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

MATHEMATICS OF OPERATIONS RESEARCH
Articles in Advance, pp. 1–30
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Load Balancing Under Strict Compatibility Constraints

Daan Rutten,[a,*] **Debankur Mukherjee**[a]

[a] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332
*Corresponding author

Contact: drutten@gatech.edu, https://orcid.org/0000-0002-4742-4201 (DR); debankur.mukherjee@isye.gatech.edu, https://orcid.org/0000-0003-1678-4893 (DM)

Received: November 29, 2020
Revised: September 1, 2021
Accepted: December 26, 2021
Published Online in Articles in Advance: April 20, 2022

MSC2020 Subject Classification: Primary: 60K25, 60F17; secondary: 60J27, 60G55, 68M20

https://doi.org/10.1287/moor.2022.1258

Abstract. Consider a system with N identical single-server queues and a number of task types, where each server is able to process only a small subset of possible task types. Arriving tasks select $d \geq 2$ random compatible servers and join the shortest queue among them. The compatibility constraints are captured by a fixed bipartite graph between the servers and the task types. When the graph is complete bipartite, the mean-field approximation is accurate. However, such dense compatibility graphs are infeasible for large-scale implementation. We characterize a class of *sparse* compatibility graphs for which the mean-field approximation remains valid. For this, we introduce a novel notion, called *proportional sparsity*, and establish that systems with proportionally sparse compatibility graphs asymptotically match the performance of a fully flexible system. Furthermore, we show that proportionally sparse random compatibility graphs can be constructed, which reduce the server degree almost by a factor $N/\ln(N)$ compared with the complete bipartite compatibility graph.

Funding: This work was supported by the National Science Foundation [Grant CCF-2113027].

Keywords: mean-field limit • power of *d* • stochastic coupling • load balancing on network • data locality • many-server asymptotics • queueing theory

1. Introduction

1.1. Background and Motivation

A canonical model for large-scale systems, such as data centers and cloud networks, consists of a large number of parallel servers with dedicated queues. Tasks arrive into the system sequentially in time and are immediately and irrevocably assigned, using some efficient load-balancing algorithm, to one of these queues, where they wait until executed. Because of ever-increasing heterogeneity in the incoming traffic, these systems typically suffer from stringent task–server compatibility constraints. Indeed, executing a task at a server requires some prestored data, and being able to serve all possible task types comes with an excessive storage capacity requirement (Wang et al. [44], Xie et al. [46]) and an overwhelming implementation complexity (Mishra et al. [20], Reiss et al. [26], Tsitsiklis and Xu [35]). Consequently, full flexibility in task allocation is not a luxury large-scale systems can afford. It is therefore important to understand the performance of load-balancing algorithms under sparser compatibility constraints, where tasks of a particular type can be served only by a relatively small number of servers, naturally viewed as neighbors in a bipartite compatibility graph between the servers and the task types.

The analysis of load-balancing algorithms for large-scale systems dates back to the seminal works by Vvedenskaya et al. [41] and Mitzenmacher [21]. Since then, using mean-field techniques, there has been significant progress in our understanding of the performance of various algorithms. However, many of these heuristics turn out to be false in the presence of compatibility constraints. A widely studied algorithm in this area is the join-the-shortest-queue with d choices (JSQ(d)) or the "power-of-d" scheme, where each arriving task is assigned to the shortest of d randomly selected queues. The JSQ(d) scheme is popular for its low-complexity implementation and excellent delay performance. However, an ill-designed compatibility structure can lead to instability or poor delay performance even for a system operating under the JSQ(d) scheme. In spirit, the nature of this observation is similar to the famous Braess's [2] paradox (Roughgarden and Tardos [27]) in networks. We discuss this example in more detail in Remark 4 below.

The lack of a thorough understanding of large-scale systems with compatibility constraints may be attributed to the scarcity of the theoretical toolbox to analyze such systems. Performance analysis of large-scale systems has flourished in the last three decades because of the abundance of sophisticated mean-field techniques and, in particular, the asymptotic analysis of density-dependent population processes (Ethier and Kurtz [10]). This has provided a firm theoretical basis to analyze these systems. In the presence of arbitrary compatibility constraints, the servers become nonexchangeable, which breaks a core assumption that lies at the foundation of the classical

mean-field framework. Our goal in this paper is to understand the effect of compatibility constraints on the performance of large-scale systems and, in particular, characterize a large class of sparse compatibility graphs that match the performance of a fully flexible system asymptotically in the large-system limit. In doing so, we will also make progress in developing new approaches to analyze such structurally constrained large-scale systems driven by stochastic inputs.

1.2. Our Contributions

Consider a system with N single-server queues and $M(N)$ task types. Tasks of each type arrive as independent Poisson processes of rate $\lambda N/M(N)$ for $\lambda < 1$. Each task requires an exponentially distributed service time with unit mean. We will be looking at a scaling regime where $N \to \infty$. The task–server compatibility is captured in terms of a bipartite graph G_N between the servers and task types; that is, a server i shares an edge in G_N with a task type j, if task type j can be processed by server i. Following the JSQ(d) policy, when a task of type j arrives, d servers that share an edge with j are sampled uniformly at random with replacement, and the task is routed to the shortest of the sampled queues. The quantity of interest is the global occupancy process $\boldsymbol{q}^N(t) = (q_1^N(t), q_2^N(t), \ldots)$, where $q_i^N(t)$ denotes the fraction of servers with queue length at least i at time t in the Nth system. Note that the case when G_N is complete bipartite corresponds to the fully flexible system. Our focus is to identify the sparsest compatibility structures that preserve the performance benefits of a fully flexible system, asymptotically as $N \to \infty$. In other words, we study the *sparsity condition for the compatibility graph, which preserves the validity of the mean-field approximation*.

It is reasonable to guess that well-designed compatibility graphs with a "sufficient" amount of expansion property should preserve the effects of full flexibility, asymptotically, in the large-system limit. However, identifying the right notion of expansion and thereby establishing precise limit laws for the process-level and steady state occupancy processes remains a notoriously challenging problem and has inspired several research works, as we will discuss below. In this work, we attempt to make progress in this direction by developing new approaches to tackle the nonexchangeability. Specifically, our results can be categorized into two groups, as follows.

1.2.1. Arbitrary Deterministic Compatibility Graphs. We start by considering an arbitrary deterministic sequence of graphs $\{G_N\}_{N\geq 1}$, indexed by the number of servers N, and define a novel notion of expansion, which we call *proportional sparsity*; see Definition 1. We show that if the sequence of compatibility graphs is proportionally sparse, then as $N \to \infty$, on any finite time interval, the occupancy process $\boldsymbol{q}^N(\cdot)$ under the JSQ(d) policy converges to the same mean-field limit as the sequence of fully flexible systems. In fact, this process-level limit result extends to a broad class of load-balancing algorithms, for which the assignment decision depends "smoothly" on the empirical queue length distribution of the compatible servers. We call such algorithms *Lipschitz continuous task assignment policies*; see Definition 3. An important step to prove the process-level limit is to show that for almost all dispatchers, the empirical queue length distribution observed in its neighborhood is close to the empirical queue length distribution observed among all servers in the system. This allows us to construct a coupling between the constrained system and the fully flexible system and establish that the ℓ_1-distance between the global occupancy processes in two systems is small uniformly over any finite time interval.

For the interchange of limits and hence the convergence of steady state, two more key ingredients that we need are ergodicity of the prelimit system (for each fixed N) and the tightness of steady states in an appropriate sense. Note that if G_N is not complete bipartite, the occupancy process $\boldsymbol{q}^N(\cdot)$ is no longer Markovian. Consequently, one needs to be careful in defining its time asymptotics and, hence, the interchange of limits. For ergodicity of the underlying Markov process, we need the graph sequence to satisfy a certain *subcriticality condition* that was first introduced in Bramson [3]; see Definition 2. The tightness, however, is technically more challenging. In particular, we need to show that the sequence of steady state occupancy is tight with respect to a certain weighted ℓ_1 norm. For this, we construct a collection of Lyapunov functions, which provide uniform tail bounds on the steady state of the global occupancy process.

Combining the above results, we conclude in Theorem 1 that if a sequence of graphs is proportionally sparse and satisfies the subcriticality condition, then both finite-time dynamics and steady state behavior of the empirical queue length process coincide with that of a fully flexible system, asymptotically as $N \to \infty$. It is worth highlighting that in the above interchange of limits, we do not impose any restrictions on how the number of task types $M(N)$ scales with N. This includes the two popular scenarios where $M(N)$ equals a constant and $M(N) = N$ as special cases.

1.2.2. Random Compatibility Graphs. The results for deterministic graph sequence provide us all the theoretical framework needed to analyze these systems. Next, we exploit these results to construct random sparse compatibility graphs with desired performance benefits. In the context of data-file placement or content replication in

large-scale systems, the degree of a server in the compatibility graph can be thought to be roughly proportional to the storage capacity requirement of that server. It is also considered to be a measure of complexity of the network. To this end, we consider two cases.

First, suppose that the servers are constrained to have degrees exactly equal to $c(N)$. In this case, construct G_N by selecting $c(N)$ task types for each server, independently uniformly at random, without replacement, from the set of all task types. For such a randomly constructed compatibility graph, we establish that the empirical queue length distribution of the system has the same asymptotic law as the fully flexible system, both at the process level and in steady state, if $c(N) \gg M(N)\ln(N)/N$ and $c(N) \gg 1$; see Theorem 2 for details.

Second, we consider a system that allows for inhomogeneous levels of flexibility for different task types. In this case, the compatibility graph is constructed by selecting each edge incident to a task type $w \in W_N$ with probability $p_w(N)$, independently of other edges. Thus, task type w will have an average degree $Np_w(N)$. In this case, we show that the empirical queue length distribution of the system has the same asymptotic law as the fully flexible system, both at the process level and in steady state, if $\min_{w\in W_N} p_w(N)$ and the ℓ_2 norm of the inverse probability vector $(1/p_w(N))_{w\in W_N}$ satisfy suitable growth conditions; see Theorem 3 for details.

To prove the results for random instances, we verify, using concentration of measure arguments, that the graph sequence satisfies both the proportional sparsity and the subcriticality conditions, under the respective growth rate conditions as $N \to \infty$.

1.3. Related Works

The effect of flexibility in the task assignment in large-scale systems was first studied by Turner [36], who considered two types of arriving customers, those that have no routing choice and those that employ the JSQ(d) strategy. It was shown that even a small amount of routing choice can lead to substantial gains in performance through resource pooling. Later, He and Down [15] also considered the diffusion limit of a similar mixed strategy model under heavy traffic.

Relatively recently, there have been a number of works analyzing load-balancing algorithms for large-scale systems, where queues themselves are interconnected by some graph topology. In these models, each queue has an independent, dedicated stream of external arrivals at rate $\lambda < 1$, and each arrival must be assigned instantaneously and irrevocably to one of the neighboring queues, including the one where it first appeared. Although these models are related to the one proposed in this paper, they cannot directly be used to capture the task–server compatibility constraints. In fact, they represent a special case of our model when $M(N) = N$ and there is a perfect matching between task types and servers; see Budhiraja et al. [4, remark 4] for a detailed discussion. Because of this structural difference, the queue length process on an undirected graph is stable for *any* graph, whereas for our model the question of stability is nontrivial by itself. Also, there is a large set of bipartite compatibility graphs that simply cannot be modeled by homogeneous arrivals on an undirected graph. In this line of work, motivated by the bike-sharing network, Gast [13] studies a system of queues connected by a ring topology. To deal with the long-range dependencies among the queue-length processes arising from the restricted graph topology, the work proposes a pair approximation to describe the steady state system. When the ordinary JSQ policy is used at each vertex, that is, when each arriving task joins the shortest of all the neighboring queues, Mukherjee et al. [23] develop a coupling-based approach to establish criteria for asymptotic optimality on fluid and diffusion scale. A key ingredient in their approach is the monotonicity of the system with respect to edge addition; that is, performance of a system gets better, in the sense of stochastic majorization of the occupancy process, if more edges are added to the underlying graph.

The scenario becomes fundamentally more challenging if the system lacks the above monotonicity. One such scenario is when the JSQ(d) policy is considered at each vertex, instead of the JSQ policy, that is, each task is assigned to the shortest queue among the one it first appears and its $d - 1$ randomly selected neighbors. Even a first-order property such as stability is nontrivial in this case. See Remark 4 for a related illustration of this nonmonotonicity. It is this nonmonotonicity that makes the scenario considered in the current article very different from those in the state-of-the-art literature. In fact, this nonmonotonicity hints that *expansion properties that are monotone with respect to edge addition cannot provide sufficient criteria for getting the same asymptotic limit law as the fully flexible system.*

Contemporaneously to the current article, Weng et al. [45] consider the join-the-fastest-shortest-queue (JFSQ) and join-the-fastest-idle-queue (JFIQ) policies for systems with task–server compatibility constraints, where the arrival rates of the task types and the service rates of the servers are heterogeneous. Specifically, Weng et al. [45] obtain finite-system bounds on the mean response time and, generalizing the Lyapunov drift method, shows that under a "well-connected" graph condition, the JFSQ and JFIQ policies can achieve the minimum steady state

response time in both the many-server regime and the sub–Halfin–Whitt regime (when the system load approaches one at a suitable rate), asymptotically as $N \to \infty$.

In a work by Budhiraja et al. [4], sufficient conditions on the graph sequence are obtained, so that the queue length process under the JSQ(d) policy has the same fluid limit on any finite time interval as the complete graph. Their method relies on an asymptotic coupling of the queue length process with an infinite-dimensional McKean–Vlasov process. The asymptotic coupling requires that the system start from a state where the queue lengths are independent and identically distributed, which is a fundamental limitation of their approach. Moreover, their approach does not easily generalize to steady state. Indeed, as mentioned by Budhiraja et al. [4, section 4], even the existence of a time asymptotic limit was not clear for this system. To circumvent these issues, we take a radically different approach in this paper and are able to establish both a process-level limit starting from an arbitrary system state and convergence of the steady state.

In another interesting line of research, Tang and Subramanian [31, 32] analyze a variant of the classical JSQ(d) policy, where the d servers are sampled through d independent nonbacktracking random walks on a high-girth graph. The motivation here is to reduce the amount of randomness used in implementing the classical JSQ(d) policy.

There has been a rich literature in the stability analysis of load-balancing algorithms of finite-sized systems. Related to the output-queued model considered in this paper, the stability analysis dates back to Stolyar [28, 29] and Chernova and Foss [11]. Building on the framework of Chernova and Foss [11], Bramson [3] analyzes stability of JSQ-type systems under a broad class of policies, including the JSQ(d) policy, where the service times and interarrival times follow general distributions. Instead of a bipartite compatibility graph, in this work, there is an independent arrival stream of tasks of rate λ_S corresponding to each subset $S \subseteq [N]$ of servers. Tasks in the arrival stream S join the shortest queue among S. Bramson [3] proposes a sufficient subcriticality condition on the arrival rates λ_S and shows that the system is ergodic under this condition. We will use the above subcriticality condition to establish ergodicity of the system for each fixed N. However, the results of Bramson [3] do not guarantee that the steady state workload in the system scales as $\Theta(N)$ as $N \to \infty$. This is crucial in the large-system limit, because it relates to the tightness of the sequence of steady state occupancy. For this, we use the Lyapunov function approach, as in Wang et al. [42, 43], and establish moment bounds (Hajek [14], Meyn and Tweedie [19]) to obtain uniform bounds on the tail of the stationary occupancy process.

A more recent work, by Cardinaels et al. [5], analyzes stability conditions and obtains performance bounds of a general model for load balancing with affinity relations. In this setup, each arriving task can be routed to either a fast, primary selection of servers or a secondary selection with a slower processing speed. Cruise et al. [8] establish stability for a similar problem where the task–server constraints are modeled as a hypergraph. In the area of redundancy scheduling under compatibility constraints, Cardinaels et al. [6] study the case when each task may only be replicated to a specific set of servers described by a compatibility graph. In the classical heavy-traffic regime (fixed number of servers and load approaches the boundary of the capacity region) and under appropriate conditions on the graph, Cardinaels et al. [6] establish that the system with graph-based redundancy scheduling operates as a multiclass single-server system.

On the scheduling side, there has been significant development in the analysis of multiserver input-queued systems with multiple task types; see Arapostathis et al. [1], Gamarnik and Stolyar [12], Hmedi et al. [16], Tsitsiklis and Xu [33–35], Yekkehkhany and Nagi [47], Yekkehkhany et al. [48], and the references therein. In this area, the work that is closest in spirit to our setup, is by Tsitsiklis and Xu [34, 35]. Here, an input-queued system is considered, where N servers are connected to rN queues by a bipartite compatibility graph, where $r > 0$ is a fixed constant that does not depend on N. Each queue i receives an independent arrival stream of rate λ_i, and tasks remain in the queue until a server becomes available. In this setup, Tsitsiklis and Xu [34, 35] establish that if the average degree of the queues $c(N) \gg \ln(N)$, then there exists a family of expander-graph-based flexibility architectures and a scheduling policy that stabilizes almost all admissible arrival rates and is throughput optimal.

Last, this work also fits into the recent line of work on load balancing for systems with multiple dispatchers (Stolyar [30], Van der Boor et al. [37], Vargaftik et al. [40], Zhou et al. [49]). The analysis in these cases is often more challenging than the classical setup. However, strict task–server compatibility is typically not considered in these works. We refer to Van der Boor et al. [38] for a recent survey on load-balancing algorithms.

1.4. Notation and Organization

The remainder of this paper is organized as follows. In Section 2, we describe the model in detail and introduce notations related to the underlying Markov chain and its state space. Section 3 lists the main theorems and discusses their ramifications. Section 4 provides the proofs of the main theorem involving the deterministic graph sequence. Most other proofs are given in the appendices. In Section 5, we present simulation experiments, both

to support the analytical results and to examine the performance of systems with compatibility constraints that are not analytically tractable. Finally, Section 6 summarizes our results and discusses directions for future research.

A complete bipartite graph with N servers and $M(N)$ dispatchers will be denoted by $K_{N,M}$. We denote by ℓ_1 the normed vector space with norm $\|\boldsymbol{q}\|_1 = \sum_{i=1}^{\infty} |q_i|$. For some positive sequence $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots)$, denote by ℓ_1^{ω} the normed vector space with the weighted-ℓ_1 norm $\|\boldsymbol{q}\|_1^{\omega} = \sum_{i=1}^{\infty} \omega_i |q_i|$. For any set V, $|V|$ denotes its cardinality. We adopt the usual notations $O(\cdot), o(\cdot), \omega(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$ to describe asymptotic comparisons. For two positive deterministic sequences $(f(n))_{n\geq 1}$ and $(g(n))_{n\geq 1}$, we write $f(n) \ll g(n)$ (respectively, $f(n) \gg g(n)$) if $f(n) = o(g(n))$ (respectively, $f(n) = \omega(g(n))$).

2. Model Description

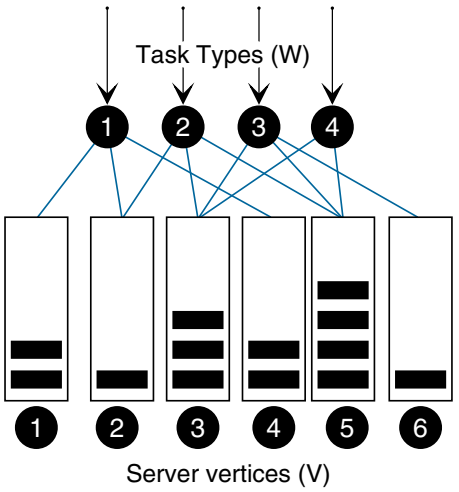
We consider a system that consists of a set of dispatchers W_N and a set of servers V_N, with $|V_N| = N$ and $|W_N| = M(N)$. Each dispatcher handles arrivals of a particular task type. Task type $w \in W_N$ can only be served by a subset $\mathcal{N}_w \subseteq V_N$ of servers. The set of servers $\mathcal{N}_w$ compatible to the task type w can naturally be viewed as neighbors of w in a bipartite graph $G_N = (V_N, W_N, E_N)$ between V_N and W_N, where $E_N \subseteq V_N \times W_N$ is the set of edges. In other words, $w \in W_N$ and $v \in V_N$ share an edge in G_N, if server v has the resources required to process task type w; see Figure 1 for an illustration of the model. Note that the fully flexible system corresponds to G_N being complete bipartite. Each server has a dedicated queue with infinite buffer capacity and operates under a nonidling service discipline that is oblivious to the actual service requirements (e.g., first come, first served). Below we will interchangeably use the terms *task type vertex* and *dispatcher*.

