_1$ and fix $y \in D$, we consider the first three terms in $f_n(\cdot)$, see Eq. (C-12). Let $Q_n(\tilde{y}) = \lfloor \overline{q}_n(\tilde{y}) \rfloor - \lfloor \underline{q}_n(\tilde{y}) \rfloor$ and $R_n(\tilde{y}) = \overline{q}_n(\tilde{y}) - \underline{q}_n(\tilde{y})$, and note that $|Q_n(\tilde{y}) - R_n(\tilde{y})| \leq 2$ for any $\tilde{y}$. Also, $R_n(\tilde{y})\log(\rho_n(\tilde{y}))$ is continuous; therefore, there exists δ_1 such that

$$|R_n(\tilde{y})\log(\rho_n(\tilde{y})) - R_n(y)\log(\rho_n(y))| \leq \epsilon/9, \quad \forall \tilde{y}\colon |\tilde{y} - y| < \delta_1.$$

Using this, for the first term in Eq. (C-12), we have

$$\begin{aligned}
\Big|Q_n(\tilde{y})\log(\rho_n(\tilde{y})) - Q_n(y)\log(\rho_n(y))\Big| &= \Big|(Q_n(\tilde{y}) - R_n(\tilde{y}))\log(\rho_n(\tilde{y})) + R_n(\tilde{y})\log(\rho_n(\tilde{y})) \\
&\quad - (Q_n(y) - R_n(y))\log(\rho_n(y)) - R_n(y)\log(\rho_n(y))\Big| \\
&\leq 2|\log(\rho_n(\tilde{y}))| + 2|\log(\rho_n(y))| \\
&\quad + |R_n(\tilde{y})\log(\rho_n(\tilde{y})) - R_n(y)\log(\rho_n(y))| \\
&\leq \frac{\epsilon}{3},
\end{aligned}$$

for all $\tilde{y}$ such that $|\tilde{y}-y|<\delta_1$. For the second term in Eq. (C-12), observe that since $\underline{q}_n(\cdot)$ is continuous there always exists $\delta_2>0$ such that for all $\tilde{y}$ with $|\tilde{y}-y|<\delta_2$ we have $|\lfloor \underline{q}_n(\tilde{y})\rfloor - \lfloor \underline{q}_n(y)\rfloor|\leq 1$. Therefore,

$$\begin{aligned}
&\Big|\sum_{k=\lfloor \underline{q}_n(\tilde{y})\rfloor+1}^{n-1}\log\Big[\frac{n}{k}\cdot\Big(1+\frac{1}{\sqrt{n-k}}\Big)\Big] \\
&-\sum_{k=\lfloor \underline{q}_n(y)\rfloor+1}^{n-1}\log\Big[\frac{n}{k}\cdot\Big(1+\frac{1}{\sqrt{n-k}}\Big)\Big]\Big| \leq \Big|\log\Big[\frac{n}{\lfloor \underline{q}_n(\tilde{y})\rfloor+1}\cdot\Big(1+\frac{1}{\sqrt{n-\lfloor \underline{q}_n(\tilde{y})\rfloor-1}}\Big)\Big]\Big| \\
&\qquad\qquad \leq \Big|\log\Big[\frac{n}{\lfloor \underline{q}_n(\tilde{y})\rfloor+1}\Big]\Big| \\
&\qquad\qquad + \Big|\log\Big(1+\frac{1}{\sqrt{n-\lfloor \underline{q}_n(\tilde{y})\rfloor-1}}\Big)\Big| \\
&\qquad\qquad \leq \frac{\epsilon}{3}.
\end{aligned}$$

