CORRECTED VERSION OF RECORD; SEE LAST PAGE OF ARTICLE

informs

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

MANAGEMENT SCIENCE

Vol. 66, No. 10, October 2020, pp. 4667–4685

ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Online Demand Fulfillment Under Limited Flexibility

Zhen Xu,[a] **Hailun Zhang,**[b,*] **Jiheng Zhang,**[c] **Rachel Q. Zhang**[c]

[a] School of Management, Fudan University, Shanghai 200433, China; [b] Institute for Data and Decision Analytics, The Chinese University of Hong Kong, Shenzhen 518172, China; [c] Department of Industrial Engineering & Decision Analytics, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong S.A.R., China

*Corresponding author

Contact: zhen_xu@fudan.edu.cn, https://orcid.org/0000-0002-7201-3383 (ZX); zhanghailun@cuhk.edu.cn, https://orcid.org/0000-0001-6116-6168 (HZ); jiheng@ust.hk, https://orcid.org/0000-0003-3025-1495 (JZ); rzhang@ust.hk (RQZ)

Received: June 16, 2018
Revised: December 23, 2018; June 20, 2019
Accepted: July 8, 2019
Published Online in Articles in Advance: June 12, 2020

https://doi.org/10.1287/mnsc.2019.3449

Abstract. We study online demand fulfillment in a class of networks with limited flexibility and arbitrary numbers of resources and request types. We show analytically that such a network is both necessary and sufficient to guarantee a performance gap independent of the market size compared with networks with full flexibility, extending the previous literature from the long chains to more general sparse networks. Inspired by the performance bound, we develop simple inventory allocation rules and guidelines for designing such network structures. Numerical experiments including one using some real data from Amazon China are conducted to confirm our findings as well as some of the flexibility principles conjectured in the literature.

History: Accepted by Chung Piaw Teo, optimization.

Funding: This work was supported by the Hong Kong Research Grants Council, University Grants Committee [Grants 16200019, 16200617, 16201417, and 16501015], and the Shenzhen Research Institute of Big Data and Shenzhen Peacock Plan [Grant KQTD2015033114415450].

Supplemental Material: The e-companion is available at https://doi.org/10.1287/mnsc.2019.3449.

Keywords: process flexibility • online retailing • dynamic resource allocation

1. Introduction

E-commerce is an exciting trend for many traditional industries, with the rising number of global Internet and smartphone users driving its growth. According to Statista, retail e-commerce sales worldwide amounted to $2.3 trillion in 2017 and e-retail revenues are projected to grow to $4.88 trillion in 2021. In particular, online sales in China accounted for 23% of total retail sales in 2017.[1] A key challenge for any online retailer is how to efficiently fulfill a large number of customer orders from different geographical locations as they arrive using its existing distribution resources. Order fulfillment at an online retailer such as Amazon China, which has a total of 12 distribution centers serving customers from hundreds of regions throughout China, involves three important decisions: (1) the distribution network, that is, which distribution center(s) can cater the demand in a given geographical region (e.g., a city or a province); (2) inventory allocation, that is, for a given distribution network, how inventory should be allocated to each distribution center; and (3) dynamic demand fulfillment rule, that is, given the network structure and actual inventory levels, which distribution center should fulfill an arriving customer order. Online demand fulfillment implies that the fulfillment decision must be made upon the arrival of each order and is irrevocable.

We now elaborate on these three decisions, how they are related and why they are difficult. The inventory allocation and dynamic fulfillment decisions are closely related to the network structure. Intuitively, for a given network structure, more inventory should be allocated to distribution centers that are serving large customer bases, and distribution centers with higher remaining inventory and smaller customer bases should be used to fulfill arriving customer orders. However, finding an optimal inventory allocation and dynamic fulfillment policy requires solving a high-dimensional dynamic programming and is analytically intractable for most real systems.

For given distribution centers and demand regions, there are many choices for designing a distribution network. A straightforward solution is to group a few nearby demand regions together, and dedicate one distribution center exclusively for this group, making the network a collection of nonoverlapping star-shaped components. It is simple and easy to operate, but it is suboptimal as some groups may starve while others have lots of leftover. Alternatively, we can make inventory at multiple distribution centers available to one demand location. Such flexibility likely will benefit the fulfillment if implemented well, though additional costs, for example, shipping and operational costs, may incur. At the extreme are systems with full flexibility if every distribution center is allowed to fulfill orders from all locations. Such systems can best match supply with demand and thus maximize sales; however, they are likely too expensive to operate, and hence impractical. The study of network structure is closely related to the process flexibility literature that dates back to the seminal work by Jordan and Graves (1995), which has shown that a little bit of flexibility in the form of the long chain structure goes a long way. However, most research along this line has focused on production systems, where the fulfillment decision is made either periodically or after all the demand is observed, that is, demand is fulfilled offline. The only exception is the work by Asadpour et al. (2019), which focuses on online demand fulfillment under the long chain structure, which is a balanced system with equal numbers of distribution centers and demand locations. However, most real e-retail systems, including Amazon China are unbalanced, making the long chain structure not applicable. Research is needed to study the design of network structures for general systems that need to fulfill customer orders as they arrive.

Although our study is motivated by online retailing, similar problems also arise from other application models where, for example, a few machines or service representatives can process many different jobs as they arrive over time and the jobs need to be assigned to the machines or representatives immediately upon arrival. We model our network structure using a bipartite graph with J types of requests (demand locations) and I types of resources (distribution centers). A resource type can serve a request type if there is an arc between them in the bipartite graph. The performance measure we are concerned with is the expected number of lost sales. We consider a class of connected networks with a positive generalized chaining gap (GCG), referred to as GCG systems and first introduced by Shi et al. (2019). Intuitively, a positive GCG means there is slack between the total (expected) demand for any subset of request types and the total supply at the resources that can serve these requests. The GCG measures the extent to which supply dominates demand and hence serves as an important indicator of the system performance under online fulfillment. For order fulfillment, we generalize the modified greedy policy in Asadpour et al. (2019). Our main contributions can be summarized as follows.