Each dispatcher receives an external stream of arrivals as a Poisson process of rate $\lambda N/M(N)$, independently of the other processes. Thus, the total arrival rate of the system is λN. The service times of the tasks are exponentially distributed with unit mean, independently of each other. For the stability of the system, assume $\lambda < 1$. Each dispatcher employs the JSQ(d) policy; that is, when a task arrives at $w \in W_N$, the dispatcher samples queue lengths at $d \geq 2$ servers uniformly at random with replacement from its neighborhood $\mathcal{N}_w$. The task is assigned to the shortest of the sampled queues.

At time $t \geq 0$, let $I_v^N(t)$ denote the queue length of server $v \in V_N$. Also, let ξ_{vw}^N denote the edge occupancy in G_N; that is, for $v \in V_N$ and $w \in W_N$, $\xi_{vw}^N = 1$ if $(v, w) \in E_N$, and $\xi_{vw}^N = 0$ otherwise. Note that $(I_v^N(\cdot), \xi_{vw} : v \in V_N, w \in W_N)$ is a Markovian state descriptor of the system. Denote this Markov process at time $t \geq 0$ by $\Phi(G_N, t)$ and the state space by $\mathcal{S}_N$. Because, for each system, we keep the graph structure fixed beforehand, we leave the edge occupancy implicit in the state and think of $\Phi(G_N, t)$ as the vector of tagged queue lengths $(I_v^N(\cdot) : v \in V_N)$ and $\mathcal{S}_N$ as $\mathbb{N}^N$.

We introduce a few shorthand notations. For a state $z \in \mathcal{S}_N$, let $\mathcal{X}_i(z) \subseteq V_N$ denote the set of servers with queue lengths exactly equal to i, $X_i(z) := |\mathcal{X}_i(z)|$ and $x_i(z) := X_i(z)/N$. Also, let $\boldsymbol{x}(z) := (x_i(z))_{i\geq 0}$. We refer to $\boldsymbol{x}(z)$ as the *global empirical queue length distribution* (*GEQD*). Furthermore, let $X_i^w(z) := |\mathcal{X}_i(z) \cap \mathcal{N}_w|$ denote the number of servers with queue length exactly equal to i in the neighborhood $\mathcal{N}_w$ of $w \in W_N$, and $x_i^w(z) := X_i^w(z)/|\mathcal{N}_w|$ and $\boldsymbol{x}^w(z) := (x_i^w(z))_{i\geq 0}$. We refer to $\boldsymbol{x}^w(z)$ as the *local empirical queue length distribution* (*LEQD*). The space of (global and

Figure 1. (Color online) A schematic overview of the system with task types W_N, servers V_N, and their compatibility relation.



local) empirical queue length distributions is denoted by

$$\mathcal{X} := \left\{ \boldsymbol{x} \in \ell_1 \Big| \sum_{i=0}^{\infty} x_i = 1 \text{ and } x_i \geq 0 \ \ \forall i \in \mathbb{N} \right\}.$$

For a state $z \in \mathcal{S}_N$, let $\mathcal{Q}_i(z)$ denote the set of servers with queue length *at least* i, $Q_i(z) := |\mathcal{Q}_i(z)|$ and $q_i(z) := Q_i(z)/N$, such that $q_i(z) = \sum_{j=i}^{\infty} x_j(z)$. Also, let $\boldsymbol{q}(z) := (q_i(z))_{i\geq 1}$. We refer to $\boldsymbol{q}(z)$ as the *global occupancy*. Also, for $w \in W_N$, let $Q_i^w(z) := |\mathcal{Q}_i(z) \cap \mathcal{N}_w|$ denote the number of servers with queue length at least i in $\mathcal{N}_w$, $q_i^w(z) := Q_i^w(z)/|\mathcal{N}_w|$, and $\boldsymbol{q}^w(z) := (q_i^w(z))_{i\geq 1}$. We refer to $\boldsymbol{q}^w(z)$ as the *local occupancy*. Local and global occupancy take values in the space

$$\mathcal{Y} := \{\boldsymbol{q} \in \ell_1 | q_i \in [0,1], q_i \geq q_j, i < j, i, j \in \mathbb{N}\}.$$

3. Main Results

3.1. Arbitrary Deterministic Graphs

We start by introducing the notion of proportional sparsity and the subcriticality condition for a deterministic sequence of bipartite graphs $\{G_N\}_{N\geq 1}$ and discuss their ramifications. Proportional sparsity provides a sufficient expansion property of the compatibility graph so that, on any finite time interval, the occupancy process of systems under a broad class of task assignment policies has the same weak limit as the fully flexible system. The subcriticality condition bounds the maximum load on any server and implies that the underlying Markov process is ergodic and the steady state global occupancy is tight in the appropriate space. Together, the proportional sparsity and the subcriticality condition imply the interchange of limits for the global occupancy process.

3.1.1. Proportional Sparsity.

The condition of proportional sparsity requires the edges in the bipartite graph to be fairly distributed in an appropriate sense.

Definition 1 (Proportionally Sparse Graph Sequences). Let $G_N = (V_N, W_N, E_N)$ be a sequence of connected graphs indexed by the number of servers $|V_N| = N$. The sequence $\{G_N\}_{N\geq 1}$ is called proportionally sparse if for each $\varepsilon > 0$,

$$\sup_{U \subseteq V_N} \left| \left\{ w \in W_N \left| \left| \frac{|\mathcal{N}_w \cap U|}{|\mathcal{N}_w|} - \frac{|U|}{N} \right| \geq \varepsilon \right. \right\} \right| / M(N) \to 0, \quad \text{as} \quad N \to \infty. \tag{1}$$

The condition of proportional sparsity ensures that for all but $o(N)$ dispatchers, the local empirical queue length distribution in its neighborhood is close, in a suitable sense, to that of the global empirical queue length distribution. The latter property will be pivotal in establishing the mean-field limit.

Remark 1. From a high level, the class of proportionally sparse graph sequences contains all graphs obtained after *two-step sparsification* of the complete bipartite graph. To see this, note that the complete bipartite graph is the only graph for which $\frac{|\mathcal{N}_w \cap U|}{|\mathcal{N}_w|} = \frac{|U|}{N}$ for all $U \subseteq V_N$ and $w \in W_N$. Now, the first step of sparsification allows for a wiggle room of ε, however small, in the above difference for all U, and the second step allows for $o(N)$-many dispatchers to have the above difference larger than ε. As we will see, after these two steps of sparsification, the class of graph sequences that satisfy this property will be large and will contain graph sequences that are much sparser than the complete bipartite graph.

Remark 2 (Proportional Sparsity and Quasi Randomness). In the dense case when $|\mathcal{N}_w|$ is $\Theta(N)$, the definition of proportional sparsity is related to the notion of quasi-random bipartite graphs. To obtain a random server network, one approach is to construct a random graph and use its structure in the network. An alternative approach is to take a deterministic graph and question whether it is sufficiently "random." For a sequence of these deterministic graphs, it is possible to verify whether it satisfies certain properties that random graphs are expected to have. A sequence of graphs satisfying such properties is called quasi-random. This notion was proposed in a seminal paper by Chung et al. [7] and has subsequently been used in developing numerous algorithmic heuristics. Lemma 1 states that quasi randomness implies proportional sparsity.

Lemma 1. *If $\{G_N\}_{N\geq 1}$ is a quasi-random sequence of graphs, it must be a proportionally sparse graph sequence.*

The proof of Lemma 1 is provided in Appendix D. However, even in the dense case, the *converse of Lemma 1 is not true*. The main reason is the inherent symmetry assumption of quasi randomness, whereas a proportionally

sparse graph sequence can have very inhomogeneous degrees. To see this, we refer to Theorem 3, which states that a broad class of sequences of inhomogeneous random graphs are proportionally sparse.

3.1.2. Subcriticality Condition. We continue by introducing a condition on the maximum load at any server. The subcriticality property will allow us to prove ergodicity of the system and tightness of the steady state occupancy process.

Definition 2 (Subcriticality Condition). Let $G_N = (V_N, W_N, E_N)$ be a graph sequence. The sequence $\{G_N\}_{N\geq 1}$ is said to satisfy the subcriticality condition if for all $N \geq 1$, $w \in W_N$ and $v_1, \ldots, v_d \in V_N$, there exists a probability distribution $\gamma_w^{v_1,\ldots,v_d}(\cdot)$ on V_N supported on $\{v_1, \ldots, v_d\}$ such that

$$\limsup_{N\to\infty} \max_{v\in V_N} \frac{N}{M(N)} \sum_{w\in W_N} |\mathcal{N}_w|^{-d} \sum_{v_1,\ldots,v_d\in\mathcal{N}_w} \gamma_w^{v_1,\ldots,v_d}(v) \leq 1. \tag{2}$$

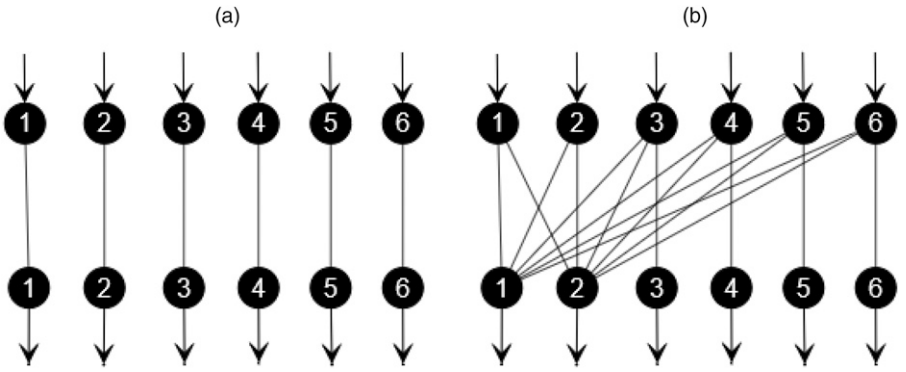
The probability distribution $\gamma_w^{v_1,\ldots,v_d}(v)$ can be interpreted as a *static* randomized task assignment policy for tasks arriving at dispatcher $w \in W_N$, when the servers $v_1, \ldots, v_d \in \mathcal{N}_w$ are selected as the d chosen servers. To this end, to understand the subcriticality condition intuitively, think of a new system where the task allocation is done as follows: when a task arrives, a dispatcher $w \in W_N$ is selected uniformly at random. Then, similar to the JSQ(d) policy, the servers $v_1, \ldots, v_d$ are sampled from the neighborhood $\mathcal{N}_w$ of w with replacement. The task is then routed to a server $v \in \{v_1, \ldots, v_d\}$ with probability $\gamma_w^{v_1,\ldots,v_d}(v)$. The subcriticality condition requires that, for any $\lambda < 1$, the load received by any server under this static task assignment policy $\gamma_w^{v_1,\ldots,v_d}(\cdot)$ is less than one. It is thus sufficient for stability to find one such static task assignment policy, depending on the underlying graph, satisfying this condition.

Remark 3. As mentioned in the introduction, Bramson [3] and, more recently, Cardinaels et al. [5] analyzed the stability of JSQ-type policies in a related framework. The subcriticality condition (2) is equivalent to the those stated in Bramson [3] and Cardinaels et al. [5], although our condition is stated for a sequence of graphs, later to be used for tightness of the steady state occupancy process, and the latter conditions are stated for a fixed system. As noted in Bramson [3, p. 1571], the subcriticality condition as stated in Definition 2 is *also a necessary condition* for stability.

Remark 4 (Nonmonotonicity). The satisfiability of neither the proportional sparsity property nor the subcriticality condition is monotone with respect to the addition of edges in the compatibility graph. As an example, consider the graph in which each dispatcher is perfectly matched to exactly one server, as in Figure 2(a). This graph satisfies the subcriticality condition. Now, alternatively, consider the graph in Figure 2(b). The latter graph contains all the edges of the graph in Figure 2(a), yet it is not hard to verify that it does not satisfy the subcriticality condition for $d = 2$. This is because for any task arriving at dispatchers 3, 4, 5, and 6, it will be assigned to either server 1 or 2 with probability 4/9. This makes the load on servers 1 and 2 higher than one, under any static task assignment policy. Similarly, the notion of proportional sparsity is also not monotone in the addition of edges. For example, adding a disproportionate number of edges between only a subset $U \subset V_N$ and the dispatchers W_N will invalidate proportional sparsity with respect to that set U.

Now we have all the ingredients to state the main result for deterministic sequences of compatibility graphs.

Figure 2. An example illustrating that increasing flexibility may not always lead to better performance.



Theorem 1. *Let $\{G_N\}_{N\geq 1}$ be a proportionally sparse graph sequence. Then, on any finite time interval $[0,T]$, the scaled occupancy process $\boldsymbol{q}(\Phi(G_N,t)) = (q_1(\Phi(G_N,t)), q_2(\Phi(G_N,t)),\dots)$ converges weakly with respect to the Skorohod-J_1 topology, as $N\to\infty$, to the process $\boldsymbol{q}^*(t) = (q_1^*(t), q_2^*(t),\dots)$, given by the unique solution of the system of ordinary differential equations (ODEs)*

$$\frac{dq_i^*(t)}{dt} = \lambda(q_{i-1}^*(t)^d - q_i^*(t)^d) - (q_i^*(t) - q_{i+1}^*(t)), \quad \text{for } i = 1,2,\dots, \tag{3}$$

provided $\boldsymbol{q}^*(0)\in\mathcal{Y}$ and $\|\boldsymbol{q}(\Phi(G_N,0)) - \boldsymbol{q}^*(0)\|_1 \to 0$ as $N\to\infty$.

Moreover, if the graph sequence $\{G_N\}_{N\geq 1}$ satisfies the subcriticality condition, then for each fixed N, the Markov process is ergodic and $\boldsymbol{q}(\Phi(G_N,\infty))$ converges weakly to $\boldsymbol{q}^*(\infty)$ as $N\to\infty$, where $\Phi(G_N,\infty)$ is a random variable distributed as the steady state of $(\Phi(G_N,t))_{t\geq 0}$ and

$$q_i^*(\infty) = \lambda^{\frac{d^i-1}{d-1}}, \quad \text{for } i = 1,2,\dots. \tag{4}$$

The proof of Theorem 1 is provided in Section 4.4.

Remark 5. As mentioned before, the process-level convergence result in Theorem 1 holds for a much broader class of assignment policies, when the assignment decision depends *smoothly* on the LEQD. We call this class the *Lipschitz continuous* task assignment policies, which is introduced in Definition 3 in Section 4.1.

Remark 6. The system of ODEs in (3) can be recognized as the mean-field limit of the classical JSQ(d) policy (Mitzenmacher [21], Vvedenskaya et al. [41]). Thus, Theorem 1 extends the validity of mean-field approximation for the class of proportionally sparse graph sequences that satisfy the subcriticality condition. In other words, as we will see in the next section, Theorem 1 shows that the performance benefits of the fully flexible system can be preserved while making the compatibility graph significantly sparser.

3.2. Randomly Designed Compatibility Graphs

We have seen two sufficient conditions on deterministic graph sequences to establish asymptotically equivalence of the JSQ(d) policy under limited and full flexibility in task allocation. Given a graph sequence, the conditions can be verified. This section will provide two simple ways of constructing a random compatibility graph, both satisfying the two conditions almost surely, in the large-system limit. Note that although the graph is random in the two cases below, once constructed, it remains fixed for the system. That is, in the terminology of random processes in a random environment, we obtain a quenched limit theorem. Below, the almost sure statements involving the sequence of random graphs $\{G_N\}_{N\geq 1}$ are with respect to any probability measure $\mathbb{P}_0$ on $\prod_N\{0,1\}^{N\times M(N)}$ such that its projection on $\{0,1\}^{N\times M(N)}$ corresponds to the distribution of G_N.

3.2.1. Hard Constraint on Server Degrees. As mentioned before, the degree of the servers in the compatibility graph is an important measure of sparsity, as it is roughly proportional to the required storage capacity. For that reason, we will first consider the case when all the servers have degree exactly equal to some fixed number $c(N)$, that is much smaller than $M(N)$, the server degree for a fully flexible system.

Theorem 2. *Let $c(N)\leq M(N)$ be a sequence of positive integers satisfying*

$$c(N)\to\infty \quad \text{and} \quad \frac{Nc(N)}{M(N)\ln(N)}\to\infty, \quad \text{as} \quad N\to\infty.$$

Also, construct G_N as follows: for each $v\in V_N$, select $c(N)$ edges from $\{(v,w)\in V_N\times W_N | v\in V_N\}$ uniformly at random, without replacement. Then the sequence $\{G_N\}_{N\geq 1}$ is proportionally sparse and satisfies the subcriticality condition, almost surely. Consequently, the conclusions of Theorem 1 hold.

The proof of Theorem 2 is technically involved. It uses concentration of measure arguments repeatedly to establish structural properties of the compatibility graphs, and it is provided in Appendix A.

Remark 7. Observe that Theorem 2 guarantees that the validity of the mean-field approximation can be retained asymptotically, by uniformly reducing the server degrees by almost a factor of $N/\ln(N)$, compared with the fully flexible system where the degree of each server is $M(N)$. Also, note that if $M(N) = O(N/\ln(N))$, then the convergence results in Theorem 2 hold for *any growth rate* of $c(N)$.

3.2.2. Inhomogeneous Levels of Flexibility. Next, we consider a system that allows for inhomogeneous levels of flexibility for different task types. In particular, the compatibility graph is constructed by selecting each edge incident to a task type $w \in W_N$ with probability $p_w(N)$, independently of other edges. Thus, in expectation, task type w has the flexibility to be assigned to $Np_w(N)$ possible servers. Theorem 3 provides a set of sufficient conditions on $(p_w(N))_{w\in W_N}$ to ensure that $\{G_N\}_{N\geq 1}$ satisfies the proportional sparsity and the subcriticality conditions.

Theorem 3. *Assume that* $(p_w(N))_{w\in W_N}$ *satisfies the following:*

$$\begin{aligned} &1. \quad \frac{N}{\ln(M(N)) + \ln(N)} \min_{w\in W_N} p_w(N) \to \infty, \\ &2. \quad M(N) \min_{w\in W_N} p_w(N) \to \infty, \text{ and} \\ &3. \quad \frac{\ln(N)}{(M(N))^2} \sum_{w\in W_N} \frac{1}{(p_w(N))^2} \to 0, \end{aligned} \tag{5}$$

as $N \to \infty$. *Also, construct* G_N *as follows: for any* $v \in V_N, w \in W_N$, *select edge* $(v,w) \in E_N$ *independently with probability* $p_w(N)$. *Then the sequence* $\{G_N\}_{N\geq 1}$ *is proportionally sparse and satisfies the subcriticality condition, almost surely. Consequently, the conclusions of Theorem 1 hold.*

As in the proof of Theorem 2, the proof of Theorem 3 also uses concentration of measure arguments, and it is provided in Appendix B.