Finally, for the third term in Eq. (C-12), since $\overline{q}_n(\cdot)$ is continuous there always exists $\delta_3>0$ such that for all $\tilde{y}$ with $|\tilde{y}-y|<\delta_3$ we have $|\lfloor \overline{q}_n(\tilde{y})\rfloor - \lfloor \overline{q}_n(y)\rfloor|\leq 1$. Therefore,

$$\Big|\sum_{k=1}^{\lfloor \overline{q}_n(\tilde{y})\rfloor-n}\log\Big[1+\frac{1}{\sqrt{k}}\Big]-\sum_{k=1}^{\lfloor \overline{q}_n(y)\rfloor-n}\log\Big[1+\frac{1}{\sqrt{k}}\Big]\Big|\leq\Big|\log\Big[1+\frac{1}{\sqrt{\lfloor \overline{q}_n(y)\rfloor-n}}\Big]\Big|\leq\frac{\epsilon}{3}$$

Putting the three inequalities jus proved together, for $\delta\leq\min\{\delta_1,\delta_2,\delta_3\}$, delivers Eq. (C-13). Next define

$$\Delta_n \triangleq \sup_{y\in D}|f_n(y^+)-f_n(y^-)|,$$

then Eq. (C-13) ensures that $\Delta_n\to 0$.

Step 2. We construct γ_n^c and show that $f_n(\gamma_n^c)\to c$. Fix $y_1\in(-(\beta_2^*-\beta_1^*),0)$ and $y_2\in(0,\beta_2^*-\beta_1^*)$, we next argue that there exists n_2 such that for all $n\geq n_2$ it holds that $f_n(y_1)>c>f_n(y_2)$. Indeed, consider first y_1 and note that $\beta_2^*+y_1\in(\beta_1^*,\beta_2^*)$. For $g(\cdot)$ as in Proposition 2 part *iii)*, one has $g(\beta_2^*+y_1)>0$. So, again by Proposition 2 part *iii)* we have that for any $\epsilon_1\in(0,g(\beta_2^*+y_1))$ there exists $n_{1,2}$ such that for all $n\geq n_{1,2}$ we have

$$c<n^{1/3}\cdot(g(\beta_1)-\epsilon_1)<f_n(y_1).$$

A similar argument that leverages the fact that $g(\beta_2^*+y_2)<0$ shows that there exists $n_{2,2}$ such that for all $n\geq n_{2,2}$ we have $f_n(y_2)<c$. We take $n_2=\max\{n_{1,2},n_{2,2}\}$ to conclude that for all $n\geq n_2$ it holds that $f_n(y_1)>c>f_n(y_2)$. To conclude consider $n\geq\max\{n_1,n_2\}$ then, by Step 1 we can always find $\gamma_n^c\in(y_1,y_2)$ such that

$$c-\frac{\Delta_n}{2}\leq f_n(\gamma_n^c)\leq c+\frac{\Delta_n}{2}$$

Taking limit at both sides and using that $\Delta_n \to 0$, we conclude that $f_n(\gamma_n^c) \to c$.

Step 3. To conclude the proof we need to argue that $\gamma_n^c \to 0$. Note from the argument above $\{\gamma_n^c\}$ is a bounded sequence. For the sake of contradiction fix $\epsilon > 0$ and suppose that

$$\limsup_{n\to\infty} \gamma_n^c > \epsilon.$$

This implies that there exists a subsequence $\{\gamma_{k(n)}^c\}$ that convergences to a point $\hat{\gamma}^c \geq \epsilon$. Let

$$\hat{\rho}_n = 1 - \frac{\beta_2^*}{n^{1/3}} - \frac{\gamma_n^c}{n^{1/3}},$$

then $k(n)^{1/3}(1-\hat{\rho}_{k(n)}) \to \beta_2^\star + \hat{\gamma}^c$. Because $g(\beta_2^\star + \hat{\gamma}^c) < 0$ from Proposition 2, for $\epsilon' > 0$ such that $g(\beta_2^\star + \hat{\gamma}^c) + \epsilon' < 0$, we can deduce that for all n large enough

$$f_{k(n)}(\gamma_{k(n)}^c) \leq n^{1/3}(g(\beta_2^\star + \hat{\gamma}^c) + \epsilon') \leq c - \epsilon'.$$