1. Asadpour et al. (2019) showed bounded performance of the long chains where $I = J$ as the market size increases when demand is fulfilled online. By bounded performance we mean that the expected number of lost sales does not diverge with the market size, which is desirable when a retailer faces large customer demand in practice. Under the same online setting, we show that a positive GCG is both necessary and sufficient for achieving bounded performance for general network structures with arbitrary I and J. Furthermore, we establish the tightness of our performance bound that is inversely proportional to the GCG. This not only extends the result in Asadpour et al. (2019) that the performance bounds are independent of the market size to general network structures, but also achieves a tighter bound when applied to the long chains. Our results are established based on a novel proof, making our work a methodologically and managerially significant contribution.

2. The performance bound inspires us to use the GCG as a proxy for the system performance when making the inventory allocation and network design decisions.

(a) We show that any connected network structure with as few as $I + J - 1$ arcs, much less than $I \times J$ arcs under full flexibility, is guaranteed to be a GCG system under our inventory allocation policies. Furthermore, such limited flexibility can achieve near-optimal performance. So our work has extended toward the online direction of the work by Shi et al. (2019), which showed that a GCG production system can achieve near-optimal performance if demand is fulfilled periodically and the capacity utilization is close to one.

(b) We develop simple principles for guiding the design of GCG networks or adding arcs to an existing GCG network in practice. First, we show that GCG networks with as few as $I + J - 1$ arcs can perform quite well and are excellent options if it is expensive to add one more arc. If we are allowed to add one more arc to a connected network with $I + J - 1$ arcs, there will be a cycle in the network and the additional arc should be added to form the largest cycle possible, which is consistent with the observation made in Jordan and Graves (1995). Second, with $I + J$ arcs, we can simply divide the resources into I groups and form a generalized long chain (GLC) with I resources and I request groups. If we have the option to add an arc to a network structure with at least $I + J$ arcs, the arc should be added to strengthen the weakest link in the exiting network.

3. Numerical studies including experiments using some real data from Amazon China are performed to confirm and verify our main findings. For instance, flexible systems that fulfill requests online may incur additional shipping costs compared with dedicated ones. We demonstrate that a little flexibility improves the performance significantly without increasing the total shipping costs too much for systems with a positive GCG.

The paper is organized as follows. After a brief literature review in Section 2, we present our detailed model in Section 3 and introduce the GCG and dynamic fulfillment policy in Section 4. We establish a performance bound for systems with a positive GCG that is independent of the market size in Section 5. Based on the insights from the bound, we derive some important principles for the inventory allocation decision and network structure design in Section 6. We extend bounded performance to systems with random batch arrivals and time varying arrivals in Section 7. We conduct numerical studies including one using some data from Amazon China in Sections 8 and 9. The paper is concluded in Section 10. The proofs of lemmas and theorems can be found in the e-companion.

2. Literature Review

The study of process flexibility structures dates back to the seminal work of Jordan and Graves (1995), who observe that a sparse chaining flexibility structure often achieves almost the same performance as the fully flexible system. Motivated by their empirical findings, most theoretical work since then has focused on explaining the power of chaining for balanced systems, with a few exceptions for unbalanced systems.

The effectiveness of a long chain in balanced systems, that is, $I = J$, compared with the effectiveness of fully flexible systems, has been investigated extensively. Chou et al. (2010) derive the ratio of the performance of the long chain to that of the full flexibility system when the system size approaches infinity and show that, for certain demand distributions, the ratio is very close to 1. Simchi-Levi and Wei (2012) consider the benefit of adding each arc as one constructs a long chain from a dedicated system and show that the benefit increases under the assumption that the request types are interchangeable, implying that the biggest benefit is always achieved when the last arc closes the chain. Furthermore, they establish that the long chain maximizes the expected sales among all flexibility designs where each node is incident to exactly two arcs. Wang and Zhang (2015) obtain a bound on the asymptotic performance of the long chain that only depends on the mean and variance of the demand distribution. Recently, Désir et al. (2016) prove the optimality of the long chain among all connected structures with the same number of arcs. Research has also been conducted on the long chain using a graph expander. Chou et al. (2011) prove that there exists a sparse graph that can achieve $(1-\epsilon)$ of the sales of a fully flexible system in the worst-case demand scenario. Chen et al. (2015) use the probabilistic expander, that is, the probability that an arc linking a supply node and a demand node is proportional to the product of their capacity and demand, to derive a theoretical bound on the number of arcs required to achieve $(1-\epsilon)$ performance of full flexibility in a symmetric system.

For unbalanced systems, that is, $I \neq J$, for which the long chain concept does not apply, analytical results are difficult to obtain and much effort has been devoted to developing flexibility design indices to measure the effectiveness of different flexibility structures starting from the JG index proposed by Jordan and Graves (1995). Other indices include the structural flexibility index in Iravani et al. (2005), the WS-APL index in Iravani et al. (2007), the g-measure in Graves and Tomlin (2003), the expansion index in Chou et al. (2008), and the plant cover index in Simchi-Levi and Wei (2015). Deng (2013) offers detailed descriptions of these indices. Researchers have also studied unbalanced systems from other perspectives. Shen and Deng (2013) propose flexibility design guidelines for symmetric demand via simulation and refined the well-known chaining guidelines if each product is manufactured at exactly two plants. Chen et al. (2019) construct a simple flexibility design to fulfill $(1-\epsilon)$ fraction of the expected total demand with high probability with an average degree of $O(\ln(1/\epsilon))$ using a probabilistic expander. Tanrisever et al. (2012) evaluate the effectiveness of different flexibility structures by simulation under a feasible production scheduling policy obtained using a sampling-based decomposition method in a multiperiod setting. Sheng et al. (2015) consider capacity portfolio investment on flexible machines and show that, under certain conditions, the optimal flexibility configuration consists only of dedicated machines and machines capable of building only two types of products. Simchi-Levi et al. (2018) study the synergy between inventory and process flexibility by considering a two-stage robust optimization problem, and use inventory allocation to mitigate demand disruption in the first stage.