Remark 8. If $p_w(N) = p(N)$ for all $w \in W_N$, then it is possible to relax condition 3 in Equation (5) on the edge probabilities to $p(N) = \omega(\ln(N)/M(N))$. For the special case $M(N) = N$, if $p(N) = o(\ln(N)/N)$, then the compatibility graph constructed as above will leave at least one dispatcher isolated with probability tending to one as N tends to infinity (see, e.g., Van der Hofstad [39]). This means the graph sequence cannot satisfy the subcriticality condition, and this growth rate condition for $p(N)$ is nearly the optimum.

4. Proofs

To prove Theorem 1, we will follow the usual interchange of limits argument. In particular, the proof consists of three key steps. First, in Section 4.1, we show that if the graph sequence is proportionally sparse, then the scaled occupancy process converges weakly to the appropriate system of ODEs, on any finite time interval. Second, in Section 4.2, we show that if a graph sequence satisfies the subcriticality condition, then for any fixed N, the system is ergodic, and the sequence of steady state global occupancy is tight in the appropriate topology. Third, in Section 4.3, we prove the global stability of the limiting system of ODEs with (4) being its fixed point. Combining the above three steps, we complete the proof of Theorem 1 in Section 4.4.

4.1. Process-Level Convergence

In this section, we will prove the process-level convergence of the occupancy process under a general class of task assignment policies. We start by specifying the class of assignment policies.

4.1.1. Task Assignment Policies. To determine which server to assign an incoming task to, each dispatcher in W_N follows a generic task assignment policy Π that works as follows. We identify the policy Π with an *assignment probability function* $\boldsymbol{p}^\Pi = (p_0^\Pi, p_1^\Pi, \dots) : \mathcal{X} \to [0,1]^\infty$. When a task arrives at a dispatcher $w \in W_N$ with LEQD $\boldsymbol{x}^w$, do the following:

a. Select a random queue length I distributed as the probability measure induced by the assignment probability function evaluated at the LEQD $\boldsymbol{x}^w$, that is, $\mathbb{P}(I = i) = p_i^\Pi(\boldsymbol{x}^w)$ for $i = 0, 1, \dots$. The random variable I is independent of any other processes and also independent across different arrival epochs.

b. Next, a server is selected uniformly at random among all servers with queue length I.

Note that the above generic task assignment policy has a number of features. For example, the assignment policy may depend only on the LEQD, and the dispatcher does not distinguish between two neighboring servers with the same queue length; that is, the assignment policy distinguishes servers based only on their queue lengths and not, for instance, the number of compatible task types of a server. It is not hard to see that the JSQ(d) policy can also be described in the above form. We restrict our analysis to Lipschitz continuous task assignment policies as given by the next definition.

Definition 3 (Lipschitz Continuous Task Assignment Policy). A policy Π is said to be Lipschitz continuous if there exists a finite positive constant K such that its assignment probability function satisfies the following. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,

$$\sum_{i=0}^{\infty} |p_i^{\Pi}(\boldsymbol{x}) - p_i^{\Pi}(\boldsymbol{y})| \leq K \sum_{i=0}^{\infty} |x_i - y_i|. \tag{6}$$

The Lipschitz continuity bounds the sensitivity of the assignment probability function. In other words, if the ℓ_1-distance between the LEQDs is small, then the probabilities of routing a task to a server with a given queue length from these states should also be close. As we will see in Lemma 2, the JSQ(d) policy is Lipschitz continuous for any fixed $d \geq 1$. An example of a policy that is *not* Lipschitz continuous is the ordinary JSQ policy, where the addition of a single task to the system, causing a change of $\Theta(1/N)$ in the LEQD, can change an assignment probability from zero to one.

Lemma 2. *For any fixed $d \geq 1$, the JSQ(d) policy is Lipschitz continuous with Lipschitz constant $2d! \times d^2$.*

The proof of Lemma 2 is provided in Appendix E. We now state the process-level convergence theorem.

Theorem 4. *Let $\{G_N\}_{N\geq 1}$ be a proportionally sparse graph sequence and Π be a Lipschitz continuous policy with assignment probability function $\boldsymbol{p}^{\Pi} = (p_0^{\Pi}, p_1^{\Pi}, \dots)$. Then, on any finite time interval $[0,T]$, the scaled occupancy process $\boldsymbol{q}(\Phi(G_N,t)) = (q_1(\Phi(G_N,t)), q_2(\Phi(G_N,t)), \dots)$ converges weakly with respect to the Skorohod-J_1 topology, as $N \to \infty$, to the process $\boldsymbol{q}^*(t) = (q_1^*(t), q_2^*(t), \dots)$, given by the unique solution of the system of ODEs*

$$\frac{dq_i^*(t)}{dt} = \lambda p_{i-1}^{\Pi}((q_j^*(t) - q_{j+1}^*(t))_{j\geq 0}) - (q_i^*(t) - q_{i+1}^*(t)) \quad \text{for} \quad i = 1, 2, \dots, \tag{7}$$

provided $\boldsymbol{q}^*(0) \in \mathcal{Y}$ and $\|\boldsymbol{q}(\Phi(G_N,0)) - \boldsymbol{q}^*(0)\|_1 \to 0$ as $N \to \infty$.

The rest of this section is devoted to the proof of Theorem 4.

Remark 9. The unique solvability of the infinite set of ODEs in Equation (7) follows from the Lipschitz property of the policy Π, using standard results in analysis (see, e.g., Deimling [9, theorem 3.2]).

There are two main ingredients to the proof of Theorem 4. First, in Proposition 1, we establish that for *almost* all dispatchers, the LEQD is close to the GEQD, uniformly over any finite time interval. Second, we couple the Nth system with graph structure G_N to a fully flexible system with the complete bipartite graph and establish in Proposition 2 a criterion for the two systems to behave similarly. Finally, we complete the proof of Theorem 4 using Propositions 1 and 2 and the Lipschitz continuity of the task assignment policy.

4.1.2. Proximity of Local and Global Empirical Queue Length Distributions.

We begin by introducing the notion of *good* and *bad* dispatchers. Loosely speaking, a good dispatcher is a dispatcher for which the LEQD is close to the GEQD.

Definition 4 (ε-Good Dispatchers). When the system is in state $z \in \mathcal{S}_N$, the dispatcher $w \in W_N$ is called ε-good if

$$\sum_{i=0}^{\infty} |x_i(z) - x_i^w(z)| < \varepsilon. \tag{8}$$

Also, a dispatcher is called ε-bad if it is not ε-good.

In the following, let $B_N^{\varepsilon}(t)$ denote the number of ε-bad dispatchers at time t, when the system is in state $\Phi(G_N,t)$.

Proposition 1. *Let $\{G_N\}_{N\geq 1}$ be a proportionally sparse graph sequence. Then for each $\varepsilon, \delta > 0$,*

$$\mathbb{P}\left(\sup_{t\in[0,T]} B_N^{\varepsilon}(t) \geq \delta M(N) \right) \to 0 \quad \text{as} \quad N \to \infty, \tag{9}$$

provided $\boldsymbol{q}^*(0) \in \mathcal{Y}$ and $\|\boldsymbol{q}(\Phi(G_N,0)) - \boldsymbol{q}^*(0)\|_1 \to 0$ as $N \to \infty$.

The key idea in the proof of Proposition 1 is to observe that the servers with queue length i at time t form a subset $U_i(t) \subseteq V_N$ and the proportional sparsity of the compatibility graph implies that *for almost all $w \in W_N$*, the fraction of neighbors within $U_i(t)$ is close to $|U_i(t)|/N$. However, we need to deal with some technical challenges.

For example, the subset of dispatchers for which the above does *not* hold may depend on i, and thus, one needs to be careful in estimating the number of ε-bad dispatchers. Also, an uniformity over $[0,T]$ needs to be established. The complete proof of Proposition 1 is provided in Appendix F.

Remark 10. It is worthwhile to highlight that Proposition 1 requires only the proportional sparsity property of the graph sequence and does *not* depend on the task assignment policy or even on the dynamics in any way. This makes the applicability of this method much broader, in analyzing certain structurally constrained large-scale dynamical systems where the process running at each vertex (queue length in our case) takes countably many values.

4.1.3. A Coupling Construction. Next, for any $N \geq 1$, we couple the queueing system on an arbitrary bipartite graph G_N with the fully flexible system, that is, corresponding to the complete bipartite graph $K_{N,M}$.

The coupling approach has been highly successful in proving large-system limit theorems (Mukherjee et al. [23, 24]). However, the biggest issue in constructing an appropriate coupling in the presence of compatibility constraints lies in the fact that if the arrivals are synchronized at each dispatcher $w \in W_N$, in two systems with different compatibility graphs, the set of neighbors of w becomes different in the two systems. Consequently, one cannot synchronize which server a task will be assigned to, and the coupling of the queues breaks down. We will now introduce a novel coupling, called *optimal coupling*, to tackle this issue.

In short, we will refer to the two systems as the G_N-system and $K_{N,M}$-system, respectively. Both systems employ a Lipschitz continuous task assignment policy Π. To describe the coupling, first, in each of the two systems, order the servers by nondecreasing queue lengths, breaking ties arbitrarily. We then couple the departure and arrivals in both systems as follows.

• **Departures.** Synchronize the departure epochs of the kth ordered servers in the two systems, that is, both systems will potentially finish serving a task at the kth ordered server at the same epoch, whenever they are nonempty for $k = 1, 2, \ldots, N$.

• **Arrivals.** Synchronize the arrival epochs of task type w in both systems, for all $w \in W_N$. At an arrival epoch of w, let $\boldsymbol{x}^w$ and $\boldsymbol{y}^w$ be the LEQDs for the G_N-system and the $K_{N,M}$-system, respectively. Note that $\boldsymbol{y}^w$ is also the GEQD for the $K_{N,M}$-system. Define $p_i = \min(p_i^{\Pi}(\boldsymbol{x}^w), p_i^{\Pi}(\boldsymbol{y}^w))$ for $i = 0, 1, \ldots$. Now, let us draw a Uniform$[0,1]$ random variable, independently of any other processes, and denote it by U. Recall the description of the task assignment policy Π from Section 4.1. The value of U will be used in both systems to generate the random variables I_1 and I_2, for the G_N-system and $K_{N,M}$-system, respectively. In the G_N-system, set $I_1 = i$, for $i = 0, 1, \ldots$, if

$$U \in \left[\sum_{j=0}^{i-1} p_j, \sum_{j=0}^{i} p_j\right) \bigcup \left[\sum_{j=0}^{\infty} p_j + \sum_{j=0}^{i-1} (p_j^{\Pi}(\boldsymbol{x}^w) - p_j), \sum_{j=0}^{\infty} p_j + \sum_{j=0}^{i} (p_j^{\Pi}(\boldsymbol{x}^w) - p_j)\right), \tag{10}$$

and assign the arriving task to a server selected uniformly at random among all servers with queue length I_1. Similarly, in the $K_{N,M}$-system, set $I_2 = i$, for $i = 0, 1, \ldots$, if

$$U \in \left[\sum_{j=0}^{i-1} p_j, \sum_{j=0}^{i} p_j\right) \bigcup \left[\sum_{j=0}^{\infty} p_j + \sum_{j=0}^{i-1} (p_j^{\Pi}(\boldsymbol{y}^w) - p_j), \sum_{j=0}^{\infty} p_j + \sum_{j=0}^{i} (p_j^{\Pi}(\boldsymbol{y}^w) - p_j)\right), \tag{11}$$

and assign the arriving task to a server selected uniformly at random among all servers with queue length I_2.

Note that the coupling preserves the marginal laws of both systems. Next, we introduce a notion that facilitates comparison of the performance of the two systems on suitable asymptotic scales.

Definition 5 (Mismatch in Queue Length). At an arrival epoch, the two coupled systems are said to mismatch in queue length if $I_1 \neq I_2$, that is, the arriving tasks are assigned to two servers of different queue lengths in the two systems. Denote by $\Delta_N(t)$ the cumulative number of times the systems mismatch in queue length up to time t.

The occupancy process and the cumulative number of mismatches in queue length are related as stated in Proposition 2. The proof of Proposition 2 is provided in Appendix G.

Proposition 2. *For any $N \geq 1$, consider the G_N-system and the $K_{N,M}$-system coupled as above. Then the following holds almost surely on the coupled probability space: for all $t \geq 0$,*

$$\sum_{i=1}^{\infty} |Q_i(\Phi(K_{N,M}, t)) - Q_i(\Phi(G_N, t))| \leq 2\Delta_N(t), \tag{12}$$

provided the inequality holds at $t = 0$.

Remark 11. The statement of Proposition 2 is similar, in spirit, to proposition 5 of Mukherjee et al. [23] and proposition 4 of Mukherjee et al. [24]. Our definition of mismatch in queue length counts the number of times a task is routed to servers with different queue lengths. In contrast, Mukherjee et al. [23, 24] order the servers by queue length, and their definition of *differing in decision* counts the number of times a task is routed to servers with different order statistics. Our notion of mismatch in queue length enables us to compare the occupancy processes of two different systems based on their LEQDs only, even when the individual queues are not coupled.

4.1.4. Proof of Theorem 4. First, it is fairly standard (see, e.g., Kurtz [18, chapter 8]) to show that $\boldsymbol{q}(\Phi(K_{N,M},t))$ converges weakly to $\boldsymbol{q}^*(t)$. Therefore, by Proposition 2, it is enough to prove that for any $\varepsilon'>0$ and $\delta'>0$, there exists $N_0\geq 1$ such that

$$\mathbb{P}\left(\sup_{t\in[0,T]}\Delta_N(t)/N\geq\varepsilon'\right)<\delta' \text{ for all } N\geq N_0. \tag{13}$$

Fix any $\varepsilon>0$, to be chosen later. Recall Definition 4 and let $A_N^\varepsilon(t)$ and $B_N^\varepsilon(t)$ denote the number of ε-good and ε-bad dispatchers in the G_N-system, respectively. Couple the G_N-system to the $K_{N,M}$-system by the optimal coupling described in Section 4.1.3. For brevity, let $\boldsymbol{x}(t):=\boldsymbol{x}(\Phi(G_N,t))$ and $\boldsymbol{x}^w(t):=\boldsymbol{x}^w(\Phi(G_N,t))$ for the G_N-system and $\boldsymbol{y}(t):=\boldsymbol{x}(\Phi(K_{N,M},t))$ for the $K_{N,M}$-system. Also, define $\rho_N(t):=\sum_{i=0}^\infty|y_i(t)-x_i(t)|$. At an arrival epoch $t\geq 0$, if a task arrives at an ε-good dispatcher $w\in W_N$, then

$$\sum_{i=0}^\infty|y_i(t-)-x_i^w(t-)|\leq\sum_{i=0}^\infty|y_i(t-)-x_i(t-)|+\sum_{i=0}^\infty|x_i(t-)-x_i^w(t-)|\leq\rho_N(t-)+\varepsilon. \tag{14}$$

Define the uniform random variable U and p_i as in the optimal coupling, and observe that the probability that the systems mismatch in queue length at such an arrival epoch is bounded by

$$\begin{aligned}&\mathbb{P}\left(U\notin\left[0,\sum_{i=0}^\infty p_i\right)\right)=1-\sum_{i=0}^\infty p_i=\sum_{i=0}^\infty(p_i^\Pi(\boldsymbol{y}(t-))-p_i)\\&\leq\sum_{i=0}^\infty|p_i^\Pi(\boldsymbol{y}(t-))-p_i^\Pi(\boldsymbol{x}^w(t-))|\leq K\sum_{i=0}^\infty|y_i(t-)-x_i^w(t-)|\leq K(\rho_N(t-)+\varepsilon),\end{aligned} \tag{15}$$

by the Lipschitz property of Π in (6). If instead a task arrives at an ε-bad dispatcher, then with probability at most one, the systems mismatch in queue length. Now, because at the arrival epochs, the random variables U are independent of any other processes, we can construct an independent unit-rate Poisson process $(Z(t))_{t\geq 0}$, so that $\Delta_N(t)$ is upper bounded by a random time change (cf. Pang et al. [25, section 2.1]) of Z as follows: for all $t\in[0,T]$,

$$\Delta_N(t)\leq Z\left(\frac{\lambda N}{M(N)}\int_0^t[A_N^\varepsilon(s-)\cdot K(\rho_N(s-)+\varepsilon)+B_N^\varepsilon(s-)\cdot 1]\mathrm{d}s\right). \tag{16}$$

Now observe that

$$\begin{aligned}\rho_N(t)&=\sum_{i=0}^\infty|x_i(\Phi(K_{N,M},t)-x_i(\Phi(G_N,t))|\\&=\sum_{i=0}^\infty|q_i(\Phi(K_{N,M},t))-q_{i+1}(\Phi(K_{N,M},t))-q_i(\Phi(G_N,t))+q_{i+1}(\Phi(G_N,t))|\\&\leq 2\sum_{i=0}^\infty|q_i(\Phi(K_{N,M},t))-q_i(\Phi(G_N,t))|\leq\frac{4\Delta_N(t)}{N},\end{aligned} \tag{17}$$

where the last inequality follows from Proposition 2. Furthermore, using Proposition 1 to bound B_N^ε on the right-hand side of (16), we can write, for any fixed $\delta>0$ to be chosen later, that there exists $N_0=N_0(\delta)\in\mathbb{N}$ such that for all $N\geq N_0$,

$$\mathbb{P}\left(\sup_{t\in[0,T]}B_N^\varepsilon(t)>\delta M(N)\right)\leq\frac{\delta}{2}. \tag{18}$$

Therefore, (17), (18), and an application of Tonelli's theorem imply that for all $N\geq N_0$ and $t\in[0,T]$,

$$\mathbb{E}\left(\frac{\Delta_N(t)}{N}\right)\leq\lambda\int_0^t\left[K\left(4\mathbb{E}\left(\frac{\Delta_N(s-)}{N}\right)+\varepsilon\right)+\frac{3\delta}{2}\right]\mathrm{d}s. \tag{19}$$

By applying Grönwall's inequality to (19), we obtain,

$$\mathbb{E}\left(\frac{\Delta_N(t)}{N}\right) \le \lambda(K\varepsilon + 3\delta/2)t\exp(4\lambda K t). \tag{20}$$

Finally, because $\Delta_N(t)$ is nonnegative and nondecreasing in t, using Markov's inequality, we write

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in[0,T]}\frac{\Delta_N(t)}{N} \ge \varepsilon'\right) &\le \mathbb{P}\left(\frac{\Delta_N(T)}{N} \ge \varepsilon'\right) \le \frac{1}{\varepsilon'}\mathbb{E}\left(\frac{\Delta_N(T)}{N}\right)\\ &\le \frac{\lambda}{\varepsilon'}(K\varepsilon + 3\delta/2)T\exp(4\lambda K T),\end{aligned} \tag{21}$$

which yields Equation (13) by choosing ε and δ small enough. □

4.2. Stability and Tightness

In this section, we will prove positive recurrence of the Markov process $\Phi(G_N,\cdot)$ and tightness of the steady state occupancy process, as stated in the next theorem.

Theorem 5. *Let $\{G_N\}_{N\ge1}$ be a sequence of graphs satisfying the subcriticality condition. Then the following hold:*

i. *There exists $N_0 \ge 1$ such that for each $N \ge N_0$, the Markov process $\Phi(G_N,\cdot)$ under the JSQ(d) policy with $d \ge 1$ is positive Harris recurrent, and thus it has a unique steady state.*

ii. *Let $\Phi(G_N,\infty)$ denote a random variable distributed as the steady state of $\Phi(G_N,\cdot)$. Fix any $r \in (1, 2/(1+\lambda))$ and a positive sequence $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots)$ satisfying*

$$\omega_{i_0+i} = \omega_{i_0} r^i \text{ for } i \ge 1 \tag{22}$$

for some $i_0 \in \mathbb{N}$. Then the sequence $\{\boldsymbol{q}(\Phi(G_N,\infty))\}_{N\ge1}$ is tight when $\mathcal{Y}$ is endowed with the ℓ_1^ω-topology.