However, from Step 1 we know that $f_{k(n)}(\gamma_{k(n)}^c) \to c$. This, together with the previous inequality yields a contradiction. The case when $\liminf_{n\to\infty} \gamma_n^c < \epsilon$ can be treated similarly and is thus omitted. Therefore, for any $\epsilon > 0$

$$\epsilon \leq \liminf_{n\to\infty} \gamma_n^c \leq \limsup_{n\to\infty} \gamma_n^c \leq \epsilon,$$

since ϵ is arbitrary we have that $\gamma_n^c \to 0$, which concludes the proof.

□

Proof of Proposition 4. We prove both statement separately.

(i) We show that

$$\lim_{n\to\infty} \mathbf{P}[Q_n(\infty) < \lfloor \overline{q}_n \rfloor - C \cdot \sqrt{\log(n)} \cdot n^{1.5\alpha}] = 0,$$

the other case in analogous. To reduce notation let $b_n = C \cdot \sqrt{\log(n)} \cdot n^{1.5\alpha}$ for some $C > 0$ that we will choose later in the proof then

$$\begin{aligned}\mathbf{P}[Q_n(\infty) < \lfloor \overline{q}_n \rfloor - b_n] &\leq \mathbf{P}[Q_n(\infty) < n] + \mathbf{P}[n \leq Q_n(\infty) \leq \lfloor \overline{q}_n \rfloor - b_n] \\ &= \mathbf{P}[Q_n(\infty) < n] + \sum_{k=n}^{\lfloor \overline{q}_n \rfloor - b_n} \pi_n(k)\end{aligned}$$

by Theorem 2 part *(i)* the first term converges to zero. For the second term we have the following upper bound

$$\begin{aligned}\sum_{k=n}^{\lfloor \overline{q}_n \rfloor - b_n} \pi_n(k) &\leq \sum_{k=n}^{\lfloor \overline{q}_n \rfloor - b_n} \prod_{\ell=k+1}^{\lfloor \overline{q}_n \rfloor} \frac{1}{\rho_n \cdot \left(1 + \frac{1}{\sqrt{\ell - n}}\right)} \\ &\overset{(a)}{\leq} \sum_{k=n}^{\lfloor \overline{q}_n \rfloor - b_n} \left[\frac{1}{\lfloor \overline{q}_n \rfloor - k} \cdot \sum_{\ell=k+1}^{\lfloor \overline{q}_n \rfloor - 1} \frac{1}{\rho_n \cdot \left(1 + \frac{1}{\sqrt{\ell - n}}\right)}\right]^{\lfloor \overline{q}_n \rfloor - k}\end{aligned}$$

$$\leq \sum_{k=n}^{\lfloor \overline{q}_n \rfloor - b_n} \Bigg[\underbrace{\frac{1}{b_n} \cdot \sum_{\ell=\lfloor \overline{q}_n \rfloor - b_n + 1}^{\lfloor \overline{q}_n \rfloor - 1} \frac{1}{\rho_n \cdot \Big(1 + \frac{1}{\sqrt{\ell - n}}\Big)}}_{s_n} \Bigg]^{\lfloor \overline{q}_n \rfloor - k}$$

$$\leq \frac{s_n^{b_n}}{1 - s_n},$$

where in (a) we used the inequality of arithmetic and geometric means. We next show the last term above converges to zero.