The work of Shi et al. (2019) is a recent breakthrough in the theoretical study of unbalanced systems. They introduce the GCG to identify effective flexibility structures. For production systems with a positive GCG, which is essentially the complete resource pooling condition (CRP) in the queueing literature (see Mandelbaum and Stolyar 2004 and references therein), they obtain an upper bound on the long-run average backlog cost under a max-weight fulfillment policy. The upper bound theoretically demonstrates that, when capacity utilization is high, the performance of a system with a positive GCG is almost the same as that of a fully flexible structure. For a given capacity profile, they also provide a simple and efficient algorithm for finding such sparse structures and show that the requirement of $I + J$ arcs is tight in general for a given GCG system. In contrast, we treat the inventory allocation as a decision and show that $I + J - 1$ arcs is sufficient to achieve bounded performance under our inventory allocation policy that guarantees a positive GCG and dynamic fulfillment policy. Under a different setting, Ding et al. (2018) show that the CRP provides a necessary and sufficient condition for global first-come, first-served in an overloaded bipartite queueing system without customer reneging.

Asadpour et al. (2019) are the first to study the performance of the long chain structure when demand is fulfilled as it arrives or online. Under a so-called ξ-Hall condition, they show bounded performance of the long chain structure. In this paper, we extend Asadpour et al. (2019) to general systems and show bounded performance for systems with a positive GCG when demand is stationary. We also discuss conditions under which bounded performance is guaranteed when demand is time varying. Our bounds are tighter than that in Asadpour et al. (2019) for the long chains.

Process flexibility has also been studied in various areas, such as limited labor cross-training in call centers (Wallace and Whitt 2005), resource portfolio investment (Bassamboo et al. 2010), and queueing networks (Gurumurthi and Benjaafar 2004, Tsitsiklis and Xu 2017). Since flexibility has the potential to increase shipping costs, research on how to develop order fulfillment policies for online retailers in order to minimize the total outbound shipping costs, for example, Xu et al. (2009) and Jasin and Sinha (2015), is also relevant.

Our work is also related to the broad class of dynamic resource allocation problems that require irrevocable decisions to be made as requests arrive sequentially. One approach to coping with sequential arrivals is to utilize approximate dynamic programming techniques, which produce tractable solutions that often exhibit satisfactory performance in practice (e.g., see Van Roy et al. 1997). On a more general level, our problem can also be viewed as an online stochastic matching and dynamic matching problem. For more information on online stochastic matching, see Feldman et al. (2009), Manshadi et al. (2012), and Jaillet and Lu (2013). Bušić et al. (2013) and Bušić and Meyn (2015) study dynamic matching problems where there is exactly one request and one supply in each period, and propose near-optimal policies to minimize the infinite-horizon average-cost.

3. Model Formulation

We consider a system with I resources and a total of K units of initial inventory for a whole selling season. Requests for inventory arrive sequentially and need to be fulfilled immediately. Requests that cannot be fulfilled are lost. We do not take into account inventory holding costs and discounting factors. Thus, the interarrival times do not matter and we only need to know the inventory profile after each arrival. Upon arrival, each request is revealed to be of type j with probability $p_j > 0$ and there is a total of J request types. We assume that $I \le J$ as in most real applications. We refer to $\mathbf{p} = (p_1, \ldots, p_J)$ as the demand vector, and define $\min_{\le j \le J}\{p_j\} \triangleq p_{\min}$ and $\max_{\le j \le J}\{p_j\} \triangleq p_{\max}$. We assume that I, J, and $\mathbf{p}$ are all given and fixed. We now describe the details of the system and operational decisions.

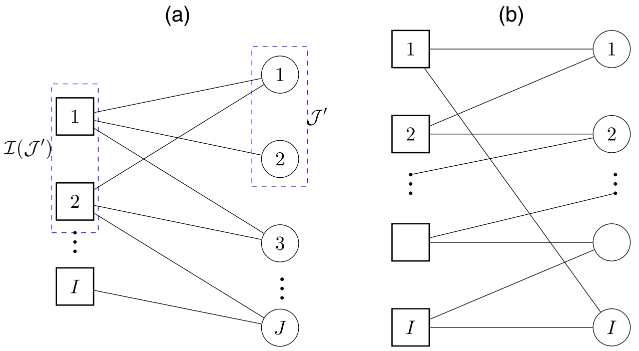
1. The flexibility structure. The flexibility structure we are concerned with can be modeled as a bipartite graph $\mathscr{A} = (\mathscr{I}, \mathscr{J}, E)$, where $\mathscr{I} = \{1, 2, \ldots, I\}$ is the set of resources, $\mathscr{J} = \{1, 2, \ldots, J\}$ is the set of request types, and E is the set of all the arcs in the network. An arc $(i, j) \in E$ if resource i is capable of fulfilling request type j. A structure has full flexibility if $\mathscr{A}$ is a complete bipartite graph with $I \times J$ arcs, that is, each resource can serve all request types, whereas the well-known long chain structure where $I = J$ has $I + J = 2I$ arcs. Figure 1(a) provides a general network structure and Figure 1(b) illustrates a long chain.

2. Inventory allocation. For a given total amount of inventory K, let $c_i K$, where $\sum_{i=1}^{I} c_i = 1$, be the amount of inventory allocated to resource i. Although almost all existing research on process flexibility assumes that the initial inventory or capacity is given and not a decision to be made, we treat inventory allocation $\mathbf{c} = (c_1, \cdots, c_I)$ as a decision. Note that, if $c_i = 0$, we can simply remove resource i from the network. Thus, when analyzing the performance of a system for a given $\mathbf{c}$, we always assume that $\min_{i \in \mathscr{I}}\{c_i\} \triangleq c_{\min} > 0$.