The positive Harris recurrence from part i of Theorem 5 follows from known techniques in Bramson [3] and Cardinaels et al. [5]. However, as we prove the tightness in part ii, the positive recurrence follows as a side result. This is given in Appendix H. In the rest of this section, we complete the proof of tightness. First, we obtain a bound on the tail of the expected global occupancy of the stationary state.

Lemma 3. *Consider a sequence $\{G_N\}_{N\ge1}$ that satisfies the subcriticality condition. There exists $N_0 \ge 1$ such that for all $N \ge N_0$ and $k \ge 1$,*

$$\sum_{i=k}^{\infty}\mathbb{E}[q_i(\Phi(G_N,\infty))] \le \frac{(1+\lambda)/2}{1-(1+\lambda)/2}\mathbb{E}[q_{k-1}(\Phi(G_N,\infty))]. \tag{23}$$

The proof of Lemma 3 is provided in Appendix H. It uses a sequence $(V_k)_{k\ge1}$ of Lyapunov functions, where bounds on the drift of V_k result in a moment bound on the tail sum of $\boldsymbol{q}(\Phi(G_N,\infty))$ starting from k. The next technical lemma states sufficient criteria for tightness with respect to the ℓ_1^ω-topology, and it is proved in Appendix I.

Lemma 4. *Fix a positive sequence $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots)$. The sequence of random variables $\{\boldsymbol{q}(\Phi(G_N,\infty))\}_{N\ge1}$ is tight when $\mathcal{Y}$ is endowed with the ℓ_1^ω-topology, if for each fixed $i \in \mathbb{N}$, $\{\omega_i q_i(\Phi(G_N,\infty))\}_{N\ge1}$ is tight in $\mathbb{R}$ and*

$$\lim_{j\to\infty}\limsup_{N\to\infty}\mathbb{P}\left(\sum_{i=j}^{\infty}\omega_i q_i(\Phi(G_N,\infty)) > \varepsilon\right) = 0. \tag{24}$$

Proof of Theorem 5ii. We will verify the sufficient conditions stated in Lemma 4. Because $\omega_i q_i(\Phi(G_N,\infty)) \in [0,\omega_i]$, it is trivially tight in $\mathbb{R}$. Now, by Markov's inequality,

$$\mathbb{P}\left(\sum_{i=j}^{\infty}\omega_i q_i(\Phi(G_N,\infty)) > \varepsilon\right) \le \frac{1}{\varepsilon}\mathbb{E}\left[\sum_{i=j}^{\infty}\omega_i q_i(\Phi(G_N,\infty))\right] = \frac{1}{\varepsilon}\sum_{i=j}^{\infty}\omega_i\mathbb{E}\left[q_i(\Phi(G_N,\infty))\right]. \tag{25}$$

Fix any $r \in (1, 2/(1+\lambda))$. In the following, let $j \ge i_0$ be such that $\omega_i = (\omega_{i_0}/r^{i_0})r^i$ for $i \ge j$. Recall from Lemma 3 that $\{\mathbb{E}[q_i(\Phi(G_N,\infty))]\}_{i\ge1}$ satisfies (23). We will bound the right-hand side of Equation (25) by maximizing its value over all sequences in ℓ_1 satisfying Equation (23). That is, we arrive at the maximization problem, for $j = 1, 2, \dots,$

$$\textsc{Sol}(j) := (\omega_{i_0}/r^{i_0})\sup_{\boldsymbol{a}\in\ell_1}\sum_{i=j}^{\infty}r^i a_i, \text{ where } \sum_{i=k}^{\infty}a_i \le \frac{(1+\lambda)/2}{1-(1+\lambda)/2}a_{k-1} \text{ for all } k \ge 1, \tag{26}$$

and its finite projection,

$$\textsc{Sol}_n(j) = (\omega_{i_0}/r^{i_0}) \sup_{a\in\mathbb{R}^n} \sum_{i=j}^{n} r^i a_i, \text{ where } \sum_{i=k}^{n} a_i \le \frac{(1+\lambda)/2}{1-(1+\lambda)/2} a_{k-1} \text{ for all } 1 \le k \le n. \tag{27}$$

Note that the solution of the maximization problem is an upper bound on the right-hand side of (25), and thus

$$\sum_{i=j}^{\infty} \omega_i \mathbb{E}\left[q_i(\Phi(G_N,\infty))\right] \le \textsc{Sol}(j) = \sup_{n\ge 1} \ \textsc{Sol}_n(j), \tag{28}$$

where the last equality follows because each sequence in ℓ_1 can be written as the limit of a sequence of vectors in $\mathbb{R}^n$ in the ℓ_1-topology. The finite projected maximization problem is linear, and hence the solution is found at one of the boundary points of the feasible region. In this case, for fixed $n \ge 1$ and $j \ge i_0$, there is only one boundary point, the point for which the constraints hold with equality:

$$a_i = \left(\frac{1+\lambda}{2}\right)^i \text{ for } 0 \le i \le n-1, \qquad a_n = \frac{1}{1-(1+\lambda)/2}\left(\frac{1+\lambda}{2}\right)^n, \tag{29}$$

such that

$$\textsc{Sol}_n(j) = (\omega_{i_0}/r^{i_0}) \sup_{a\in\mathbb{R}^n} \sum_{i=j}^{n} r^i a_i = \frac{\omega_{i_0}}{r^{i_0}} \frac{r^j}{1-(1+\lambda)/2}\left(\frac{1+\lambda}{2}\right)^j. \tag{30}$$

Therefore,

$$\lim_{j\to\infty} \limsup_{N\to\infty} \mathbb{P}\left(\sum_{i=j}^{\infty} \omega_i q_i(\Phi(G_N,\infty)) > \varepsilon\right) \le \lim_{j\to\infty} \frac{1}{\varepsilon} \frac{\omega_{i_0}}{r^{i_0}} \frac{r^j}{1-(1+\lambda)/2}\left(\frac{1+\lambda}{2}\right)^j = 0. \quad \square \tag{31}$$

4.3. Global Stability

The last key property needed in establishing the interchange of limits is the global stability of the process-level limit in (3). As stated in the next theorem, global stability shows that the limiting system of ODEs converges to a fixed point if started from suitable states.

Theorem 6. *For any positive sequence* $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots)$, *let us define*

$$\Psi_\omega(t) := \sum_{i=1}^{\infty} \omega_i |q_i^*(t) - q_i^*(\infty)|$$

and $i_0 := \min\{i \ge 1 \mid \lambda(2q_i^*(\infty)+1) < \frac{1+\lambda}{2}\}$. *For each* $\lambda < 1$, *there exists a choice of* $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots)$ *satisfying*

$$\omega_{i_0+i} = \omega_{i_0} r^i \text{ for } i \ge 1 \tag{32}$$

for some $r \in (1, 2/(1+\lambda))$ *such that* $\Psi_\omega(t)$ *converges exponentially to zero as* $t \to \infty$, *if* $\Psi_\omega(0) < \infty$.

The proof relies on the global stability result of the classical mean-field limit of the JSQ(d) policy, as stated in Mitzenmacher [22, theorem 3.6]. The details can be found in Appendix C.

4.4. Proof of Theorem 1

We now have all the necessary results to prove Theorem 1.

Proof of Theorem 1. The process-level convergence of the occupancy process under proportional sparsity follows from Theorem 4. For the convergence of steady states, we follow the usual interchange of limits argument.

Fix $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots)$ as in Equation (32) for r chosen as in Theorem 6. By Theorem 5, $\{\boldsymbol{q}(\Phi(G_{N,}\infty))\}_{N\ge 1}$ is tight when $\mathcal{Y}$ is endowed with the ℓ_1^ω-topology, and hence, any subsequence has a convergent further subsequence. Let $\{\boldsymbol{q}(\Phi(G_{N_n,}\infty))\}_{n\ge 1}$ be such a convergent subsequence with $\{N_n\}_{n\ge 1} \subseteq \mathbb{N}$, and assume that $\boldsymbol{q}(\Phi(G_{N_n},\infty))$ converges weakly to some distribution $\hat{\pi}$, as $n \to \infty$. Observe that $\hat{\pi}$ must be supported on

$$\mathcal{Y}_\omega := \left\{ \boldsymbol{q} \in \ell_1 \,\middle|\, \sum_{i=1}^{\infty} \omega_i q_i < \infty,\ q_i \in [0,1], q_i \ge q_j, i < j, i, j \in \mathbb{N} \right\}.$$

Now, imagine starting the system in the steady state as $\Phi(G_{N_n},0) \sim \Phi(G_{N_n},\infty)$. Then $\Phi(G_{N_n},t) \sim \Phi(G_{N_n},\infty)$ for all $t \in [0,T]$, and also, by Theorem 4, it follows that

$$\boldsymbol{q}(\Phi(G_{N_n},t)) d\boldsymbol{q}^*(t) \text{ as } n \to \infty, \tag{33}$$

uniformly on $[0, T]$. Therefore,

$$\boldsymbol{q}^*(t) \overset{d}{=} \hat{\pi} \text{ for all } t \in [0, T]. \tag{34}$$

It follows that $\hat{\pi}$ is an invariant measure of the limiting dynamics. To this end, the global stability result in Theorem 6 implies that $\hat{\pi}$ must be the Dirac measure at the fixed point $\boldsymbol{q}^*(\infty)$ of $\boldsymbol{q}^*(t)$. This completes the proof of interchange of limits. □

5. Simulation Experiments

In this section, we perform extensive simulation experiments to evaluate our results and to gain insights into the cases that are not included in our theoretical framework, for example, when the compatibility graph may not be proportionally sparse.

5.1. Verification of Process-Level Convergence

Figure 3(a) shows the evolution of the occupancy process for $M(N) = N = 10^2$ and $M(N) = N = 10^4$, where the compatibility graphs are single instances of bipartite Erdős–Rényi random graphs (ERRGs) with edge probability $(\ln(N))^2/N$. Although the average number of neighbors is significantly less than the number of neighbors in a fully flexible system, the simulation illustrates that the occupancy processes closely follow the limiting dynamics of a fully flexible system. The sample path trajectories for $N = 10^2$ further exhibit the validity of the asymptotic results for fairly small values of N.

5.2. Influence of the Level of Connectivity

In Figure 3(b), we examine the dependence of the average queue length in steady state on the average server degree in G_N. In each case, we have assumed $M(N) = N$ and the compatibility graph is a single instance of the bipartite ERRG. When the average degree scales as $(\ln(N))^2$, the graph sequence satisfies the conditions in Theorem 1, and the average queue length can be seen to converge to the fixed point $\boldsymbol{q}^*(\infty)$ given by (4). Among the graph sequences that do not satisfy the conditions in Theorem 3 are the cases when the average degree is (i) a constant and (ii) equal to $\ln(N)$. The simulation shows that in case (i), the behavior differs significantly from the fully flexible case, and in case (ii) the average queue length converges to the fixed point. The latter is an edge case not captured by Theorem 3.

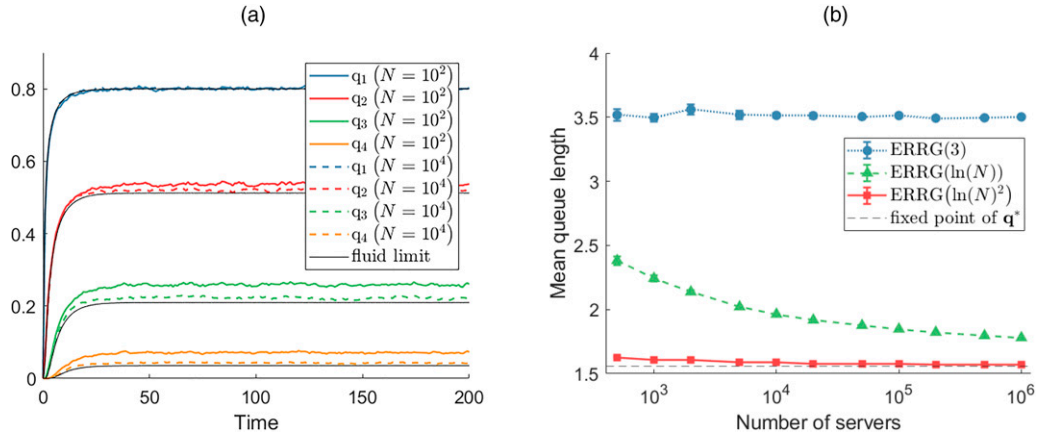
5.3. Influence of the Arrival Rate

Figure 4(a) shows the dependence of the average queue length in steady state on the arrival rate λ. Although the average queue length converges to the fixed point $\boldsymbol{q}^*(\infty)$ for all $\lambda < 1$, the rate of convergence is faster for smaller arrival rates. The difference in convergence rate is especially notable for $\lambda = 0.8$.

5.4. Influence of the Service-Time Distribution

Figure 4(b) shows the average queue length in steady state when service times are either exponentially distributed, deterministic (fixed at one), or distributed as a power law with exponent three. Note that these simulations

Figure 3. (Color online) Performance of the JSQ(2) policy on a bipartite Erdős–Rényi graph for $\lambda = 0.8$.



Notes. (a) The occupancy process for compatibility graphs with average degree $(\ln(N))^2$ for $N = 10^2, 10^4$. (b) The average queue length in steady state for varying degrees.

go beyond the theoretical results for the exponential distribution. In each of the above three cases, Figure 4(b) shows that the behavior of the steady state average queue length with a sparse bipartite ERRG is close to that of the corresponding fully flexible system. This hints at a possible universality result, where one would expect to see an analog of Theorem 1 under a general service time distribution, although the behavior for different service time distributions will be different.

5.5. Influence of Geometry

In the broader context of spatial queueing systems, servers and dispatchers are often geometrically constrained. The compatibility graph in this case arises because of the proximity of servers and dispatchers, and such a geometrically constrained graph often does not satisfy the proportional sparsity condition. In Figure 5(a), we investigate using the well-known *random geometric graph* model, if and when the performance of such networks is asymptotically indistinguishable from the fully flexible model. A random geometric graph is built by placing each of the N servers and $M(N)$ dispatchers at a uniformly selected location in $[0,1]^2$. As Figure 5(a) illustrates, dispatchers are then connected to all servers within a certain radius. Dispatchers will therefore have only local connections to servers in their proximity.

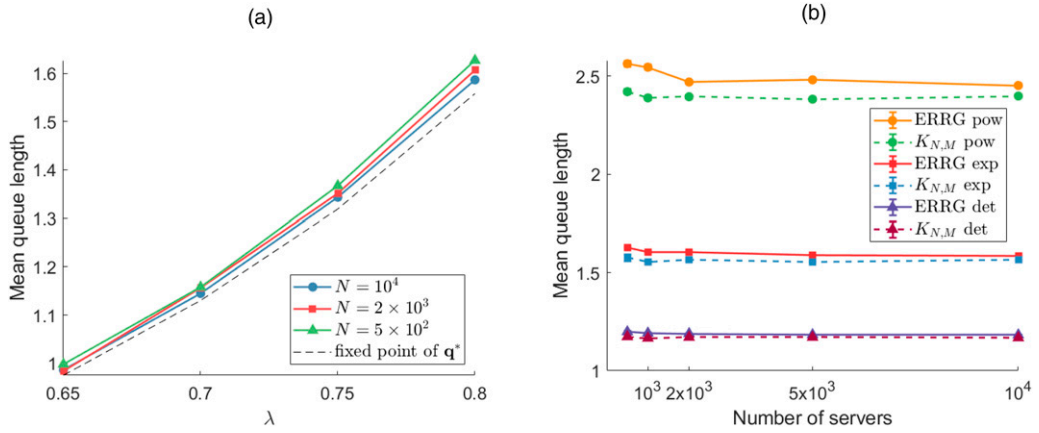
Figure 5(b) shows a comparison between the average steady state queue length when the compatibility graphs are either single instances of a bipartite ERRG or a random geometric graph, respectively. Notice that the average steady state queue length of the random geometric graph appears to converge to the fixed point $\boldsymbol{q}^*(\infty)$ if the average degree is $(\ln(N))$ or $(\ln(N))^2$, although the proportional sparsity condition is not satisfied. Extending our results for such spatial queueing systems would be an important future research direction.

6. Conclusion

In this paper, we studied the impact of task–server compatibility constraints on the performance of the JSQ(d) algorithm in large-scale systems. The results extend the validity of the mean-field approximation far beyond the fully flexible scenario, allowing room for a much sparser class of compatibility graphs to asymptotically match the performance of a fully flexible system. We have also provided explicit constructions of random compatibility graphs that are shown to almost surely belong to the above class, while being significantly sparser than the complete bipartite graph. In the context of large-scale data centers, this translates into significant reductions in implementation complexity and storage capacity, because both are roughly proportional to the server degrees in the compatibility graph. Extensive simulation experiments corroborate the theoretical results. More interestingly, some simulation experiments seem to suggest that the mean-field approximation may be valid for an even larger class of spatial graphs, where our expansion criterion does not hold. This would be an interesting future research direction.

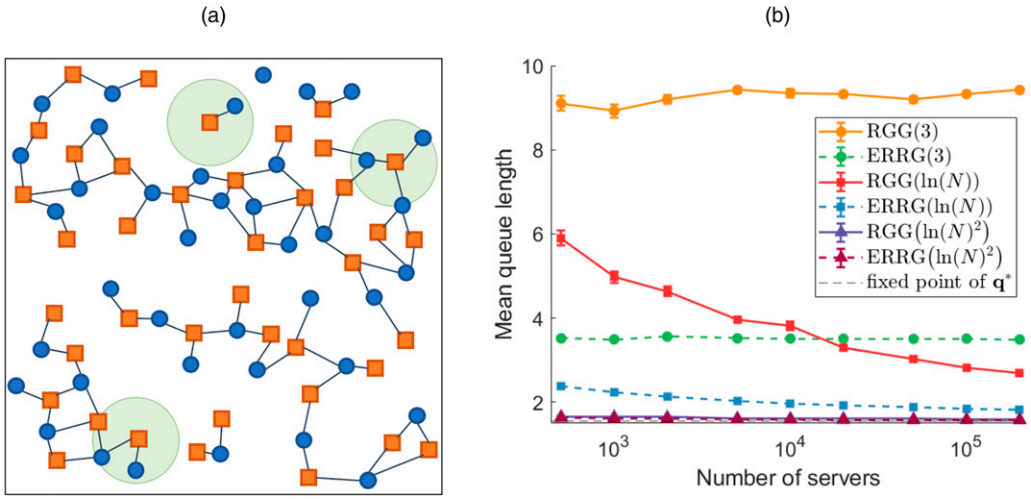
Also, in an ongoing work, we are exploring the heterogeneous case where the arrival and service rates may depend on the task type and the server, respectively, to see how the sufficient condition on the graph sequence depends on the heterogeneity in the arrival rates or the service rates.

Figure 4. (Color online) Performance of the JSQ(2) policy on a bipartite Erdős–Rényi random graph. Here, pow means a power law with exponent three, exp means an exponential distribution and det means deterministic (fixed at one).



Notes. (a) The average queue length in steady state for $(\ln(N))^2$ average degree and varying arrival rates. (b) The average queue length in steady state for $(\ln(N))^2$ average degree and varying service distributions.

Figure 5. (Color online) Performance of large-scale systems with a spatially constrained compatibility graph.



Notes. (a) A random geometric graph (RGG) is built by connecting dispatchers (square) to servers (circle) that are within a specific distance to each other. (b) The average queue length in steady state of a bipartite Erdős-Rényi random graph for varying degrees, d = 2, and $\lambda = 0{:}8$, in comparison to a RGG with same average degree.