Recall that $\overline{q}_n = n + z_n^2$ where $z_n = \frac{\rho_n}{(1-\rho_n)}$. We have

$$\begin{aligned}
s_n &= \frac{1}{\rho_n \cdot b_n} \cdot \sum_{\ell=\lfloor \overline{q}_n \rfloor - b_n + 1}^{\lfloor \overline{q}_n \rfloor - 1} \frac{1}{\Big(1 + \frac{1}{\sqrt{\ell - n}}\Big)} \\
&\leq \frac{1}{\rho_n \cdot b_n} \cdot \int_{\overline{q}_n - b_n}^{\overline{q}_n} \frac{1}{\Big(1 + \frac{1}{\sqrt{x - n}}\Big)} dx \\
&= \frac{1}{\rho_n \cdot b_n} \cdot \Big[-2\sqrt{x-n} + 2\log(\sqrt{x-n} + 1) + x - n \Big] \Big|_{\overline{q}_n - b_n}^{\overline{q}_n} \\
&= \frac{1}{\rho_n \cdot b_n} \cdot \Big[-2z_n + 2\log(z_n + 1) + 2\sqrt{z_n^2 - b_n} - 2\log(\sqrt{z_n^2 - b_n} + 1) + b_n \Big],
\end{aligned}$$

denote this last term by $\tilde{s}_n$. Then

$$\begin{aligned}
\tilde{s}_n &= \frac{1}{\rho_n \cdot b_n} \cdot \Big[-2z_n + 2\Big(\frac{1}{z_n} + O(n^{-2\alpha})\Big) + 2z_n \Big(1 - \frac{b_n}{2z_n^2} - \frac{b_n^2}{8z_n^4} + O\Big(\frac{b_n^3}{z_n^6}\Big)\Big) \\
&\quad - 2\Big(\sqrt{1 - \frac{b_n}{z_n^2}} + \frac{1}{z_n} - 1 + O\Big(\frac{b_n^2}{z_n^4}\Big)\Big) + b_n \Big] \\
&= \frac{1}{\rho_n \cdot b_n} \cdot \Big[2z_n \Big(- \frac{b_n}{2z_n^2} - \frac{b_n^2}{8z_n^4} \Big) - 2\Big(- \frac{b_n}{2z_n^2} \Big) + b_n \Big] + O(n^{-2\alpha} \log(n)) \\
&= \Big[1 + \frac{(1-\rho_n)^3}{\rho_n^3} - \frac{b_n (1-\rho_n)^3}{4\rho_n^4} \Big] + O(n^{-2\alpha} \log(n)).
\end{aligned}$$

Hence, $\tilde{s}_n \to 1$ and

$$b_n \cdot (1 - \tilde{s}_n) = \underbrace{\frac{(1-\rho_n)^3}{\rho_n^3}}_{O(n^{-3\alpha})} \cdot \underbrace{\frac{b_n^2}{4\rho_n}}_{O(n^{3\alpha}\log(n))} + O(n^{-\alpha/2} \log(n)^{3/2}) = O(\log(n)). \tag{C-14}$$

From this we can deduce that $b_n \cdot (1 - \tilde{s}_n) \to +\infty$ (which implies that $\tilde{s}_n \leq 1$) and

$$b_n \cdot (1 - \tilde{s}_n)^2 = O(\log(n)) \cdot (1 - \tilde{s}_n) = O(\log(n)) \cdot O(n^{-3\alpha/2} \sqrt{\log(n)}) \longrightarrow 0 \quad \text{as } n \to \infty.$$

Putting all this together yields, for n large enough,

$$\sum_{k=n}^{\lfloor \overline{q}_n \rfloor - b_n} \pi_n(k) \leq \frac{s_n^{b_n}}{1 - s_n}$$

$$
\begin{aligned}
&\leq \frac{\tilde{s}_n^{b_n}}{1-\tilde{s}_n} \\
&= \frac{\exp\Big(-b_n\cdot(1-\tilde{s}_n)+O(b_n(1-\tilde{s}_n)^2)\Big)}{1-\tilde{s}_n} \\
&\overset{Eq.(C\text{-}14)}{=} \frac{\exp\Big(-\frac{(1-\rho_n)^3 b_n^2}{4\rho_n^4}+O(n^{-\alpha/2}\log(n)^{3/2})\Big)}{1-\tilde{s}_n} \\
&= \frac{n^{-\frac{n^{3\alpha}(1-\rho_n)^3C^2}{4\rho_n^4}}\exp\Big(O(n^{-\alpha/2}\log(n)^{3/2})\Big)}{1-\tilde{s}_n},
\end{aligned}
$$