3. Dynamic fulfillment policy. Upon an arrival, the resource needed to fulfill the request must be determined based on the flexibility structure $\mathscr{A}$ and system status. Since it is difficult to find an optimal dynamic fulfillment policy, we will extend the greedy fulfillment policy for the long chain proposed by Asadpour et al. (2019) to unbalanced systems.

We will refer to $(\mathscr{A}, \mathbf{c}, \mathbf{p})$ as a system and the goal is to better match supply with demand, that is, satisfy customer demand as much as possible. We take the fully flexible version of a system as the benchmark

Figure 1. (Color online) Network Structures



A general network structure

A long chain

and consider the difference in the expected number of lost sales between the two systems after all requests have arrived as the performance measure. The performance difference is most significant when the total demand is exactly equal to the total capacity K, as discussed in Asadpour et al. (2019). Thus we will follow Asadpour et al. (2019) and focus on the impact of limited flexibility by assuming that the total expected demand is K. In this case, there would not be any lost sales under full flexibility no matter how inventory is allocated and how requests are fulfilled, and the performance measure reduces to the expected number of lost sales of a given system $(\mathscr{A}, \mathbf{c}, \mathbf{p})$.

4. The GCG Systems and a Dynamic Fulfillment Policy

In this section, we will first define the GCG of a given system $(\mathscr{A}, \mathbf{c}, \mathbf{p})$ and show that a positive GCG is a necessary condition for the expected number of lost sales to remain bounded despite increases in K. We then introduce the dynamic fulfillment policy.

4.1. Generalized Chaining Gap

Let $\mathcal{J}(i) = \{j : (i,j) \in E\}$ and $\mathcal{J}(\mathcal{I}') = \cup_{i\in\mathcal{I}'}\mathcal{J}(i)$ be the sets of request types that can be fulfilled by resource i and by a resource in $\mathcal{I}' \subseteq \mathcal{I}$, respectively. Similarly, let $\mathcal{I}(j) = \{i : (i,j) \in E\}$ and $\mathcal{I}(\mathcal{J}') = \cup_{j\in\mathcal{J}'}\mathcal{I}(j)$ be the sets of resources that can fulfill type j requests and a request in $\mathcal{J}' \subseteq \mathcal{J}$, respectively.

For a subset of requests $\mathcal{J}'$, $\eta^{\mathcal{J}'} = \sum_{i\in\mathcal{I}(\mathcal{J}')} c_i - \sum_{j\in\mathcal{J}'} p_j$ represents the system's ability to fulfill $\mathcal{J}'$. In Figure 1(a), request types in $\mathcal{J}' = \{1,2\}$ can only be fulfilled by resources in $\mathcal{I}(\mathcal{J}') = \{1,2\}$ and $\eta^{\mathcal{J}'} = (c_1 + c_2) - (p_1 + p_2)$. The GCG is then defined as

$$\eta \triangleq \min_{\mathcal{J}' \subsetneq \mathcal{J},\, \mathcal{J}' \neq \emptyset} \{\eta^{\mathcal{J}'}\} \tag{1}$$

measuring the ability of the system to fulfill all subsets of requests. Our GCG is a special case of that considered by Shi et al. (2019) when the total demand is equal to the total capacity K. A system $(\mathscr{A}, \mathbf{c}, \mathbf{p})$ with a positive GCG is referred to as a GCG system in which there is slack between the total expected demand from request types in $\mathcal{J}' \subset \mathcal{J}$ and the total inventory that can be used to fulfill the requests in $\mathcal{J}'$. In Figure 1(a), a positive GCG implies that $c_1 + c_2 > p_1 + p_2$.

It can be easily shown that a GCG system $(\mathscr{A}, \mathbf{c}, \mathbf{p})$ described previously has the following important properties:

- The GCG $\eta \le p_{\min}$ since $\eta \le \sum_{i\in\mathcal{I}(\mathcal{J}')} c_i - \sum_{j\in\mathcal{J}'} p_j \le 1 - (1 - p_j) = p_j$ when $\mathcal{J}' = \mathcal{J} \setminus j$ for any $j \in \mathcal{J}$.
- The network $\mathscr{A}$ is connected and has at least $I + J - 1$ arcs. The long chain in Asadpour et al. (2019) where $I = J$ and $\mathbf{c} = \mathbf{p}$ is a GCG system with $I + J$ arcs.
- For any given connected network structure $\mathscr{A}$ and demand vector $\mathbf{p}$, there always exists an inventory allocation $\mathbf{c}$ such that $(\mathscr{A}, \mathbf{c}, \mathbf{p})$ is a GCG network. For instance, we have a GCG network if we allocate p_j amount of inventory evenly to all the resources in $\mathcal{I}(j)$ for all $j \in \mathcal{J}$.

Lastly, we show that a positive GCG is a necessary condition for the expected number of lost sales to not diverge with the market size K.

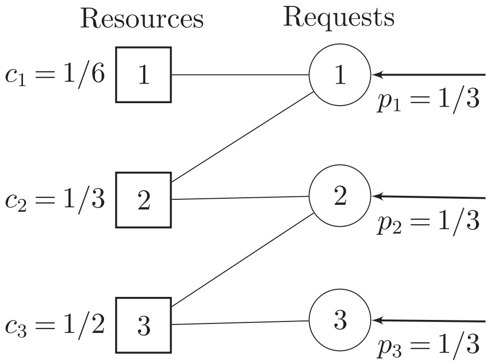
Lemma 1. *If a system $(\mathscr{A}, \mathbf{c}, \mathbf{p})$ has a nonpositive GCG, that is, $\eta \le 0$, then the expected number of lost sales diverges with K under any feasible fulfillment policy.*

4.2. The Load Deviation Fulfillment Policy for a GCG System $(\mathscr{A}, \mathbf{c}, \mathbf{p})$

The fulfillment decision when a request arrives clearly requires consideration of not only the network structure $\mathscr{A}$, the capacity vector $\mathbf{c}$, and the demand vector $\mathbf{p}$, but also the system status upon arrival, for example, the current inventory at all resources and the number of remaining arrivals. Thus, it is difficult to optimize the fulfillment decision. Let us first examine two simple fulfillment policies.