Appendix A. Hard Constraint on Server Degrees

To establish Theorem 2, it suffices to prove that the sequence of random graphs as described in Theorem 2 satisfies the conditions of proportional sparsity and subcriticality, almost surely. Throughout this section, $\{G_N\}_{N\geq 1}$ will denote the sequence of random graphs as described in Theorem 2, where each server has degree $c(N)$.

Verification of Proportional Sparsity

We start by verifying the proportional sparsity condition, as stated in Proposition A.1. Lemmas A.1–A.3 below provide necessary technical results for the proof of Proposition A.1.

Lemma A.1. *Let X_N be a Binomial$(N,p(N))$ random variable. If $p(N)\to 0$ as $N\to\infty$, then for any fixed $\delta\in(0,1/2)$, there exists $N_0\geq 1$ and $a>0$ such that for all $N\geq N_0$,*

$$\mathbb{P}(X_N\geq \delta N)\leq (a(p(N))^{\delta})^N. \tag{A.1}$$

Proof. Fix $\delta\in(0,1/2)$. Let $N_0\geq 1$ such that $p(N)\leq\delta$ for all $N\geq N_0$. By the Chernoff bound for binomials (Janson et al. [17, equation 2.4]), we get, for all $N\geq N_0$,

$$\begin{aligned}
&\mathbb{P}(X_N\geq\delta N)\\
&\leq \exp\left(-N\left((\delta+p(N))\ln\left(\frac{\delta+p(N)}{p(N)}\right)-(1-p(N)-\delta)\ln\left(\frac{1-p(N)}{1-p(N)-\delta}\right)\right)\right)\\
&\leq \exp\left(-N\left(\delta\ln\left(\frac{\delta}{p(N)}\right)-(1-\delta)\ln\left(\frac{1-p(N)}{1-2\delta}\right)\right)\right)\\
&= \exp\left(-\delta\ln\left(\frac{1}{p(N)}\right)+\left(\delta\ln\left(\frac{1}{\delta}\right)+(1-\delta)\ln\left(\frac{1}{1-2\delta}\right)\right)\right)^N=(a(p(N))^{\delta})^N,
\end{aligned} \tag{A.2}$$

where $a=\exp\left(\delta\ln\left(\frac{1}{\delta}\right)+(1-\delta)\ln\left(\frac{1}{1-2\delta}\right)\right)$. □

Define the neighborhood of a server $v\in V_N$ as $\mathcal{N}_v:=\{w\in W_N|(v,w)\in E_N\}$. The next lemma provides sufficient conditions on the server degree $c(N)$ to ensure that the number of edges from the servers to any subset of dispatchers is close to its mean, almost surely.

Lemma A.2. *For any $X\subseteq W_N$ and $\varepsilon>0$, define*

$$B_N(X):=\left\{v\in V_N \;\middle|\; \left||\mathcal{N}_v\cap X|-\frac{c(N)|X|}{M(N)}\right|\geq\frac{\varepsilon c(N)|X|}{M(N)}\right\}. \tag{A.3}$$

If $c(N)=\omega(1)$ and $c(N)=\omega\left(\frac{M(N)}{N}\right)$, then for all $0<\varepsilon\leq 3/2$, $0<\delta<1/2$, and $\eta>0$, almost surely there exists $N_0\geq 1$ such that, for all $N\geq N_0$, $|B_N(X)|\leq\delta N$, for all $X\subseteq W_N$ with $|X|\geq\eta M(N)$.

Proof. Fix $0 < \varepsilon \le 3/2$, $0 < \delta < 1/2$, and $\eta > 0$. The number of edges between $v \in V_N$ and $X \subseteq W_N$ is a hypergeometric random variable. By the Chernoff bound for hypergeometric random variables (Janson et al. [17, corollary 2.3]), we have, for any $X \subseteq W_N$ with $|X| \ge \eta M(N)$,

$$\mathbb{P}\left(\left|\mathcal{N}_v \cap X| - \frac{c(N)|X|}{M(N)}\right| \ge \frac{\varepsilon c(N)|X|}{M(N)}\right) \le 2\exp\left(-\frac{\varepsilon^2 c(N)|X|}{3M(N)}\right)$$
$$\le 2\exp\left(-\frac{\varepsilon^2 \eta c(N)}{3}\right) =: p(N). \tag{A.4}$$

For any fixed X, the events in the definition of $B_N(X)$ are independent over $v \in V_N$. The random variable $|B_N(X)|$ is therefore stochastically upper bounded by a binomial $(N, p(N))$ random variable for all X. By Lemma A.1 and because $c(N) \to \infty$ as $N \to \infty$, there exists $N_1 \ge 1$ and $a > 0$ such that

$$\mathbb{P}(|B_N(X)| \ge \delta N) \le \left(ap(N)^\delta\right)^N = \exp\left(-N\left(\frac{\varepsilon^2 \eta \delta c(N)}{3} - \ln(2a)\right)\right) \le \exp\left(-\frac{\varepsilon^2 \eta \delta c(N) N}{6}\right), \tag{A.5}$$

for all $N \ge N_1$ and $X \subseteq W_N$. Because $c(N)N/M(N) \to \infty$ and $c(N) \to \infty$ as $N \to \infty$, there exists $N_2 \ge 1$ such that

$$\mathbb{P}(\exists X \subseteq W_N \text{ with } |X| \ge \eta M(N) \text{ such that } |B_N(X)| \ge \delta N) \le 2^{M(N)} \exp\left(-\frac{\varepsilon^2 \eta \delta c(N) N}{6}\right)$$
$$= \exp\left(\ln(2) M(N) - \frac{\varepsilon^2 \eta \delta c(N) N}{6}\right) \le \exp\left(-\frac{\varepsilon^2 \eta \delta c(N) N}{12}\right) \le \exp(-N) \tag{A.6}$$

for all $N \ge N_2$. Hence, the proof is completed by the first Borel–Cantelli lemma. □

We will now use Lemma A.2 to prove that for *most dispatchers*, the number of servers sharing an edge with the dispatcher is close to the mean.

Lemma A.3. *Define*

$$A_N := \{w \in W_N \mid |\mathcal{N}_w| \le (1-\varepsilon)\mathbb{E}[|\mathcal{N}_w|]\}. \tag{A.7}$$

If $c(N) = \omega(1)$ and $c(N) = \omega\left(\frac{M(N)}{N}\right)$, then for all $\varepsilon > 0$ and $\delta > 0$, almost surely there exists $N_0 \ge 1$ such that $|A_N| \le \delta M(N)$, for all $N \ge N_0$.

Proof. Fix $\varepsilon > 0$ and $\delta > 0$. Assume for the sake of contradiction that there is a sequence in $\mathbb{N}$ for which $|A_N| \ge \delta M(N)$ for all N in this sequence. The number of edges between V_N and A_N is then

$$\sum_{w \in A_N} |\mathcal{N}_w| \le \frac{(1-\varepsilon)c(N)N|A_N|}{M(N)}. \tag{A.8}$$

However, by Lemma A.2, there exists $N_0 \ge 1$ such that

$$\sum_{v \in V_N} |\mathcal{N}_v \cap A_N| \ge \sum_{v \in B_N(A_N)^c} |\mathcal{N}_v \cap A_N|$$
$$\ge (N - |B_N(A_N)|)\frac{(1-\varepsilon_1)c(N)|A_N|}{M(N)} \ge (1-\delta_1)\frac{(1-\varepsilon_1)c(N)N|A_N|}{M(N)}, \tag{A.9}$$

for all $N \ge N_0$, which contradicts Equation (A.8) for ε_1 and δ_1 small enough. □

Proposition A.1. *Assume $c(N) = \omega(1)$ and $c(N) = \omega\left(\frac{M(N)}{N}\right)$. Then the sequence of graphs $\{G_N\}_{N \ge 1}$ is proportionally sparse almost surely.*

Proof. Let $\varepsilon_1, \varepsilon_2 > 0$ and define

$$A_N := \{w \in W_N \mid |\mathcal{N}_w| \le (1-\varepsilon_1)\mathbb{E}[|\mathcal{N}_w|]\},$$
$$B_N(X) := \left\{v \in V_N \middle| \left|\mathcal{N}_v \cap X| - \frac{c(N)|X|}{M(N)}\right| \ge \frac{\varepsilon_2 c(N)|X|}{M(N)}\right\} \text{ for } X \subseteq W_N.$$

Fix $\varepsilon > 0$ and $\delta > 0$. The goal will be to prove proportional sparsity, that is, that there exists $N_0 \ge 1$ such that

$$\sup_{U \subseteq V_N} \left|\left\{w \in W_N \middle| \frac{|\mathcal{N}_w \cap U|}{|\mathcal{N}_w|} - \frac{|U|}{N}\right| \ge \varepsilon\right\}\right| \le 2\delta M(N) \text{ for } N \ge N_0. \tag{A.10}$$

Lemma A.3 showed that $|A_N| \le \delta M(N)$ for N large enough. It is therefore sufficient to prove that there exists $N_1 \ge 1$ such that

$$\sup_{U \subseteq V_N} \left|\left\{w \in A_N^c \middle| \frac{|\mathcal{N}_w \cap U|}{|\mathcal{N}_w|} - \frac{|U|}{N}\right| \ge \varepsilon\right\}\right| \le \delta M(N) \text{ for } N \ge N_1. \tag{A.11}$$

Consider a subset $U \subseteq V_N$. We proceed by contradiction. First, consider the case that there is a subset of dispatchers that is overconnected to U. In other words, assume that there exists a subset $X \subseteq A_N^c \subseteq W_N$ with $|X| \geq \delta M(N)$ such that for all $w \in X$,

$$\frac{|\mathcal{N}_w \cap U|}{|\mathcal{N}_w|} \geq \frac{|U|}{N} + \varepsilon. \tag{A.12}$$

We now distinguish two cases based on the cardinality of U. Choose $\eta := \varepsilon\delta/2$ and assume that $|U| \geq \eta N$. By Lemma A.2, the number of edges between U and X is

$$\begin{aligned}\sum_{v \in U \cap B_N(X)} |\mathcal{N}_v \cap X| + \sum_{v \in U \cap B_N(X)^c} |\mathcal{N}_v \cap X| &\leq c(N)|B_N(X)| + \frac{(1+\varepsilon_2)c(N)|X||U|}{M(N)} \\ &\leq \frac{(1+2\varepsilon_2)c(N)|X||U|}{M(N)},\end{aligned} \tag{A.13}$$

for all N large enough such that $|B_N(X)| \leq \varepsilon_2\delta\eta N$. At the same time, the number of edges between U and X is

$$\begin{aligned}\sum_{w \in X} |\mathcal{N}_w \cap U| &\geq \sum_{w \in X} |\mathcal{N}_w| \left(\frac{|U|}{N} + \varepsilon\right) \geq (1-\varepsilon_1)\frac{c(N)N}{M(N)}\left(\frac{|U|}{N} + \varepsilon\right)|X| \\ &= (1-\varepsilon_1)(1+\varepsilon)\frac{c(N)|X||U|}{M(N)},\end{aligned} \tag{A.14}$$

which contradicts Equation (A.13) for ε_1 and ε_2 small enough.

Assume now that $|U| < \eta N$. The number of edges between U and X is

$$\sum_{v \in U} |\mathcal{N}_v \cap X| \leq \eta c(N)N = \varepsilon\delta c(N)N/2. \tag{A.15}$$

Then, the number of edges between U and X is

$$\begin{aligned}&\sum_{w \in X} |\mathcal{N}_w \cap U| \geq \sum_{w \in X} |\mathcal{N}_w| \left(\frac{|U|}{N} + \varepsilon\right) \geq (1-\varepsilon_1)\frac{c(N)N}{M(N)}\left(\frac{|U|}{N} + \varepsilon\right)|X| \\ &\geq \varepsilon\delta(1-\varepsilon_1)c(N)N,\end{aligned} \tag{A.16}$$

which contradicts Equation (A.15) for ε_1 small enough.

Second, consider the case that there is a subset of dispatchers which is underconnected to U. In other words, assume that there exists $X \subseteq A_N^c \subseteq W_N$ with $|X| \geq \delta M(N)$ such that for all $w \in X$,

$$\frac{|\mathcal{N}_w \cap U|}{|\mathcal{N}_w|} \leq \frac{|U|}{N} - \varepsilon. \tag{A.17}$$

Then we consider the complement of U to find

$$\frac{|\mathcal{N}_w \cap U^c|}{|\mathcal{N}_w|} = 1 - \frac{|\mathcal{N}_w \cap U|}{|\mathcal{N}_w|} \geq 1 - \frac{|U|}{N} + \varepsilon = \frac{|U^c|}{N} + \varepsilon, \tag{A.18}$$

which leads to a contradiction in the same way as before. Hence, the sequence $\{G_N\}_{N\geq 1}$ as described in Theorem 2 is proportionally sparse almost surely.

Verification of Subcriticality

We now prove that $\{G_N\}_{N\geq 1}$ satisfies the subcriticality condition in Definition 2, almost surely. Note that it is enough to verify that $\{G_N\}_{N\geq 1}$ satisfies (2) for some choice of $\gamma_w^{v_1,\dots,v_d}(v)$. In particular, if $\gamma_w^{v_1,\dots,v_d}(v)$ is the uniform distribution, that is, $\gamma_w^{v_1,\dots,v_d}(v) = 1/|\{v_1,\dots,v_d\}|$ for $v \in \{v_1,\dots,v_d\}$, then the condition reduces to

$$\limsup_{N\to\infty} \max_{v \in V_N} \frac{N}{M(N)} \sum_{w \in W_N} \frac{1\{(v,w) \in E_N\}}{|\mathcal{N}_w|} \leq 1. \tag{A.19}$$

In the rest of this section, we will verify that $\{G_N\}_{N\geq 1}$ satisfies (A.19) almost surely.

Remark A.1. The condition of subcriticality as in (A.19) is similar to condition 1(ii) in Budhiraja et al. [4]. Translated to the current framework, condition 1(ii) in Budhiraja et al. [4] reads

$$\max_{v \in V_N} \frac{N}{M(N)} \left| \sum_{w \in W_N} \frac{\mathbb{1}\{(v,w) \in E_N\}}{|\mathcal{N}_w|} - 1 \right| \to 0 \text{ as } N \to \infty. \tag{A.20}$$

The subcriticality condition in (A.19) is less restrictive than the condition in Budhiraja et al. [4].

Lemma A.4. *Assume $c(N) = \omega(M(N)\ln(N)/N)$. Then the graph sequence $\{G_N\}_{N\geq 1}$ satisfies (A.19), and hence the subcriticality condition, almost surely.*

Proof. Let $\zeta(N) := \sqrt{\ln(N)c(N)N/M(N)}$ and notice that

$$\frac{\zeta(N)}{c(N)N/M(N)} = \sqrt{\frac{\ln(N)}{c(N)N/M(N)}} \to 0 \text{ as } N \to \infty. \tag{A.21}$$

We know that $|\mathcal{N}_w|$ is binomial with parameters N and $c(N)/M(N)$. By the Chernoff bound for binomials (Janson et al. [17, theorem 2.1]),

$$\mathbb{P}\left(|\mathcal{N}_w| \le \frac{c(N)N}{M(N)} - a\zeta(N)\right) \le \exp\left(-\frac{a^2\zeta(N)^2}{2c(N)N/M(N)}\right) \le \exp\left(-\frac{a^2\ln(N)}{2}\right) = N^{-a^2/2}. \tag{A.22}$$

Hence,

$$\mathbb{P}\left(\exists w \in W_N \text{ such that } |\mathcal{N}_w| \ge \frac{c(N)N}{M(N)} - a\zeta(N)\right) \le N^{1-a^2/2}. \tag{A.23}$$

As the probabilities are summable over N for arbitrary $a^2 > 4$, the first Borel–Cantelli lemma proves the events can only happen finitely many times. Hence, there exists $N_0 \ge 1$ such that

$$|\mathcal{N}_w| \ge c(N)N/M(N) - a\zeta(N) \text{ for all } w \in W_N \text{ and } N \ge N_0. \tag{A.24}$$

Therefore,

$$\begin{aligned}\max_{v\in V_N} \frac{N}{M(N)} \sum_{w\in W_N} \frac{\mathbb{1}\{(v,w)\in E_N\}}{|\mathcal{N}_w|} &\le \frac{N}{M(N)} \frac{c(N)}{\min_{W\in V_N}|\mathcal{N}_w|} \\ &\le \frac{N}{M(N)} \frac{c(N)}{c(N)N/M(N) - a\zeta(N)} = 1 + \frac{a\zeta(N)}{c(N)M(N)/N - a\zeta(N)} \to 1 \text{ as } N \to \infty. \quad \square\end{aligned}$$

Proof of Theorem 2. The proof of Theorem 2 follows immediately from Proposition A.1 and Lemma A.4. $\square$

Appendix B. Inhomogeneous Levels of Flexibility

Throughout this section, $\{G_N\}_{N\ge1}$ will denote the sequence of random graphs as described in Theorem 3, where $(p_w(N))_{w\in W_N}$ is the connection probability vector. To establish Theorem 3, it suffices to prove that $\{G_N\}_{N\ge1}$ satisfies the conditions of proportional sparsity and subcriticality, almost surely.

Verification of Proportional Sparsity

We start by verifying the proportionally sparsity condition. Define $\bar{p}(N) := \min_{w\in W_N} p_w(N)$.