observe that the exponential term above converges to 1. Also, the denominator is $O(n^{-3\alpha/2}\sqrt{\log(n)})$ while $\frac{n^{3\alpha}(1-\rho_n)^3C^2}{4\rho_n^4}\to\beta^3C^2/4$. So if we choose C such that $\beta^3C^2/4>3\alpha/2$ then we have that

$$
\lim_{n\to\infty}\sum_{k=n}^{\lfloor \bar{q}_n\rfloor - b_n}\pi_n(k)=0,
$$

as desired.

(ii) We show that

$$
\lim_{n\to\infty}\mathbf{P}[Q_n(\infty)<\lfloor \underline{q}_n\rfloor - C\cdot\sqrt{\log(n)}\cdot\sqrt{n}]=0,
$$

the other case is analogous. To reduce notation let $b_n = C\cdot\sqrt{\log(n)\cdot n}$ for some $C>0$ that we identify later then

$$
\begin{aligned}
\mathbf{P}[Q_n(\infty)<\lfloor \underline{q}_n\rfloor - b_n] &= \sum_{k=0}^{\lfloor \underline{q}_n\rfloor - b_n-1}\pi_n(k) \\
&= \sum_{k=0}^{\lfloor \underline{q}_n\rfloor - b_n-1}\prod_{\ell=k+1}^{\lfloor \underline{q}_n\rfloor}\frac{1}{\rho_n}\cdot\frac{\ell}{n}\cdot\frac{1}{\Big(1+\frac{1}{\sqrt{n-\ell}}\Big)} \\
&\overset{(a)}{\leq} \sum_{k=0}^{\lfloor \underline{q}_n\rfloor - b_n-1}\prod_{\ell=k+1}^{\lfloor \underline{q}_n\rfloor-1}\frac{1}{\rho_n}\cdot\frac{\ell}{n}\cdot\frac{1}{\Big(1+\frac{1}{\sqrt{n-\ell}}\Big)} \\
&\overset{(b)}{\leq} \sum_{k=0}^{\lfloor \underline{q}_n\rfloor - b_n-1}\left\{\frac{1}{\lfloor \underline{q}_n\rfloor-k-1}\cdot\sum_{\ell=k+1}^{\lfloor \underline{q}_n\rfloor-1}\frac{1}{\rho_n}\cdot\frac{\ell}{n}\cdot\frac{1}{\Big(1+\frac{1}{\sqrt{n-\ell}}\Big)}\right\}^{\lfloor \underline{q}_n\rfloor-k-1} \\
&\overset{(c)}{\leq} \sum_{k=0}^{\lfloor \underline{q}_n\rfloor - b_n-1}\Bigg\{\underbrace{\frac{1}{b_n}\cdot\sum_{\ell=\lfloor \underline{q}_n\rfloor-b_n}^{\lfloor \underline{q}_n\rfloor-1}\frac{1}{\rho_n}\cdot\frac{\ell}{n}\cdot\frac{1}{\Big(1+\frac{1}{\sqrt{n-\ell}}\Big)}}_{s_{1n}}\Bigg\}^{\lfloor \underline{q}_n\rfloor-k-1} \\
&\leq \frac{s_{1n}^{b_n}}{(1-s_{1n})},
\end{aligned}
$$

where in (a) we use that

$$
\frac{1}{\rho_n}\cdot\frac{\lfloor \underline{q}_n\rfloor}{n}\cdot\frac{1}{\Big(1+\frac{1}{\sqrt{n-\lfloor \underline{q}_n\rfloor}}\Big)}\leq 1,
$$