1. A priority policy: Each type of request has a primary resource and is fulfilled by another resource only if the primary resource is out of stock. Such a policy may lead to lost sales that increase in K, as demonstrated in the following example. Consider the system in Figure 2, where resource j is the primary resource for request type j and sample paths of the demand where there is an roughly equal number of requests (by roughly we mean that the difference is no more than $O(\sqrt{K})$ from each type after the first $K/2$ arrivals). Then, resource 1 will have little or no inventory, and resources 2 and 3 will have roughly $K/6$ and $K/3$ inventory, respectively, after the first $K/2$ arrivals. Since resource 2 is the only resource for request type 1 as well as the primary resource for request type 2 for the remaining $K/2$ arrivals, the expected number of lost sales will be in the order of K. Since the probability that demand takes such sample

Figure 2. A Simple Example with $I = J = 3$



paths does not vanish as $K \to \infty$ by the central limit theorem, the expected total number of lost sales is at least in the order of K.

2. A random fulfillment policy: Requests of type j are randomly fulfilled by resources in $\mathcal{I}(j)$ with positive remaining inventory according to a certain distribution, for example, with equal probability. That is, a type j request is first randomly assigned to a resource in $\mathcal{I}(j)$. It will be fulfilled by this resource if it has positive inventory or by another resource in $\mathcal{I}(j)$ otherwise. For the example in Figure 2, if a type 2 request is assigned to resource 3 (which occurs with probability $1/2$), it will be fulfilled by resource 2 (if possible) if resource 3 is out of stock. Let $Y_k = 1$ if the kth request is assigned to resource 3 and $Y_k = 0$ otherwise. Then, Y_k is a Bernoulli random variable with mean $1/2$. When resource 3 runs out of stock, which happens after $k' = \min\{k : \sum_{i=1}^{k} Y_i \geq K/2\}$ arrivals, the number of remaining requests of type 3 follows a binomial distribution with $([K - k']^+, 1/3)$ and will be lost. Since

$$\begin{aligned}
\mathbb{E}[K - k']^+ &= \int_0^K \mathbb{P}(K - k' \geq k)dk = \int_0^K \mathbb{P}(k' \leq k)dk \\
&= K - \int_0^K \mathbb{P}(k' \geq k)dk \\
&= K - \sum_{k=1}^{K} \mathbb{P}(Y_1 + \ldots + Y_k \leq K/2) \\
&\geq \sqrt{K} - \sum_{k=K-\sqrt{K}}^{K} \mathbb{P}(Y_1 + \ldots + Y_k \leq K/2) \\
&\geq \sqrt{K} - \sqrt{K}\mathbb{P}(Y_1 + \ldots + Y_{K-\sqrt{K}} \leq K/2) \\
&\sim \sqrt{K}(1 - \mathbb{P}(Z \leq 1)),
\end{aligned}$$

where $\sim$ follows from the central limit theorem and Z is the standard normal random variable. $\mathbb{E}[K - k']^+$ and hence the total expected number of lost sales are at least in the order of $\sqrt{K}$.

Thus, we will generalize the modified greedy policy in Asadpour et al. (2019), which was designed for the long chains where $I = J$ and each request can be fulfilled by exactly two resources. Let $L_i(k)$ be the number of requests that have been assigned (which we will explain later) to resource i, referred to as the load of resource i, and $X_i(k) = L_i(k) - c_i k$ be its deviation from the average load of resource i after k arrivals. Let $X_i(0) = 0$ for all i. A positive (negative) load deviation indicates a higher (lower) ideal rate of demand for inventory at a resource. As the $(k+1)$th request, for example, of type j, arrives, it is assigned to a resource in $\mathcal{I}(j)$ that has the smallest load deviation regardless of its inventory status, denoted by $i^*(j)$, and the load at this resource is updated as $L_{i^*(j)}(k+1) = L_{i^*(j)}(k) + 1$ while $L_i(k+1) = L_i(k)$ for $i \neq i^*(j)$. If there are multiple resources with the same smallest load deviation, we simply pick one randomly. Thus, the load deviation evolves as

$$\begin{aligned}
X_i(k+1) &= L_i(k+1) - c_i(k+1) \\
&= X_i(k) - c_i + \begin{cases} 1, & \text{if } i = i^*(j), \\ 0, & \text{otherwise,} \end{cases}
\end{aligned} \tag{2}$$

and, for any $k \leq K$, $\sum_{i=1}^{I} L_i(k) = k$ and $\sum_{i=1}^{I} X_i(k) = 0$. Note that resource $i^*(j)$ may not have inventory, in which case, the request will be fulfilled by a resource in $\mathcal{I}(j)$ in an arbitrary manner, or lost if none of the resources in $\mathcal{I}(j)$ has inventory.

We can remove resources from the system one by one over time as they run out of inventory and the network structure changes in k. However, doing so would greatly complicate the presentation of the analysis. Thus, for the ease of presentation, we will keep all resources in the network at all times and allow the assignment of requests to resources with zero inventory, that is, $L_i(k) \geq c_i K$, even though these requests would most likely be fulfilled by another resource with inventory. Thus, $L_i(k)$ can be understood as the number of requests that would have been fulfilled by resource i after k arrivals had there been enough inventory.

We would like to point out that a fulfillment policy based on the relative magnitude of the load deviation $X_i(k)/c_i, i = 1, \ldots, I$, also works well and bounded performance is guaranteed by the same bound in the next section for GCG systems. As a matter of fact, numerical examples indicate that this weighted load deviation policy may perform even better than the load deviation policy.