Lemma B.1. *Assume $\bar{p}(N) = \omega(1/N)$ and $\bar{p}(N) = \omega(1/M(N))$. Then the sequence of graphs $\{G_N\}_{N\ge1}$ is proportionally sparse almost surely.*

Proof. Fix $\varepsilon > 0$, $w \in W_N$ and $U \subseteq V_N$. Let $B_w(U)$ denote the event that a dispatcher $w \in W_N$ is bad with respect to the set $U \subseteq V_N$, defined as

$$B_w(U) := \left\{\left|\frac{|\mathcal{N}_w \cap U|}{|\mathcal{N}_w|} - \frac{|U|}{N}\right| \ge \varepsilon\right\}. \tag{B.1}$$

Define $\zeta_w(N) := p_w(N)N$ and $\eta := |U|/N$. By the law of total probability,

$$\begin{aligned}&\mathbb{P}(B_w(U)) \\ &\le \mathbb{P}(B_w(U) \,|\, ||\mathcal{N}_w \cap U| - \eta\zeta_w(N)| < \varepsilon_1\zeta_w(N) \text{ and } ||\mathcal{N}_w| - \zeta_w(N)| < \varepsilon_2\zeta_w(N)) \\ &\quad + \mathbb{P}(||\mathcal{N}_w \cap U| - \eta\zeta_w(N)| \ge \varepsilon_1\zeta_w(N)) + \mathbb{P}(||\mathcal{N}_w| - \zeta_w(N)| \ge \varepsilon_2\zeta_w(N)).\end{aligned} \tag{B.2}$$

We will bound each term on the right-hand side of (B.2). The first term becomes equal to zero by choosing $\varepsilon_1 = \varepsilon/6$ and $\varepsilon_2 = \min(\varepsilon/6, 1/2)$ such that

$$\frac{|\mathcal{N}_w \cap U|}{|\mathcal{N}_w|} - \eta \le \frac{\eta\zeta_w(N) + \varepsilon_1\zeta_w(N)}{\zeta_w(N) - \varepsilon_2\zeta_w(N)} - \eta \le \frac{\varepsilon_1 + \varepsilon_2}{1 - \varepsilon_2} < \varepsilon \tag{B.3}$$

and

$$\frac{|\mathcal{N}_w \cap U|}{|\mathcal{N}_w|} - \eta \ge \frac{\eta\zeta_w(N) - \varepsilon_1\zeta_w(N)}{\zeta_w(N) + \varepsilon_2\zeta_w(N)} - \eta \ge -\frac{\varepsilon_1 + \varepsilon_2}{1 + \varepsilon_2} > -\varepsilon. \tag{B.4}$$

By applying the Chernoff bound for binomials (Janson et al. [17, theorem 2.1]) on the second term, we obtain

$$\begin{aligned}\mathbb{P}(|\mathcal{N}_w \cap U| - \eta\zeta_w(N)| \geq (\varepsilon_1/\eta)\eta\zeta_w(N)) \\ \leq 2\exp\left(-\frac{\eta\zeta_w(N)(\varepsilon_1/\eta)^2}{3}\right) \leq 2\exp\left(-\frac{\bar{p}(N)N\varepsilon_1^2}{3}\right),\end{aligned} \tag{B.5}$$

and applying the Chernoff bound (Janson et al. [17, theorem 2.1) on the third term yields

$$\mathbb{P}(|\mathcal{N}_w| - \zeta_w(N)| \geq \varepsilon_2 p_w(N)N) \leq 2\exp\left(-\frac{\zeta_w(N)\varepsilon_2^2}{3}\right) \leq 2\exp\left(-\frac{\bar{p}(N)N\varepsilon_2^2}{3}\right). \tag{B.6}$$

Therefore,

$$\mathbb{P}(B_w(U)) \leq 2\exp\left(-\frac{\bar{p}(N)N\varepsilon_1^2}{3}\right) + 2\exp\left(-\frac{\bar{p}(N)N\varepsilon_2^2}{3}\right), \tag{B.7}$$

uniformly over $w \in W_N$ and $U \subseteq V_N$. Fix $0 < \delta < 1/2$ and define $\alpha(\delta)$ and $\beta(\delta)$ as

$$\alpha(\delta) := \frac{4}{\delta}, \beta(\delta) := \exp\left(-\frac{2}{\delta}\left(\delta\ln\left(\frac{1}{\delta}\right) + (1-\delta)\ln\left(\frac{1}{1-2\delta}\right)\right)\right). \tag{B.8}$$

Because $\bar{p}(N)N \to \infty$ as $N \to \infty$, Equation (B.7) converges to zero as N becomes large. Hence, for all $\delta > 0$, there exists $N_0 \geq 1$ such that $\mathbb{P}(B_w(U)) \leq \beta(\delta)$ for all $N \geq N_0$. Similarly, because $\bar{p}(N)M(N) \to \infty$ as $N \to \infty$, it follows that for all $\delta > 0$, there exists $N_1 \geq 1$ such that $\mathbb{P}(B_w(U)) \leq \exp\left(-\frac{\alpha(\delta)N}{M(N)}\right)$ for all $N \geq N_1$. As the events $\{B_w(U)\}_{w\in W_N}$ are independent, for all $\delta > 0$ and $N \geq \max(N_0, N_1)$, the sum of their indicators $\sum_{w\in W_N} \mathbb{1}(B_w(U))$ is stochastically dominated by a binomial $(M(N), \theta)$ random variable, where

$$\theta := \min\left(\exp\left(-\frac{\alpha(\delta)N}{M(N)}\right), \beta(\delta)\right). \tag{B.9}$$

By the Chernoff bound for binomials (Janson et al. [17, theorem 2.1]),

$$\begin{aligned}&\mathbb{P}\left(\sum_{w\in W_N} \mathbb{1}(B_w(U)) \geq \delta M(N)\right) \\ &\leq \exp\left(-M(N)\left(\delta\ln\left(\frac{\delta}{\theta}\right) - (1-\delta)\ln\left(\frac{1-\theta}{1-2\delta}\right)\right)\right) \\ &\leq \exp\left(-\frac{\delta M(N)}{2}\ln\left(\frac{1}{\theta}\right) + M(N)\left(\delta\ln\left(\frac{1}{\delta}\right) + (1-\delta)\ln\left(\frac{1}{1-2\delta}\right) - \frac{\delta}{2}\ln\left(\frac{1}{\theta}\right)\right)\right) \\ &\leq \exp\left(-\frac{\delta M(N)}{2}\frac{\alpha(\delta)N}{M(N)} + M(N)\left(\delta\ln\left(\frac{1}{\delta}\right) + (1-\delta)\ln\left(\frac{1}{1-2\delta}\right) - \frac{\delta}{2}\ln\left(\frac{1}{\beta(\delta)}\right)\right)\right) \\ &\leq \exp(-2N)\end{aligned} \tag{B.10}$$

for all $N \geq \max(N_0, N_1)$, and hence,

$$\mathbb{P}\left(\sup_{U\subseteq V_N}\sum_{w\in W_N} \mathbb{1}(B_w(U)) \geq \delta M(N)\right) \leq \sum_{U\subseteq V_N} \mathbb{P}\left(\sum_{w\in W_N} \mathbb{1}(B_w(U)) \geq \delta M(N)\right) \leq \exp(-N). \tag{B.11}$$

As the probabilities are summable over N, the first Borel–Cantelli lemma proves that the sequence of random graphs is proportionally sparse almost surely. □

Verification of Subcriticality

We now prove that $\{G_N\}_{N\geq 1}$ satisfies the subcriticality condition in Definition 2, almost surely. Note, as in Appendix A, we will verify that $\{G_N\}_{N\geq 1}$ satisfies (A.19) almost surely. Recall that $\bar{p}(N) = \min_{w\in W_N} p_w(N)$.

Lemma B.2. *Assume* $\bar{p}(N) = \omega((\ln(M(N)) + \ln(N))/N)$ *and* $\|\boldsymbol{p}(N)^{-1}\|_2^{-2} = \omega(\ln(N)/M(N)^2)$. *Then the sequence of graphs* $\{G_N\}_{N\geq 1}$ *satisfies the subcriticality condition almost surely.*

Proof. Let $\zeta_w(N) := \sqrt{(\ln(M(N)) + \ln(N))p_w(N)N}$, and notice that

$$\frac{\zeta_w(N)}{p_w(N)N} \leq \sqrt{\frac{\ln(M(N)) + \ln(N)}{\bar{p}(N)N}} \to 0 \text{ as } N \to \infty. \tag{B.12}$$

We know that $|\mathcal{N}_w|$ is a binomial $(N, p_w(N))$ random variable. By the Chernoff bound for binomials (Janson et al. [17, theorem 2.1]),

$$\begin{aligned}\mathbb{P}(|\mathcal{N}_w| \geq p_w(N)N - c_1\zeta_w(N)) &\leq \exp\left(-\frac{c_1^2\zeta_w(N)^2}{2p_w(N)N}\right)\\ &\leq \exp\left(-\frac{c_1^2(\ln(M(N)) + \ln(N))}{2}\right) = M(N)^{-c_1^2/2}\cdot N^{-c_1^2/2}.\end{aligned} \tag{B.13}$$

Hence,

$$\mathbb{P}(\exists w \in W_N \text{ such that } |\mathcal{N}_w| \geq p_w(N)N - c_1\zeta_w(N)) \leq M(N)^{1-c_1^2/2}\cdot N^{-c_1^2/2}. \tag{B.14}$$

As the probabilities are summable over N for arbitrary $c_1^2 > 2$, because of the first Borel–Cantelli lemma, almost surely, there exists $N_0 \geq 1$ such that for all $N \geq N_0$,

$$|\mathcal{N}_w| \geq p_w(N)N - c_1\zeta_w(N) \text{ for all } w \in W_N. \tag{B.15}$$

Define the functions f_v and g for $v \in V_N$ as

$$f_v(\boldsymbol{z}) = \frac{N}{M(N)}\sum_{w\in W_N}\frac{z_w}{|\mathcal{N}_w\backslash\{v\}| + z_w}, \qquad g(\boldsymbol{z}) = \frac{N}{M(N)}\sum_{w\in W_N}\frac{z_w}{p_w(N)N - c_1\zeta_w(N)}, \tag{B.16}$$

for vectors $\boldsymbol{z} \in \{0,1\}^{M(N)}$. Note that for $N \geq N_0$,

$$\begin{aligned}f_v(\boldsymbol{z}) - g_v(\boldsymbol{z}) &= \frac{N}{M(N)}\sum_{w\in W_N} z_w\left(\frac{1}{|\mathcal{N}_w\backslash\{v\}| + 1} - \frac{1}{p_w(N)N - c_1\zeta_w(N)}\right)\\ &\leq \frac{N}{M(N)}\sum_{w\in W_N} z_w\left(\frac{1}{|\mathcal{N}_w|} - \frac{1}{p_w(N)N - c_1\zeta_w(N)}\right) \leq 0,\end{aligned} \tag{B.17}$$

and therefore $f_v(\boldsymbol{z}) \leq g_v(\boldsymbol{z})$. Define the edge indicators $Z_{v,w} = \mathbb{1}\{(v,w) \in E_N\}$ for $v \in V_N$ and $w \in W_N$. Note that the random variables $\{Z_{v,w}\}_{w\in W_N}$ are independent. Furthermore, there exists $N_1 \geq 1$ such that if two vectors $\boldsymbol{z}, \boldsymbol{z}' \in \{0,1\}^{M(N)}$ differ only in the wth coordinate, then

$$|g_v(\boldsymbol{z}) - g_v(\boldsymbol{z}')| = \frac{N}{M(N)}\frac{1}{p_w(N)N - c_1\zeta_w(N)} \leq \frac{2}{p_w(N)M(N)}, \tag{B.18}$$

for $N \geq N_1$. Finally, note that

$$\mathbb{E}\left[g_v\left(\{Z_{v,w}\}_{w\in W_N}\right)\right] \to 1 \text{ as } N \to \infty. \tag{B.19}$$

We apply the Azuma–Hoeffding inequality (Janson et al. [17, corollary 2.27]) and use the condition $\|\boldsymbol{p}(N)^{-1}\|_2^{-2} = \omega(\ln(N)/M(N)^2)$ to obtain

$$\begin{aligned}\mathbb{P}\left(f_v\left(\{Z_{v,w}\}_{w\in W_N}\right) \geq 1 + 2\varepsilon\right) &\leq \mathbb{P}\left(g_v\left(\{Z_{v,w}\}_{w\in W_N}\right) \geq \mathbb{E}\left[g_v\left(\{Z_{v,w}\}_{w\in W_N}\right)\right] + \varepsilon\right)\\ &\leq \exp\left(-\frac{\varepsilon^2}{2\sum_{w\in W_N}\frac{4}{p_w(N)^2M(N)^2}}\right) = \exp\left(-\frac{\varepsilon^2M(N)^2}{8\|\boldsymbol{p}(N)^{-1}\|_2^2}\right) \leq \exp\left(-\frac{c_2\varepsilon^2\ln(N)}{8}\right)\end{aligned} \tag{B.20}$$

for all $c_2 > 0$ and N large enough. Therefore,

$$\mathbb{P}\left(\exists v \in V_N \text{ such that } \max_{v\in V_N}\frac{N}{M(N)}\sum_{w\in W_N}\frac{\mathbb{1}\{(v,w)\in E_N\}}{|\mathcal{N}_w|} \geq 1 + 2\varepsilon\right) \leq N^{1-\frac{c_2\varepsilon^2}{8}}. \tag{B.21}$$

As the probabilities are summable over N for c_2 large enough, the first Borel–Cantelli lemma proves that the graph sequence satisfies the subcriticality condition almost surely. □

Proof of Theorem 3. The proof of Theorem 3 follows immediately from Lemmas B.1 and B.2. □

Appendix C. Global Stability Analysis

The proof of global stability is based on the proof of theorem 3.6 in Mitzenmacher [22]. Recall that $\Psi_\omega(t) := \sum_{i=1}^{\infty}\omega_i|q_i^*(t) - q_i^*(\infty)|$ and $i_0 := \min\{i \geq 1 | \lambda(2q_i^*(\infty) + 1) < \frac{1+\lambda}{2}\}$.

Proof of Theorem 6. To establish the theorem, it is sufficient to show that there exists $1 < r < 2/(1+\lambda)$ such that $\boldsymbol{\omega} \in \mathbb{R}^\infty$ satisfies the condition from the proof of theorem 3.6 in Mitzenmacher [22]. That is, if there exists $\delta > 0$ such that

$$\omega_{i+1} \leq \omega_i + \frac{\omega_i(1-\delta) - \omega_{i-1}}{\lambda(2q_i^*(\infty) + 1)} \quad \text{for} \quad i \geq 1, \tag{C.1}$$

then Ψ_ω converges exponentially to zero if $\Psi(0) < \infty$. As suggested in the proof of theorem 3.6 in Mitzenmacher [22], the weights are broken up into two subsequences starting with $\omega_0 = 0$ and $\omega_1 = 1$. For $1 \le i \le i_0 - 1$, we set

$$\omega_{i+1} := \omega_i + \frac{\omega_i(1-\delta) - \omega_{i-1}}{3} \le \omega_i + \frac{\omega_i(1-\delta) - \omega_{i-1}}{\lambda(2q_i^*(\infty)+1)}. \tag{C.2}$$

Note that the subsequence $\omega_0, \omega_1, \ldots, \omega_{i_0}$ consists of finitely many terms. Hence, there exists $\delta_0 > 0$ such that this subsequence is increasing for $\delta \le \delta_0$. Applying Equation (C.1) to the (i_0+1) th term yields

$$\omega_{i_0+1} := \omega_{i_0} r \le \omega_{i_0} + \frac{\omega_{i_0}(1-\delta) - \omega_{i_0-1}}{\lambda(2\pi_{i_0}+1)}, \tag{C.3}$$

from which it follows that

$$r \le 1 + \frac{1}{\omega_{i_0}} \frac{\omega_{i_0}(1-\delta) - \omega_{i_0-1}}{\lambda(2\pi_{i_0}+1)} =: R(\delta). \tag{C.4}$$

Note that $R(\delta)$ increases as δ decreases. Hence, there exists $\delta_1 > 0$ such that $R(\delta) > 1 + \epsilon$ for $\delta \le \delta_1$ and some $\epsilon > 0$. Define

$$r(\delta) := \frac{1}{2}\left(1 + \frac{2-2\delta}{1+\lambda} - \sqrt{\left(1 + \frac{2-2\delta}{1+\lambda}\right)^2 - \frac{8}{1+\lambda}}\right), \tag{C.5}$$

and $\omega_{i_0+i} := \omega_{i_0} r(\delta)^i$ for $i \ge 1$ such that

$$\omega_{i+1} = \omega_i + \frac{2\omega_i(1-\delta) - 2\omega_{i-1}}{1+\lambda} \le \omega_i + \frac{\omega_i(1-\delta) - \omega_{i-1}}{\lambda(2q_i^*(\infty)+1)}, \tag{C.6}$$

for $i \ge i_0 + 1$. Note that $r(\delta) \to 1$ as $\delta \to 0$. Hence, there exists $\delta_2 > 0$ such that $1 < r(\delta) < \min(2/(1+\lambda), 1+\epsilon)$ for $\delta \le \delta_2$. Finally, let $\delta := \min(\delta_0, \delta_1, \delta_2)$ and $r = r(\delta)$ such that Equation (C.1) is satisfied. □

Appendix D. Proportional Sparsity and Quasi Randomness

Proof of Lemma 1. Note that the quasi-random graph can be characterized by the discrepancy condition (Chung et al. [7, property 4]), which, for bipartite graphs, states that a graph sequence $\{G_N\}_{N\ge1}$ is quasi-random if there exists a fixed $0 < p < 1$ such that for all $U_N \subseteq V_N$ and $B_N \subseteq W_N$,

$$\left| \sum_{w \in B_N} |\mathcal{N}_w \cap U_N| - p|U_N||B_N| \right| = o(NM(N)). \tag{D.1}$$

We proceed by contradiction. Assume, if possible, that there exists $\varepsilon > 0$ and a sequence of choices for N for which there exists $U \subseteq V_N$ such that

$$\left| \left\{ w \in W_N \middle| \left| \frac{|\mathcal{N}_w \cap U|}{|\mathcal{N}_w|} - \frac{|U|}{N} \right| \ge \varepsilon \right\} \right| \ge 2\delta M(N), \tag{D.2}$$

for $\delta > 0$. Define the overconnected set of dispatchers as

$$B_1 := \left\{ w \in W_N \middle| \frac{|\mathcal{N}_w \cap U|}{|\mathcal{N}_w|} - \frac{|U|}{N} \ge \varepsilon \right\}, \tag{D.3}$$

and assume without loss of generality that $|B_1| \ge \delta M(N)$. Let $p := \sum_{w\in B_1} |\mathcal{N}_w|/(|B_1|N)$ be the average connection probability. Now,

$$\sum_{w \in B_1} |\mathcal{N}_w \cup U| \ge \left(\frac{|U|}{N} + \varepsilon\right) \sum_{w \in B_1} |\mathcal{N}_w| = p|U||B_1| + \varepsilon p|B_1|N \ge p|U||B_1| + \varepsilon\delta pNM(N). \tag{D.4}$$

Define the underconnected set of dispatchers to U^c as

$$\begin{aligned} B_2 &:= \left\{ w \in W_N \middle| \frac{|\mathcal{N}_w \cap U^c|}{|\mathcal{N}_w|} - \frac{|U^c|}{N} \le -\varepsilon \right\} \\ &= \left\{ w \in W_N \middle| \left(1 - \frac{|\mathcal{N}_w \cap U|}{|\mathcal{N}_w|}\right) - \left(1 - \frac{|U|}{N}\right) \le -\varepsilon \right\} \\ &= \left\{ w \in W_N \middle| \frac{|\mathcal{N}_w \cap U|}{|\mathcal{N}_w|} - \frac{|U|}{N} \ge \varepsilon \right\} = B_1. \end{aligned} \tag{D.5}$$

With the same reasoning as before,

$$\sum_{w \in B_2} |\mathcal{N}_w \cup U^c| \le \left(\frac{|U^c|}{N} - \varepsilon\right) \sum_{w \in B_2} |\mathcal{N}_w| \le p|U^c||B_1| - \varepsilon\delta pNM(N). \tag{D.6}$$

As the deviations are of the order $NM(N)$, it now follows that there is no choice for p for which Equation (D.1) is satisfied. □

Appendix E. Lipschitz Continuity of JSQ(d)

Proof of Lemma 2. Let $\boldsymbol{x} \in \mathcal{X}$ be the queue length distribution, and let $q_i = \sum_{j=i}^{\infty} x_j$ be the corresponding occupancy process. Let Π be the JSQ(d) policy. The assignment probability function $\boldsymbol{p}^{\Pi} = (p_0^{\Pi}, p_1^{\Pi}, \dots) : \mathcal{X} \to [0,1]^{\infty}$ is given by

$$p_i^{\Pi}(\boldsymbol{x}) = q_i^d - q_{i+1}^d = (x_i + q_{i+1})^d - q_{i+1}^d = \sum_{k=0}^{d-1} \binom{d}{k} x_i^{d-k} q_{i+1}^k \quad \text{for } i = 1,2,\dots. \tag{E.1}$$

Let $\boldsymbol{y} \in \mathcal{X}$ be another queue length distribution with corresponding occupancy process $r_i = \sum_{j=i}^{\infty} y_j$. Apply the triangle inequality and Lipschitz continuity of $f(z) = z^k$,

$$\begin{aligned}
&\sum_{i=0}^{\infty} |p_i^{\Pi}(\boldsymbol{y}) - p_i^{\Pi}(\boldsymbol{x})| \le \sum_{i=0}^{\infty} \sum_{k=0}^{d-1} \binom{d}{k} |y_i^{d-k} r_{i+1}^k - x_i^{d-k} q_{i+1}^k| \\
&\le d! \sum_{i=0}^{\infty} \sum_{k=0}^{d-1} \left(|y_i^{d-k} r_{i+1}^k - x_i^{d-k} r_{i+1}^k| + |r_{i+1}^k - q_{i+1}^k| |x_i^{d-k}| \right) \\
&\le d! \sum_{k=0}^{d-1} \sum_{i=0}^{\infty} |y_i^{d-k} - x_i^{d-k}| + d! \sum_{k=0}^{d-1} \max_{i \ge 1} |r_i^k - q_i^k| \sum_{i=0}^{\infty} |x_i^{d-k}| \\
&\le d! \sum_{k=0}^{d-1} (d-k) \sum_{i=0}^{\infty} |y_i - x_i| + d! \cdot d^2 \max_{i \ge 1} |r_i - q_i| \\
&\le d! \cdot d^2 \sum_{i=0}^{\infty} |y_i - x_i| + d! \cdot d^2 \max_{i \ge 1} \left| \sum_{j=i}^{\infty} \left(x_j - y_j \right) \right| \le 2d! \cdot d^2 \sum_{i=0}^{\infty} |y_i - x_i|. \quad \square
\end{aligned}$$

Appendix F. Proof of Proposition 1

To prove Proposition 1, we first need to show that the tail of the occupancy process is small uniformly on any finite time interval, for all large enough N. This is stated in the next lemma.