in (b) the inequality of arithmetic and geometric means, and in (c) the fact that the term we are summing in the second summation is decreasing in ℓ is decreasing for $\ell \leq \underline{q}_n + b_n$. In order to simplify notation let $\tilde{z}_n^2 = n - \lfloor \underline{q}_n \rfloor$. Let us analyze s_{1n},

$$
\begin{aligned}
s_{1n} &= \frac{1}{b_n} \cdot \sum_{\ell=\lfloor \underline{q}_n \rfloor - b_n}^{\lfloor \underline{q}_n \rfloor - 1} \frac{1}{\rho_n} \cdot \frac{\ell}{n} \cdot \frac{1}{\left(1 + \frac{1}{\sqrt{n-\ell}}\right)} \\
&\leq \frac{1}{b_n} \cdot \int_{\lfloor \underline{q}_n \rfloor - b_n}^{\lfloor \underline{q}_n \rfloor} \frac{1}{\rho_n} \cdot \frac{x}{n} \cdot \frac{1}{\left(1 + \frac{1}{\sqrt{n-x}}\right)} dx \\
&= \frac{1}{b_n n \rho_n} \cdot \left[\frac{1}{6} \Big(-3n^2 + n(8\sqrt{n-x} + 6) + 4x\sqrt{n-x} - 12\sqrt{n-x} + 3x^2 - 6x \Big) \right. \\
&\left. - 2(n-1)\log(\sqrt{n-x} + 1) \right] \Bigg|_{\lfloor \underline{q}_n \rfloor - b_n}^{\lfloor \underline{q}_n \rfloor} \\
&= \frac{1}{b_n n \rho_n} \cdot \left[\frac{12n - 4\tilde{z}_n^2 - 12}{6} \Big(\tilde{z}_n - \sqrt{\tilde{z}_n^2 + b_n} \Big) + \frac{4}{6} b_n \sqrt{\tilde{z}_n^2 + b_n} + (n - \tilde{z}_n^2) b_n \right. \\
&\left. - \frac{b_n^2}{2} - 2(n-1)\log\Big(\frac{\tilde{z}_n + 1}{\sqrt{\tilde{z}_n^2 + b_n} + 1} \Big) \right],
\end{aligned}
$$

If we denote this last expression $\tilde{s}_{1n}$ then for $(1 - \tilde{s}_{1n})$ we have that

$$
\begin{aligned}
(1 - \tilde{s}_{1n}) &= 1 - \frac{1}{b_n n \rho_n} \left[\frac{12n - 4\tilde{z}_n^2 - 12}{6} \tilde{z}_n \Big(1 - \sqrt{1 + \frac{b_n}{\tilde{z}_n^2}} \Big) + (n - \tilde{z}_n^2) b_n - \frac{b_n^2}{2} \right] + o\Big(\sqrt{\frac{1}{n \log(n)}} \Big) \\
&= 1 - \frac{1}{b_n n \rho_n} \left[\frac{12n - 4\tilde{z}_n^2 - 12}{6} \Big(-\frac{b_n}{2\tilde{z}_n} + \frac{b_n^2}{8\tilde{z}_n^3} \Big) + (n - \tilde{z}_n^2) b_n - \frac{b_n^2}{2} \right] + o\Big(\sqrt{\frac{1}{n \log(n)}} \Big) \\
&= 1 - \frac{1}{b_n n \rho_n} \left[2n \Big(-\frac{b_n}{2\tilde{z}_n} + \frac{b_n^2}{8\tilde{z}_n^3} \Big) + (n - \tilde{z}_n^2) b_n - \frac{b_n^2}{2} \right] + o\Big(\sqrt{\frac{1}{n \log(n)}} \Big) \\
&\overset{\text{Eq.(C-8)}}{=} 1 - \Big(1 - \frac{\tilde{z}_n^2}{n} \Big) \frac{1}{\rho_n} + \Big(\frac{(1 - \rho_n)}{\rho_n} - \frac{z_n^2}{n \rho_n} \Big) \frac{1}{\rho_n} \cdot \frac{z_n}{\tilde{z}_n} - \frac{b_n}{4\rho_n \tilde{z}_n^3} + \frac{b_n}{2\rho_n n} + o\Big(\sqrt{\frac{1}{n \log(n)}} \Big) \\
&= - \underbrace{\frac{b_n}{4\rho_n \tilde{z}_n^3}}_{O(\sqrt{\frac{\log(n)}{n^{2-3\alpha}}})} + \underbrace{\frac{b_n}{2\rho_n n}}_{O(\sqrt{\frac{\log(n)}{n}})} + o\Big(\sqrt{\frac{1}{n \log(n)}} \Big),
\end{aligned}
$$