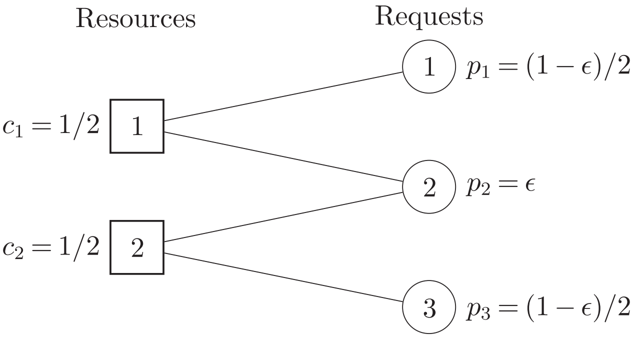
5. Bounded Performance of GCG Systems

In this section, we establish an upper bound on the expected number of lost sales for any GCG system $(\mathscr{A}, \mathbf{c}, \mathbf{p})$ if requests are fulfilled according to the load deviation fulfillment policy. This bound is independent of the market size K, implying that a positive GCG is not only necessary but also sufficient to guarantee bounded performance. By bounded performance we mean that the expected number of lost sales does not diverge with the market size K.

Theorem 1. *The expected number of lost sales of a GCG system are bounded from above by* $\ln 64 \cdot \max\{\frac{1}{c_{\min}}, \frac{I}{\eta}\}$, *independent of the market size K.*

The upper bound only depends on I, $c_{\min}$, and the GCG η. When $c_{\min} < \eta/I$, which occurs when at least one resource is allocated relatively low inventory compared with others, the upper bound is inversely proportional to $c_{\min}$ but independent of other system parameters. Otherwise, the upper bound is increasing in the number of resources I and decreasing in the GCG η. For the long chain with $\mathbf{c} = \mathbf{p}$ in

Figure 3. Illustration of the Tightness of the Upper Bound



Asadpour et al. (2019), $\eta = p_{\min} = c_{\min}$ and our bound reduces to $\ln 64 \cdot I/\eta$, which is tighter than $2I/\eta \ln(1 + 18I^2/\eta^2)$ provided by Asadpour et al. (2019).

Tightness of the upper bound. Note that the upper bound is in the order of η^{-1} as $\eta \to 0$. Consider the example in Figure 3, where $\mathbf{c} = (1/2, 1/2)$ and $\mathbf{p} = ((1-\epsilon)/2, \epsilon, (1-\epsilon)/2)$ for small $\epsilon > 0$, so $\eta = \epsilon/2$. Since the total number of type 1 requests, denoted as $\mathcal{D}_1$, follows a binomial distribution with (K, p_1), the expected total number of lost sales of request type 1 alone is at least

$$\begin{aligned}\mathbb{E}[\mathcal{D}_1 - Kc_1]^+ &= \mathbb{E}[\mathcal{D}_1 - K(p_1 + \epsilon/2)]^+ \\ &= \sqrt{Kp_1(1-p_1)}\mathbb{E}\left[\frac{\mathcal{D}_1 - Kp_1}{\sqrt{Kp_1(1-p_1)}} - \frac{\sqrt{K}\epsilon}{2\sqrt{p_1(1-p_1)}}\right]^+ \end{aligned} \tag{3}$$

under any feasible fulfillment policy. Note that $\frac{\mathcal{D}_1 - Kp_1}{\sqrt{Kp_1(1-p_1)}}$ converges to the standard normal as $K \to \infty$, and $p_1 \to 1/2$ as $\epsilon \to 0$. When $\epsilon \to 0$, for $K = C/\epsilon^2$ where C is a constant so that $\sqrt{K}\epsilon$ is a constant, the right-hand side of Equation (3) is at least in the order of ϵ^{-1} or η^{-1} and our upper bound is indeed tight.

5.1. Overview of the Proof

At a high level, we follow Asadpour et al. (2019) by first establishing a bound of the expected number of lost sales in Lemma 2 and then trying to bound the expectation of the potential function, which is achieved by showing that the potential function exhibits a contraction property.

Lemma 2. *The expected total number of lost sales under the Load Deviation Fulfillment Policy (LDFP) is bounded from above by*

$$\mathbb{E}\left[\sum_{i=1}^{I} \max\{X_i(K), 0\}\right] \le IC \ln\left(\frac{1}{I}\mathbb{E}[\Phi(\mathbf{X}(K))] + 1\right),$$

where $\Phi(\mathbf{X}(k)) = \sum_{i=1}^{I} e^{X_i(k)/C}$ *is a potential function.*

We say that the potential function exhibits a contraction property if

$$\begin{aligned}&\mathbb{E}[\Phi(\mathbf{X}(k+1)) \mid \mathbf{X}(k)] \le (1-a)\Phi(\mathbf{X}(k)) + b, \\ &\quad \text{for some } 0 < a < 1, b > 0,\end{aligned} \tag{4}$$

which immediately implies that $\mathbb{E}[\Phi(\mathbf{X}(K))] \le b/a$ by induction under the initial condition $\mathbb{E}[\Phi(\mathbf{X}(0))] \le b/a$.

The approach used by Asadpour et al. (2019) to establish Equation (4) for the long chains relies heavily on their symmetric network structure and does not apply to general network structures. Thus, we need a completely new approach to establishing Equation (4) for general network structures. Since Equation (4) is equivalent to

$$\mathbb{E}[\Phi(\mathbf{X}(k+1)) \mid \mathbf{X}(k)] - \Phi(\mathbf{X}(k)) \le -a\Phi(\mathbf{X}(k)) + b, \tag{5}$$

we first obtain the following property of the potential function through the Taylor expansion of the left-hand side of Equation (5).

Lemma 3. *For any* $C > 1$,

$$\begin{aligned}&\mathbb{E}[\Phi(\mathbf{X}(k+1)) \mid \mathbf{X}(k)] - \Phi(\mathbf{X}(k)) \\ &\quad \le \frac{2}{C^2}\sum_{i=1}^{I} c_i e^{X_i(k)/C} - \left(\frac{1}{C} + \frac{1}{C^2}\right)\Gamma,\end{aligned} \tag{6}$$

where

$$\Gamma = \left(\sum_{i=1}^{I} c_i e^{X_i(k)/C} - \sum_{j=1}^{J} p_j e^{X_{i^*(j)}(k)/C}\right) \tag{7}$$

and $i^*(j)$ *is the resource assigned to the* $(k+1)$*th arrival if it is of type* j.