Lemma F.1. *If the starting states satisfy* $\|\boldsymbol{q}(\Phi(G_N,0)) - \boldsymbol{q}^*(0)\|_1 \to 0$ *as* $N \to \infty$, *for some* $\boldsymbol{q}^*(0) \in \mathcal{Y}$, *then for each* $\varepsilon > 0$, $\delta > 0$, *and* $T > 0$, *there exist* $i_0 \ge 1$ *and* $N_1 \ge 1$, *possibly depending on* λ, $\boldsymbol{q}^*(0)$, ε, δ, *and* T, *such that,*

$$\mathbb{P}\left(\sup_{t \in [0,T]} q_{i_0}(\Phi(G_N,t)) \ge \varepsilon \right) < \delta \quad \textit{for all} \quad N \ge N_1. \tag{F.1}$$

Note that Lemma F.1 does not depend on any condition on the graph sequence, nor does it require the Lipschitz continuity of the task assignment policy.

Proof of Lemma F.1. Fix $\varepsilon > 0$ and $\delta > 0$. As $\boldsymbol{q}^*(0) \in \ell_1$, there exists $j_0 = j_0(\boldsymbol{q}^*(0)) \ge 1$ such that $q_{j_0}^*(0) < \varepsilon/4$. By convergence of the starting states, there exists $N_0 \ge 1$ such that

$$\mathbb{P}(q_{j_0}(\Phi(G_N,0)) \ge \varepsilon/2) \le \mathbb{P}(|\boldsymbol{q}(\Phi(G_N,0)) - \boldsymbol{q}^*(0)|_1 \ge \varepsilon/4) < \delta/2 \quad \text{for} \quad N \ge N_0. \tag{F.2}$$

Let $i_0 = j_0 + \lceil 4\lambda T/\varepsilon \rceil$. Then,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t \in [0,T]} q_{i_0}(\Phi(G_N,t)) \ge \varepsilon \right) \\
&\le \mathbb{P}\left(\sup_{t \in [0,T]} q_{i_0}(\Phi(G_N,t)) \ge \varepsilon | q_{j_0}(\Phi(G_N,0)) < \varepsilon/2 \right) + \mathbb{P}(q_{j_0}(\Phi(G_N,0)) \ge \varepsilon/2).
\end{aligned} \tag{F.3}$$

By the choice of j_0, the second term is smaller than $\delta/2$ for $N \ge N_0$. By the condition $q_{j_0}(\Phi(G_N,0)) < \varepsilon/2$, the number of servers with at least j_0 tasks is upper bounded by $\varepsilon N/2$ at time zero. Hence, to reach $q_{i_0}(\Phi(G_N,t)) \ge \varepsilon$, there must be at least $(\varepsilon N - \varepsilon N/2)(i_0 - j_0) \ge 2\lambda NT$ tasks added to the system. Therefore, there exists $N_0' \ge 1$ such that

$$\mathbb{P}\left(\sup_{t \in [0,T]} q_{i_0}(\Phi(G_N,t)) \ge \varepsilon \middle| q_{j_0}(\Phi(G_N,0)) < \varepsilon/2 \right) \le \mathbb{P}(Z(\lambda NT) \ge 2\lambda NT) < \delta/2 \tag{F.4}$$

for $N \ge N_0'$, where $Z(\cdot)$ is a unit-rate Poisson process and $Z(\lambda NT)$ denotes the total number of arrivals into the system up to time T. Choosing $N_1 = \max(N_0, N_0')$ completes the proof of the lemma. $\square$

Proof of Proposition 1. We can upper bound the probability in Equation (9) by repeatedly applying the triangle inequality and splitting the probabilities in separate terms. For brevity, let $\boldsymbol{x}(t) := \boldsymbol{x}(\Phi(G_N,t))$ and $\boldsymbol{x}^w(t) := \boldsymbol{x}^w(\Phi(G_N,t))$, and

similarly, $\boldsymbol{q}(t) := \boldsymbol{q}(\Phi(G_N, t))$ and $\boldsymbol{q}^w(t) := \boldsymbol{q}^w(\Phi(G_N, t))$. Then,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in[0,T]}\left|\left\{w\in W_N|\sum_{i=0}^{\infty}|x_i(t)-x_i^w(t)|\geq\varepsilon\right\}\right|\geq\delta M(N)\right)\\
&\leq\mathbb{P}(\sup_{t\in[0,T]}\left|\left\{w\in W_N|\sum_{i=0}^{i_0-1}|x_i(t)-x_i^w(t)|\geq\varepsilon/4\right\}\right|\\
&\quad+\sup_{t\in[0,T]}\left|\left\{w\in W_N|\sum_{i=i_0}^{\infty}x_{i_0}^w(t)\geq\varepsilon/2\right\}\right|\\
&\quad+\sup_{t\in[0,T]}\left|\left\{w\in W_N|\sum_{i=i_0}^{\infty}x_{i_0}(t)\geq\varepsilon/4\right\}\right|\geq\delta M(N))\qquad\text{(F.5)}\\
&\leq\mathbb{P}\left(\sup_{t\in[0,T]}\left|\left\{w\in W_N|\sum_{i=0}^{i_0-1}|x_i(t)-x_i^w(t)|\geq\varepsilon/4\right\}\right|>\delta M(N)/4\right)\\
&\quad+\mathbb{P}\left(\sup_{t\in[0,T]}\left|\left\{w\in W_N|q_{i_0}^w(t)\geq\varepsilon/2\right\}\right|>\delta M(N)/2\right)\\
&\quad+\mathbb{P}\left(\sup_{t\in[0,T]}|\left\{w\in W_N|q_{i_0}(t)\geq\varepsilon/4\right\}|>\delta M(N)/4\right)\\
&\leq\sum_{i=0}^{i_0-1}\mathbb{P}\left(\sup_{t\in[0,T]}|\left\{w\in W_N||x_i(t)-x_i^w(t)|\geq\varepsilon/4i_0\right\}|>\delta M(N)/4i_0\right)\\
&\quad+\mathbb{P}\left(\sup_{t\in[0,T]}\left|\left\{w\in W_N||q_{i_0}(t)-q_{i_0}^w(t)|\geq\varepsilon/4\right\}\right|>\delta M(N)/4\right)\\
&\quad+2\mathbb{P}\left(\sup_{t\in[0,T]}\mathbb{1}\{q_{i_0}(t)\geq\varepsilon/4\}>0\right),\qquad\text{(F.6)}
\end{aligned}$$

for an arbitrary choice of $i_0 \geq 1$, where in the second step we use $q_{i_0}(t) = \sum_{i=i_0}^{\infty} x_i(t)$. By Markov's inequality,

$$\begin{aligned}
&\sum_{i=0}^{i_0-1}\mathbb{P}\left(\sup_{t\in[0,T]}|\left\{w\in W_N||x_i(t)-x_i^w(t)|\geq\varepsilon/4i_0\right\}|>\delta M(N)/4i_0\right)\\
&\leq\frac{4i_0}{\delta M(N)}\sum_{i=0}^{i_0-1}\mathbb{E}\left[\sup_{t\in[0,T]}|\left\{w\in W_N||x_i(t)-x_i^w(t)|\geq\varepsilon/4i_0\right\}|\right]\\
&\leq\frac{4i_0^2}{\delta M(N)}\sup_{U\subseteq V_N}\left|\left\{w\in W_N\left|\frac{|\mathcal{N}_w\cap U|}{|\mathcal{N}_w|}-\frac{|U|}{N}\right|\geq\varepsilon/4i_0\right\}\right|,\qquad\text{(F.7)}
\end{aligned}$$

where the last step follows because the servers with queue length i form a subset of V_N at every time $t \geq 0$. Similarly, the second term on the right-hand side of (F.6) can be bounded by

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in[0,T]}\left|\left\{w\in W_N||q_{i_0}(t)-q_{i_0}^w(t)|\geq\varepsilon/4\right\}\right|>\delta M(N)/4\right)\\
&\leq\frac{4}{\delta M(N)}\mathbb{E}\left[\sup_{t\in[0,T]}\left|\left\{w\in W_N||q_{i_0}(t)-q_{i_0}^w(t)|\geq\varepsilon/4\right\}\right|\right]\\
&\leq\frac{4}{\delta M(N)}\sup_{U\subseteq V_N}\left|\left\{w\in W_N\left|\frac{|\mathcal{N}_w\cap U|}{|\mathcal{N}_w|}-\frac{|U|}{N}\right|\geq\varepsilon/4\right\}\right|.\qquad\text{(F.8)}
\end{aligned}$$

Fix $\varepsilon' > 0$. By Lemma F.1, we can choose $i_0 \geq 1$ and $N_0 \geq 1$ such that

$$\mathbb{P}\left(\sup_{t\in[0,T]}q_{i_0}(t)\geq\varepsilon/4\right)\leq\varepsilon'/4\quad\text{for}\quad N\geq N_0.\qquad\text{(F.9)}$$

Because the graph sequence is proportionally sparse, by Definition 1, there exists $N_1 \geq 1$ such that

$$\sup_{U\subseteq V_N}\left|\left\{w\in W_N\left|\frac{|\mathcal{N}_w\cap U|}{|\mathcal{N}_w|}-\frac{|U|}{N}\right|\geq\varepsilon/4i_0\right\}\right|<\frac{\delta}{4i_0^2+4}\frac{\varepsilon' M(N)}{2}\quad\text{for all}\quad N\geq N_1.\qquad\text{(F.10)}$$

Therefore, plugging the bound from (F.10) in (F.8) and (F.7) and plugging this together with (F.9) in (F.6), we obtain

$$\mathbb{P}\left(\sup_{t\in[0,T]} B_N^{\varepsilon}(t) \geq \delta M(N)\right) \leq \frac{4i_0^2+4}{\delta M(N)} \sup_{U\subseteq V_N} \left|\left\{w \in W_N \middle| \frac{|\mathcal{N}_w \cap U|}{|\mathcal{N}_w|} - \frac{|U|}{N}\right| \geq \varepsilon/4i_0\right\}\right|$$

$$+ 2\mathbb{P}\left(\sup_{t\in[0,T]} q_{i_0}(t) \geq \varepsilon/4\right) < \varepsilon'/2 + \varepsilon'/2 = \varepsilon' \quad \text{for all} \quad N \geq \max(N_0, N_1). \quad \square \tag{F.11}$$

Appendix G. Proof of Proposition 2

We now provide the proof of Proposition 2. The proof follows by induction on t.

Proof of Proposition 2. We prove the inequality in Equation (12) by induction on the event times. Assume the inequality holds before the current time epoch t. We continue by distinguishing two cases, depending on whether t is an arrival or a departure epoch. Recall that the G_N-system and $K_{N,M}$-system refer to the systems where the compatibility graphs are G_N and $K_{N,M}$, respectively. Also, recall that, because of the optimal coupling, the servers in both systems are ordered by nondecreasing queue lengths, and departures happen simultaneously at the jth ordered server in the two systems, whenever they are nonempty for $j = 1, 2, \ldots, N$.

First, assume t to be a departure epoch at the jth ordered server. Our goal is to bound the increase of $\sum_{i=1}^{\infty} |Q_i(\Phi(K_{N,M}, t)) - Q_i(\Phi(G_N, t))|$ by zero, because $\Delta_N(t)$ remains unchanged during the departure. Let I_1 and I_2 denote the queue lengths just before time t at the jth ordered server for the G_N-system and $K_{N,M}$-system, respectively. After the departure, both $Q_{I_1}(\Phi(G_N, t))$ and $Q_{I_2}(\Phi(K_{N,M}, t))$ decrease by one, whereas the rest of the terms on the left-hand side of (12) remain unchanged. We will consider three possibilities:

1. If $I_1 = I_2 = i_0$, then Q_{i_0} decreases by one in both systems, and hence the sum in the left-hand side of (12) remains unchanged.
2. Now assume $I_2 < I_1$. This implies that $Q_{I_1}(\Phi(G_N, t-)) > Q_{I_1}(\Phi(K_{N,M}, t-))$, and hence $|Q_{I_1}(\Phi(K_{N,M}, t)) - Q_{I_1}(\Phi(G_N, t))|$ decreases by one during the departure. Because $|Q_{I_2}(\Phi(K_{N,M}, t)) - Q_{I_2}(\Phi(G_N, t))|$ increases by at most one, the sum on the left-hand side of (12) remains unchanged.
3. The case $I_2 > I_1$ is similar to the $I_2 < I_1$ case.

Next, assume t to be an arrival epoch. In this case, we further distinguish two cases:

1. If the arriving tasks are routed to servers with unequal queue lengths in the two systems, then there is a mismatch in queue length at time t and $\Delta_N(t)$ increases by one. Let I_1 and I_2 denote the queue lengths of the servers the tasks are routed to for the G_N-system and $K_{N,M}$-system, respectively. During the arrival, $Q_{I_1+1}(\Phi(G_N, t))$ and $Q_{I_2+1}(\Phi(K_{N,M}, t))$ increase by one, whereas the rest of the terms on the left-hand side of (12) remain unchanged. Hence, $\sum_{i=1}^{\infty} |Q_i(\Phi(K_{N,M}, t)) - Q_i(\Phi(G_N, t))|$ increases by at most two and, because $\Delta_N(t)$ increases by one, the right-hand side of (12) increases by two as well.
2. If the arriving tasks are routed to servers with equal queue lengths in the two systems, then there is no mismatch in queue length at time t and $\Delta_N(t)$ remains unchanged. Therefore, our goal in this case is to bound the increase in the sum $\sum_{i=1}^{\infty} |Q_i(\Phi(K_{N,M}, t)) - Q_i(\Phi(G_N, t))|$ by zero. Let $I_1 = I_2 = i_0$ denote the queue length of the servers the tasks are routed to. During the arrival, Q_{i_0+1} increases by one in both systems, and hence, $|Q_{i_0+1}(\Phi(K_{N,M}, t)) - Q_{i_0+1}(\Phi(G_N, t))|$ remains unchanged. As the rest of the terms on the left-hand side of (12) remain unchanged as well, the increase in the left-hand side of (12) is bounded by zero.

Therefore, in all the above cases, the inequality in (12) is preserved at time epoch t. This completes the proof of Proposition 2. $\square$

Appendix H. Proof of Stability

As mentioned in the introduction, to prove stability and tightness of the steady state occupancy process, we use the Lyapunov function approach, as in Wang et al. [42, 43] and establish moment bounds (Hajek [14], Meyn and Tweedie [19]) to obtain uniform bounds on the tail of the stationary occupancy process. To apply techniques from Meyn and Tweedie [19] in discrete time, we consider the state of the system at *event times* $t_0 = 0 < t_1 < t_k$; that is, for all $i \geq 1$, t_i is either an arrival or a potential departure epoch. The Markov process can now be viewed as a uniformized Markov chain. We will later relate the behavior of this uniformized Markov chain to the behavior of the original process.

For any real-valued function $V : \mathcal{S}_N \to \mathbb{R}$ defined on the state space, denote the expected increase $\Delta V(\cdot)$ as

$$\Delta V(z) := \mathbb{E}[V(\Phi(G_N, t_1)) - V(\Phi(G_N, t_0)) | \Phi(G_N, t_0) = z]. \tag{H.1}$$

We investigate positive Harris recurrence of the uniformized chain by employing the next theorem.

Theorem H.1 (Meyn and Tweedie [19, Theorem 11.0.1]). *Suppose that Φ is a Markov chain. The Markov chain Φ is a positive Harris recurrent chain if and only if there exists some petite set $C \subseteq \mathcal{S}_N$, $b < \infty$ and some nonnegative function $V : \mathcal{S}_N \to \mathbb{R}$ satisfying*

$$\Delta V(z) \leq -1 + b\, \mathbb{1}\{z \in C\} \quad \textit{for all} \quad z \in \mathcal{S}_N. \tag{H.2}$$

The goal is to find an appropriate petite set C and a suitable Lyapunov function V.

The Lyapunov Function

Define a sequence of Lyapunov functions $V_k : \mathcal{S}_N \to \mathbb{R}$ indexed by $k = 1, 2, \dots$ as

$$V_k(z) := \sum_{i=k}^{\infty} \sum_{j=i}^{\infty} Q_j(z) = \frac{1}{2} \sum_{i=k}^{\infty} (i-k+1)(i-k+2) \cdot X_i(z). \tag{H.3}$$

The sequence of Lyapunov functions, instead of a single Lyapunov function, will be necessary to bound the tail of the occupancy of the stationary state. This in turn will establish that the steady state of the occupancy process is tight in the appropriate space.

Lemma H.1. *Consider a sequence $\{G_N\}_{N \geq 1}$ of graphs that satisfies the subcriticality condition. For each $\varepsilon > 0$, there exists $N_0 \geq 1$ such that under the JSQ(d) policy, for all $N \geq N_0$,*

$$\Delta V_k(z) \leq \frac{1}{\lambda + 1} \left((1+\varepsilon)\lambda q_{k-1}(z) - (1 - (1+\varepsilon)\lambda) \sum_{i=k}^{\infty} q_i(z) \right) \quad \text{for all } z \in \mathcal{S}_N. \tag{H.4}$$

Note that because of Lemma H.1, if the tail of the occupancy is heavy compared with $q_{k-1}(z)$, then the increase of the Lyapunov function is negative. In other words, if queues are long, a task is more likely to join a shorter queue than to join one of these long queues. This is a consequence of the JSQ(d) policy, which prefers shorter queues. The subcriticality condition on the graph is necessary here to ensure tasks are able to reach the shorter queues.