hence, $\tilde{s}_n \to 1$ and

$$
b_n \cdot (1 - \tilde{s}_n) = b_n \cdot \Big(-\frac{b_n}{4\rho_n \tilde{z}_n^3} + \frac{b_n}{2\rho_n n} + o\Big(\sqrt{\frac{1}{n \log(n)}} \Big) \Big) = O(\log(n)). \tag{C-15}
$$

From this we can deduce that $b_n \cdot (1 - \tilde{s}_n) \to +\infty$ (which implies that $\tilde{s}_n \leq 1$) and

$$
b_n \cdot (1 - \tilde{s}_n)^2 = O(\log(n)) \cdot (1 - \tilde{s}_n) = O(\log(n)) \cdot O(\sqrt{\frac{\log(n)}{n}}) \longrightarrow 0 \quad \text{as } n \to \infty.
$$

Putting all this together yields, for n large enough

$$\begin{aligned}\sum_{k=0}^{\lfloor q_n \rfloor - b_n - 1} \pi_n(k) &\leq \frac{s_n^{b_n}}{1-s_n} \\ &\leq \frac{\tilde{s}_n^{b_n}}{1-\tilde{s}_n} \\ &= \frac{\exp\Big(-b_n \cdot (1-\tilde{s}_n) + O(b_n(1-\tilde{s}_n)^2)\Big)}{1-\tilde{s}_n} \\ &\overset{Eq.(C\text{-}15)}{=} \frac{\exp\Big(-\Big(-\frac{b_n^2}{4\rho_n \tilde{z}_n^3} + \frac{b_n^2}{2\rho_n n}\Big) + O(\sqrt{\frac{\log(n)^3}{n}})\Big)}{1-\tilde{s}_n} \\ &= \frac{n^{-\left(-\frac{C^2 n}{4\rho_n \tilde{z}_n^3} + \frac{C^2 n}{2\rho_n n}\right)} \exp\Big(O(\sqrt{\frac{\log(n)^3}{n}})\Big)}{1-\tilde{s}_n},\end{aligned}$$

observe that the exponential term above converges to 1. Also, the denominator is $O(\sqrt{\frac{\log(n)}{n}})$ while

$$-\frac{C^2 n}{4\rho_n \tilde{z}_n^3} + \frac{C^2 n}{2\rho_n n} \to \frac{C^2}{2}\Big(1 - \mathbf{1}_{\{\alpha=1/3\}} \cdot \frac{1}{2(\beta \cdot r(\beta))^{3/2}}\Big),$$

where $r(\beta) = \lim_{n\to\infty} r_{0,n}(\rho_n)$, and the term in brackets in the expression above is strictly positive when $\alpha = 1/3$ and $\beta > \beta_1^*$. So if we choose C such that

$$\frac{C^2}{2}\Big(1 - \mathbf{1}_{\{\alpha=1/3\}} \cdot \frac{1}{2(\beta \cdot r(\beta))^{3/2}}\Big) > \frac{1}{2}$$

then we have that

$$\lim_{n\to\infty} \sum_{k=0}^{\lfloor q_n \rfloor - b_n - 1} \pi_n(k) = 0,$$

as desired. $\square$