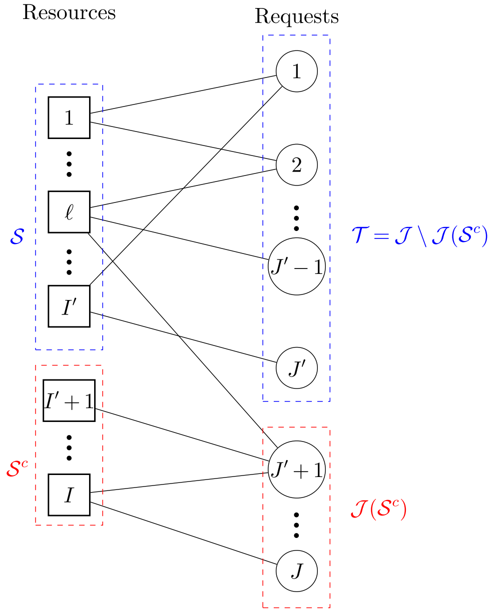
With Lemma 3, establishing Equation (5) becomes finding an upper bound of the right-hand side of Equation (6), that is, a desired lower bound of Γ and an upper bound of $\sum_{i=1}^{I} c_i e^{X_i(k)/C}$ for any $C > 1$. When C is relatively large, the second term of the right-hand side of Equation (6) dominates the first term and we focus on finding a desired lower bound of Γ, that is,

$$\Gamma \ge a'\Phi(\mathbf{X}(k)) + b' \tag{8}$$

for some $a' > 0$.

Note that Γ is a summation of terms associated with all the resources and demand locations. Recall that $i^*(j)$ is the resource with the lowest load deviation in $\mathcal{I}(j)$; thus, if we define $\mathcal{S}^c = \{i \in \mathcal{I} : X_i(k) = X_{\min}(k)\}$ where $X_{\min}(k)$ denotes the value of the lowest load deviation among all resources, then $X_i(k) = X_{i^*(j)}(k) = X_{\min}(k)$ for all $i \in \mathcal{S}^c$ and $j \in \mathcal{J}(\mathcal{S}^c)$. For the terms in Γ associated

Figure 4. (Color online) Illustration of the Sets $\mathscr{S}, \mathscr{S}^c, \mathscr{J}(\mathscr{S}^c)$ and $\mathscr{T}$


Resources
Requests
1
⋮
ℓ
⋮
I'
$I'+1$
⋮
I
1
2
⋮
$J'-1$
J'
$J'+1$
⋮
J
$\mathcal{S}$
$\mathcal{S}^c$
$\mathcal{T} = \mathcal{J} \setminus \mathcal{J}(\mathcal{S}^c)$
$\mathcal{J}(\mathcal{S}^c)$

with the set of resources $\mathscr{S}^c$ and the set of demand locations $\mathscr{J}(\mathscr{S}^c)$, the following holds:

$$\sum_{i\in\mathscr{S}^c} c_i e^{X_i(k)/C} - \sum_{j\in\mathscr{J}(\mathscr{S}^c)} p_j e^{X_{i^*(j)}(k)/C} = e^{X_{\min}(k)/C}\left(\sum_{i\in\mathscr{S}^c} c_i - \sum_{j\in\mathscr{J}(\mathscr{S}^c)} p_j\right).$$

Thus, we only need to bound the terms associated with the rest of the resources $\mathscr{S} = \mathscr{I} \setminus \mathscr{S}^c$ and demand locations $\mathscr{T} = \mathscr{J} \setminus \mathscr{J}(\mathscr{S}^c)$; that is,

$$\Gamma_{(\mathscr{S},\mathscr{T})} = \sum_{i\in\mathscr{S}} c_i e^{X_i(k)/C} - \sum_{j\in\mathscr{T}} p_j e^{X_{i^*(j)}(k)/C}.$$

For an illustration of the network partition, see Figure 4.

To find a lower bound of $\Gamma_{(\mathscr{S},\mathscr{T})}$, note that the GCG of the subnetwork involving $\mathscr{S}$ and $\mathscr{T}$ is at least η by definition. That is, there is at least η amount of slack inventory to fulfill requests in $\mathscr{T}$. We can then apply the max-flow min-cut theorem to establish a desired lower bound in the form of Equation (8). In the same process, we can also obtain an upper bound of $\sum_{i=1}^{I} c_i e^{X_i(k)/C}$. With these bounds, we can derive the contraction property of the potential function and obtain the desired performance bound by an appropriate choice of C. A detailed proof of Theorem 1 is in the e-companion.

It is worth mentioning that even though we consider general network structures, our proof involves fewer steps in the bounding process than that in Asadpour et al. (2019). Although they partitioned the resources with positive load deviations into multiple subgroups and then established a bound for the terms associated with each of the groups, we partition the whole network (including the resources and requests) into two subnetworks and only need to establish a bound for the terms in one subnetwork. In addition to fewer steps in the bounding process (each step of bounding loosens the bounds), they relied on local minimum and maximum load deviations in each subgroup, while we applied the max-flow min-cut theorem to the subnetwork involving $\mathscr{S}$ and $\mathscr{T}$. We attribute these differences to our tighter bounds.

6. Inventory Allocation and Network Design

In this section, we consider the decisions on inventory allocation $\mathbf{c}$ for a given connected network structure $\mathscr{A}$ and demand vector $\mathbf{p}$ in Section 6.1 and network structure in Section 6.2 with the goal of minimizing the expected number of lost sales. Since the objective function is elusive due to the complexity of the problem, we need to find a proxy for it first and will seek inspiration from the upper bound in Section 5, $\ln 64 \cdot \max\{\frac{1}{c_{\min}}, \frac{I}{\eta}\}$.