Proof of Lemma H.1. By conveniently employing Poisson thinning as an argument, we can formulate an alternative description of the queueing system. Tasks in the system arrive as a Poisson process with rate λN. At the arrival of a task, a dispatcher $w \in W_N$ is uniformly chosen at random. Next, servers $v_1, \dots, v_d \in \mathcal{N}_w$ are selected uniformly at random with replacement. The task is then routed to the server with the shortest queue out of $v_1, \dots, v_d$. In this formulation, the probability of a task being routed to the subset of servers with at least i tasks, $\mathcal{Q}_i$, is

$$\frac{1}{M(N)} \sum_{w \in W_N} |\mathcal{N}_w|^{-d} \sum_{v_1, \dots, v_d \in \mathcal{N}_w} \left(\mathbb{1}\{\{v_1, \dots, v_d\} \subseteq \mathcal{Q}_i\} \cdot 1 + (1 - \mathbb{1}\{\{v_1 \dots, v_d\} \subseteq \mathcal{Q}_i\}) \cdot 0 \right). \tag{H.5}$$

Because the graph sequence satisfies the subcriticality condition, we know there exist a probability distribution $\gamma_w^{v_1, \dots, v_d}(v)$ on V_N supported on $\{v_1, \dots, v_d\}$ for each $w \in W_N$ and $v_1, \dots, v_d \in V_N$ such that

$$\max_{v \in V_N} \frac{N}{M(N)} \sum_{w \in W_N} |\mathcal{N}_w|^{-d} \sum_{v_1, \dots, v_d \in \mathcal{N}_w} \gamma_w^{v_1, \dots, v_d}(v) \leq 1 + \epsilon, \tag{H.6}$$

for N large enough. If $\{v_1, \dots, v_d\} \subseteq \mathcal{Q}_i$, then

$$\sum_{v \in \mathcal{Q}_i} \gamma_w^{v_1, \dots, v_d}(v) = \sum_{v \in \{v_1, \dots, v_d\}} \gamma_w^{v_1, \dots, v_d}(v) = 1, \tag{H.7}$$

and hence,

$$\begin{aligned}
&\frac{1}{M(N)} \sum_{w \in W_N} |\mathcal{N}_w|^{-d} \sum_{v_1, \dots, v_d \in \mathcal{N}_w} \left(\mathbb{1}\{\{v_1, \dots, v_d\} \subseteq \mathcal{Q}_i\} \cdot 1 + (1 - \mathbb{1}\{\{v_1, \dots, v_d\} \subseteq \mathcal{Q}_i\}) \cdot 0 \right) \\
&\quad \leq \frac{1}{M(N)} \sum_{w \in W_N} |\mathcal{N}_w|^{-d} \sum_{v_1, \dots, v_d \in \mathcal{N}_w} \sum_{v \in \mathcal{Q}_i} \gamma_w^{v_1, \dots, v_d}(v) \\
&\quad = \sum_{v \in \mathcal{Q}_i} \frac{1}{M(N)} \sum_{w \in W_N} |\mathcal{N}_w|^{-d} \sum_{v_1, \dots, v_d \in \mathcal{N}_w} \gamma_w^{v_1, \dots, v_d}(v) \leq \sum_{v \in \mathcal{Q}_i} \frac{1+\varepsilon}{N} = (1+\varepsilon) q_i,
\end{aligned} \tag{H.8}$$

where the last inequality follows from the subcriticality condition for N large enough. The increase of the Lyapunov function is then bounded as

$$\begin{aligned}
\Delta V_k(z) &:= \mathbb{E}[V_k(\Phi(G_N, t_1)) - V_k(\Phi(G_N, t_0)) | \Phi_{t_0} = z] \\
&= \sum_{i=k}^{\infty} \mathbb{E}\left[\sum_{j=i}^{\infty} \Big(Q_j(\Phi(G_N, t_1)) - Q_j(\Phi(G_N, t_0)) \Big) \middle| \Phi_{t_0} = z \right] \\
&= \sum_{i=k}^{\infty} \left(\frac{\lambda N}{\lambda N + N} \mathbb{P}(\text{task routed to } \mathcal{Q}_{i-1}(z)) - \frac{N}{\lambda N + N} \frac{Q_i(z)}{N} \right) \\
&\leq \frac{1}{\lambda + 1} \sum_{i=k}^{\infty} \big((1+\varepsilon)\lambda q_{i-1}(z) - q_i(z) \big) \\
&= \frac{1}{\lambda + 1} \left((1+\varepsilon)\lambda q_{k-1}(z) - (1 - (1+\varepsilon)\lambda) \sum_{i=k}^{\infty} q_i(z) \right). \quad \square
\end{aligned} \tag{H.9}$$

Proof of Positive Harris Recurrence

We will gradually progress toward proving positive Harris recurrence using Theorem H.1, starting with the definition of an appropriate petite set. The set $C_\Gamma \subseteq \mathcal{S}_N$ is defined as

$$C_\Gamma := \left\{ z \in \mathcal{S}_N \middle| \sum_{i=1}^{\infty} Q_i(z) \le \Gamma \right\}. \tag{H.10}$$

We formally state the definition of a petite set below.

Definition H.1 (Petite Set). A set $C \subseteq \mathcal{S}_N$ is a *petite set* if and only if there exists a distribution $a = \{a(n)\}$ on $\mathbb{N}$ and a nontrivial measure μ on $\mathcal{S}_N$ such that

$$\sum_{n=0}^{\infty} P^n(z, A) a(n) \ge \mu(A) \tag{H.11}$$

for all $z \in C$ and $A \subseteq \mathcal{S}_N$, where $P^n(z, A)$ is the n-step transition probability from z to A.

Lemma H.2 states that C_Γ introduced in (H.10) is a petite set. The proof is fairly straightforward because C_Γ is finite, however, we provide it for the sake of completeness.

Lemma H.2. *The set $C_\Gamma \subseteq \mathcal{S}_N$ defined as in (H.10) is petite for any $\Gamma \ge 0$.*

Proof. To establish petiteness, it is sufficient to show that there exists a nontrivial measure μ on $\mathcal{S}_N$ such that

$$\sum_{n=1}^{\infty} e^{-n} P^n(z, A) \ge \mu(A), \tag{H.12}$$

for all $z \in C_\Gamma$, $A \subseteq \mathcal{S}_N$, where $P^n(z, A)$ is the n-step transition probability from z to A. Define μ as

$$\mu(a) := \min_{z \in C_\Gamma} \sum_{n=1}^{\infty} e^{-n} P^n(z, a) \text{ for } a \in \mathcal{S}_N, \tag{H.13}$$

and let $\mu(A) = \sum_{a \in A} \mu(a)$, which is well defined because the state space is countable. Trivially, Equation (H.12) is satisfied. Moreover, because each state $a \in \mathcal{S}_N$ is reachable and $|C_\Gamma| < \infty$, the minimum will be nonzero and the measure is nontrivial. □

The next lemma establishes that the Lyapunov function defined in the previous section satisfies the condition of Theorem H.1.

Lemma H.3. *Consider a sequence $\{G_N\}_{N \ge 1}$ of graphs that satisfies the subcriticality condition. There exists $N_0 \ge 1$ such that for the JSQ(d) policy, the function V_1 as defined in Equation (H.3) satisfies*

$$\Delta V_1(z) \le -1 + 2\, \mathbb{1}\{z \in C\} \quad \text{for all} \quad z \in \mathcal{S}_N, \tag{H.14}$$

for $C = C_\Gamma$ with $\Gamma = 3\frac{1+\lambda}{1-\lambda} N$ and all $N \ge N_0$.

Choose ε small such that $(1+\varepsilon)\lambda \le (\lambda+1)/2 < 1$. By Lemma H.1, we know that

$$\Delta V_1(z) \le \frac{1}{\lambda+1} \left((1+\varepsilon)\lambda - (1-(1+\varepsilon)\lambda) \sum_{i=1}^{\infty} q_i(z) \right), \tag{H.15}$$

for N large enough. Then, for $z \in \mathcal{S}_N$,

$$\Delta V_1(z) \le \frac{(1+\varepsilon)\lambda}{\lambda+1} \le \frac{1}{2} \frac{\lambda+1}{\lambda+1} < -1 + 2. \tag{H.16}$$

Also, for $z \notin C_\Gamma$ with $\Gamma = 3\frac{1+\lambda}{1-\lambda} N$,

$$\Delta V_1(z) \le \frac{(1+\varepsilon)\lambda}{1+\lambda} - \frac{1-(1+\varepsilon)\lambda}{1+\lambda} \sum_{i=1}^{\infty} q_i(z) < \frac{1}{2} \frac{1+\lambda}{1+\lambda} - \frac{1}{2} \frac{1-\lambda}{1+\lambda} \frac{\Gamma}{N} \le -1. \tag{H.17}$$

Lemmas H.2 and H.3 together with Theorem H.1 imply positive Harris recurrence of the uniformized Markov chain. This implies that also the continuous time chain is positive Harris recurrent and has a stationary distribution.

Proof of Moment Bound

Finally, we prove the moment bound in Lemma 3.

Proof of Lemma 3. By the definition of the steady state,

$$\mathbb{E}[\Delta V_k(\Phi(G_N, \infty))] = 0. \tag{H.18}$$

Choose ε small such that $(1+\varepsilon)\lambda \le (1+\lambda)/2 < 1$. Using Lemma H.1, we conclude for N large enough that

$$\mathbb{E}\left[(1+\varepsilon)\lambda q_{k-1}(z) - (1-(1+\varepsilon)\lambda)\sum_{i=k}^{\infty} q_i(z)\right] \ge 0, \tag{H.19}$$

which can be rewritten to show the lemma. □

Appendix I. Criteria for Tightness in the ℓ_1^{ω}-Topology

Proof of Lemma 4. Define the map $F : \ell_1 \to \ell_1^{\omega}$ as

$$F(\boldsymbol{s}) = \left(\frac{s_1}{\omega_1}, \frac{s_2}{\omega_2}, \dots\right). \tag{I.1}$$

The map F is a bijective isometry as for $\boldsymbol{s}, \boldsymbol{t} \in \ell_1$,

$$|F(\mathbf{s}) - F(\mathbf{t})|_1^{\omega} = \sum_{i=1}^{\infty} \omega_i \left|\frac{s_i}{\omega_i} - \frac{t_i}{\omega_i}\right| = \sum_{i=1}^{\infty} |s_i - t_i| = \|\mathbf{s} - \mathbf{t}\|_1. \tag{I.2}$$

For convenience, we denote $\boldsymbol{\omega} \cdot \boldsymbol{s} := F^{-1}(\boldsymbol{s})$. We first claim that the tightness of the sequence $\{\boldsymbol{q}(\Phi(G_N, \infty))\}_{N\ge1}$ in ℓ_1^{ω} is equivalent to tightness of $\{\boldsymbol{\omega} \cdot \boldsymbol{q}(\Phi(G_N, \infty))\}_{N\ge1}$ in ℓ_1. Indeed, if $\{\boldsymbol{\omega} \cdot \boldsymbol{q}(\Phi(G_N, \infty))\}_{N\ge1}$ is tight in ℓ_1, then we can find a compact set $K \subseteq \ell_1$ such that

$$\mathbb{P}(\boldsymbol{q}(\Phi(G_N, \infty)) \notin F(K)) = \mathbb{P}(\boldsymbol{\omega}\boldsymbol{q}(\Phi(G_N, \infty)) \notin K) < \epsilon. \tag{I.3}$$

As the mapping preserves compactness due to continuity, $F(K)$ is a compact set in ℓ_1^{ω}, and hence $\{\boldsymbol{q}(\Phi(G_N, \infty))\}_{N\ge1}$ is tight in ℓ_1^{ω}. Conversely, if $\{\boldsymbol{q}(\Phi(G_N, \infty))\}_{N\ge1}$ is tight in ℓ_1^{ω}, then there is a compact set $K \subseteq \ell_1^{\omega}$ such that Equation (I.3) holds and $F^{-1}(K)$ is compact in ℓ_1. Hence, the claim is proved.

Now, by lemma 2 of Mukherjee et al. [24], tightness in ℓ_1 is equivalent to showing tightness with respect to the product topology and (24). This completes the proof of Lemma 4. □

References

[1] Arapostathis A, Hmedi H, Pang G (2021) On uniform exponential ergodicity of Markovian multiclass many-server queues in the Halfin–Whitt regime. *Math. Oper. Res.* 46(2):772–796.
[2] Braess D (1968) Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung* 12(1):258–268.
[3] Bramson M (2011) Stability of join the shortest queue networks. *Ann. Appl. Probab.* 21(4):1568–1625.
[4] Budhiraja A, Mukherjee D, Wu R (2019) Supermarket model on graphs. *Ann. Appl. Probab.* 29(3):1740–1777.
[5] Cardinaels E, Borst SC, van Leeuwaarden JSH (2019) Job assignment in large-scale service systems with affinity relations. Working paper, Eindhoven University of Technology, Eindhoven, Netherlands.
[6] Cardinaels E, Borst S, van Leeuwaarden JSH (2020) Redundancy scheduling with locally stable compatibility graphs. Working paper, Eindhoven University of Technology, Eindhoven, Netherlands.
[7] Chung FRK, Graham RL, Wilson RM (1989) Quasi-random graphs. *Combinatorica* 9(4):345–362.
[8] Cruise J, Jonckheere M, Shneer S (2020) Stability of JSQ in queues with general server-job class compatibilities. *Queueing Systems* 95(3–4):271–279.
[9] Deimling K (2006) *Ordinary Differential Equations in Banach Spaces*. Lecture Notes in Mathematics, vol. 596 (Springer, Berlin).
[10] Ethier SN, Kurtz TG (2009) *Markov Processes: Characterization and Convergence* (John Wiley & Sons, Hoboken, NJ).
[11] Foss SG, Chernova NI (1998) On the stability of a partially accessible multi-station queue with state-dependent routing. *Queueing Systems* 29(1):55–73.
[12] Gamarnik D, Stolyar AL (2012) Multiclass multiserver queueing system in the Halfin–Whitt heavy traffic regime: Asymptotics of the stationary distribution. *Queueing Systems* 71(1–2):25–51.
[13] Gast N (2015) The power of two choices on graphs: The pair-approximation is accurate. Squillante M (Workshop chair), ed. *Proc. Workshop Math. Performance Model. Anal.* (Association for Computing Machinery, New York), 69–71.
[14] Hajek B (1982) Hitting-time and occupation-time bounds implied by drift analysis with applications. *Adv. Appl. Probab.* 14(3):502–525.
[15] He YT, Down DG (2008) Limited choice and locality considerations for load balancing. *Performance Evaluation* 65(9):670–687.
[16] Hmedi H, Arapostathis A, Pang G (2019) Uniform stability of a class of large-scale parallel server networks. Working paper, University of Texas at Austin.
[17] Janson S, Luczak T, Rucinski A (2000) *Random Graphs* (John Wiley & Sons, Hoboken, NJ).
[18] Kurtz TG (1981) *Approximation of Population Processes* (SIAM, Philadelphia).
[19] Meyn SP, Tweedie RL (1993) *Markov Chains and Stochastic Stability* (Springer, London).
[20] Mishra AK, Hellerstein JL, Cirne W, Das CR (2010) Toward characterizing cloud backend workloads: Insights from Google compute clusters. *Performance Evaluation Rev.* 37(4):34–41.
[21] Mitzenmacher M (1996) The power of two choices in randomized load balancing. Unpublished PhD thesis, University of California, Berkeley.
[22] Mitzenmacher M (2001) The power of two choices in randomized load balancing. *IEEE Trans. Parallel Distributed Systems* 12(10):1094–1104.
[23] Mukherjee D, Borst SC, Van Leeuwaarden JSH (2018) Asymptotically optimal load balancing topologies. Chaintreau A, Golubchik L, Zhang Z-L, eds. *Proc. ACM Measurement Anal. Comput. Systems* (Association for Computing Machinery, New York), 14.
[24] Mukherjee D, Borst SC, van Leeuwaarden JSH, Whiting PA (2018) Universality of power-of-d load balancing in many-server systems. *Stochastic Systems* 8(4):265–292.

[25] Pang G, Talreja R, Whitt W (2007) Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surveys* 4:193–267.
[26] Reiss C, Tumanov A, Ganger GR, Katz RH, Kozuch MA (2012) Heterogeneity and dynamicity of clouds at scale: Google trace analysis. Carey M (Program Chairs), Hand S, eds. *Proc. Third ACM Sympos. Cloud Comput.* (Association for Computing Machinery, New York), 7.
[27] Roughgarden T, Tardos E (2002) How bad is selfish routing? *J. ACM* 49(2):236–259.
[28] Stolyar AL (1995) On the stability of multiclass queueing networks: A relaxed sufficient condition via limiting fluid processes. *Markov Processes Related Fields* 1(4):491–512.
[29] Stolyar AL (2005) Optimal routing in output-queued flexible server systems. *Probab. Engrg. Inform. Sci.* 19(2):141–189.
[30] Stolyar AL (2017) Pull-based load distribution among heterogeneous parallel servers: The case of multiple routers. *Queueing Systems* 85(1):31–65.
[31] Tang D, Subramanian VG (2019) Derandomized load balancing using random walks on expander graphs. Working paper, University of California, Berkeley.
[32] Tang D, Subramanian VG (2019) Random walk based sampling for load balancing in multi-server systems. Chaintreau A, Golubchik L, Zhang Z-L, eds. *Proc. ACM Measurement Anal. Comput. Systems* (Association for Computing Machinery, New York), 14.
[33] Tsitsiklis JN, Xu K (2013) On the power of (even a little) resource pooling. *Stochastic Systems* 2(1):1–66.
[34] Tsitsiklis JN, Xu K (2013) Queueing system topologies with limited flexibility. Harchol-Balter M (General Chair), ed. *Proc. ACM SIGMETRICS Internat. Conf. Measurement Model. Comput. Systems* (Association for Computing Machinery, New York), 167–178.
[35] Tsitsiklis JN, Xu K (2017) Flexible queueing architectures. *Oper. Res.* 65(5):1398–1413.
[36] Turner SR (1998) The effect of increasing routing choice on resource pooling. *Probab. Engrg. Inform. Sci.* 12(1):109–124.
[37] Van der Boor M, Borst SC, van Leeuwaarden JSH (2017) Load balancing in large-scale systems with multiple dispatchers. Akan OB (General Chairs), Sivakumar RS, eds. *Proc. IEEE Conf. Comput. Comm.* (Institute of Electrical and Electronics Engineers, Piscataway, NJ), 1–9.
[38] Van der Boor M, Borst SC, Van Leeuwaarden JSH, Mukherjee D (2018) Scalable load balancing in networked systems: Universality properties and stochastic coupling methods. Sirakov B, de Souza PN, Viana M, eds. *Proc. Internat. Congress Mathematicians* (World Scientific, Singapore), 3893–3923.
[39] Van der Hofstad R (2017) *Random Graphs and Complex Networks*, vol. 1 (Cambridge University Press, Cambridge, UK).
[40] Vargaftik S, Keslassy I, Orda A (2020) LSQ: Load balancing in large-scale heterogeneous systems with multiple dispatchers. *IEEE/ACM Trans. Networking* 28(3):1186–1198.
[41] Vvedenskaya ND, Dobrushin RL, Karpelevich FI (1996) Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii* 32(1):20–34.
[42] Wang W, Maguluri ST, Srikant R, Ying L (2018) Heavy-traffic delay insensitivity in connection-level models of data transfer with proportionally fair bandwidth sharing. *ACM SIGMETRICS Performance Evaluation Rev.* 45(3):232–245.
[43] Wang W, Maguluri ST, Srikant R, Ying L (2018) Heavy-traffic insensitive bounds for weighted proportionally fair bandwidth sharing policies. Preprint, submitted August 6, https://arxiv.org/abs/1808.02120.
[44] Wang W, Zhu K, Ying L, Tan J, Zhang L (2016) MapTask scheduling in MapReduce with data locality: Throughput and heavy-traffic optimality. *IEEE/ACM Trans. Networking* 24(1):190–203.
[45] Weng W, Zhou X, Srikant R (2020) Optimal load balancing with locality constraints. Chaintreau A, Golubchik L, Zhang Z-L, eds. *Proc. ACM Measurement Anal. Comput. Systems* (Association for Computing Machinery, New York), 45.
[46] Xie Q, Yekkehkhany A, Lu Y (2016) Scheduling with multi-level data locality: Throughput and heavy-traffic optimality. Song M (General Chair), Westphal C (General Chair), eds. *Proc. 35th Annual IEEE Internat. Conf. Comput. Comm.* (Institute of Electrical and Electronics Engineers, Piscataway, NJ), 1–9.
[47] Yekkehkhany A, Nagi R (2020) Blind GB-PANDAS: A blind throughput-optimal load balancing algorithm for affinity scheduling. *IEEE/ACM Trans. Networking* 28(3):1199–1212.
[48] Yekkehkhany A, Hojjati A, Hajiesmaili MH (2018) GB-PANDAS: Throughput and heavy-traffic optimality analysis for affinity scheduling. *Performance Evaluation Rev.* 45(3):2–14.
[49] Zhou X, Shroff N, Wierman A (2021) Asymptotically optimal load balancing in large-scale heterogeneous systems with multiple dispatchers. *Perform. Eval.* 145:102146.