Since η is not a monotone function of $c_{\min}$, the bound may increase or decrease in $c_{\min}$. For a given network structure, suppose that $c_{\min}$ is low, for example, $c_{\min} < \frac{\eta}{2I}$. We can move some inventory from resources with higher inventory to raise $c_{\min}$ to $c'_{\min} = \frac{\eta}{2I}$ with the corresponding GCG, η'. It is easy to show that $\eta' \ge \eta$ or $\frac{\eta}{2} \le \eta' < \eta$. Thus, the bound under the new inventory allocation, $\ln 64 \cdot \max\{\frac{1}{c'_{\min}}, \frac{I}{\eta'}\}$, reduces to $\ln 64 \cdot \frac{I}{\eta'} \le \ln 64 \cdot \frac{2I}{\eta} < \ln 64 \cdot \max\{\frac{1}{c_{\min}}, \frac{I}{\eta}\}$, the bound under inventory allocation $\mathbf{c}$. On the other hand, the bound is obviously in the order of $\frac{I}{\eta}$ for inventory allocations with $c_{\min} \ge \frac{\eta}{2I}$. Given the tightness of the performance bound established after introducing Theorem 1, a higher η is likely to indicate better performance, an insight consistent with that from Shi et al. (2019) for production systems and confirmed by our numerical study in Section 8.1. Thus, we will use η as a proxy for the objective function when making the inventory allocation and network structure decisions.

6.1. Inventory Allocation Decision

In this section, we examine the inventory allocation decision $\mathbf{c}$ that maximizes the GCG for any connected network $\mathscr{A} = (\mathscr{I}, \mathscr{J}, E)$ and demand vector $\mathbf{p}$. We first present a lower bound of the highest GCG possible, denoted as η^*.

Lemma 4. *For any given connected network $\mathscr{A}$ and demand vector $\mathbf{p}$, $\eta^* \ge \min_{j\in\mathscr{J}} \frac{p_j}{|\mathscr{I}(j)|}$.*

Since a connected network requires at least $I + J - 1$ arcs, we will consider the inventory allocation decisions for networks with at least $I+J-1$ arcs. Before proceeding, we would like to exclude a trivial case where a resource is dedicated to a single request type in a GCG system. If resource i is dedicated to request type j, that is, $|\mathcal{J}(i)|=1$, request type j must have access to at least one more resource as the network would be disconnected otherwise. Since inventory at resource i has little flexibility, one should not allocate any to it absent of capacity or geographical constraints and the problem reduces to one with $I-1$ resources. Otherwise, we simply allocate the minimum required inventory to resource i and the problem reduces to allocating the rest of the inventory $1-c_i$ to $I-1$ resources for the remaining demand in the system. Thus, we only need to consider GCG networks where $|\mathcal{J}(i)|\geq 2$ for all $i\in\mathcal{I}$; that is, each resource must serve at least two types of requests. This is certainly true in most real applications where $I \ll J$.

6.1.1. Networks with $I+J-1$ and $I+J$ Arcs. Let $d(j)$ denote the number of connected subnetworks after request type j and the arcs associated with it are removed. Then, $d(j)=|\mathcal{I}(j)|$ for networks with $I+J-1$ arcs, and $d(j)=|\mathcal{I}(j)|-1$ or $|\mathcal{I}(j)|$ for networks with $I+J$ arcs. This is because, with only $I+J-1$ arcs, a network does not contain a cycle and removing request type j divides the network into exactly $|\mathcal{I}(j)|$ connected, nonoverlapping subnetworks. Adding one more arc increases the connectivity of a network and hence may reduce $d(j)$, by at most 1.

Proposition 1. *For a connected network structure $\mathcal{A}=(\mathcal{I},\mathcal{J},E)$ and demand vector $\mathbf{p}$, $\eta^*=\min_{j\in\mathcal{J}}\frac{p_j}{d(j)}$ is achieved under the following inventory allocations.*

- *If $|E|=I+J-1$, allocate p_j amount of inventory evenly to all the resources in $\mathcal{I}(j)$.*
- *If $|E|=I+J$, allocate $\frac{p_j}{d(j)}$ amount of inventory to each of the $d(j)$ subnetworks and then allocate $\frac{p_j}{d(j)}$ evenly to the resources in each subnetwork that belong to $\mathcal{I}(j)$.*

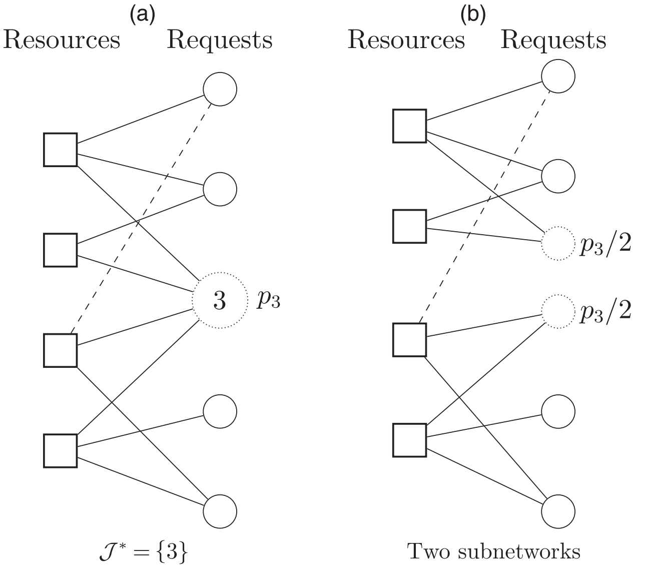
For networks with $I+J-1$ arcs, Proposition 1 reveals a very simple inventory allocation that maximizes the GCG and η^* can be any value in $[\frac{p_{\min}}{I}, p_{\min}]$. For example, $\eta^*=\frac{p_{\min}}{I}$ if $J-1$ types of requests each have a single supplier and one request type with the lowest demand enjoys full supplier flexibility. On the other hand, $\eta^*=p_{\min}$ if $p_{\max}\geq Ip_{\min}$, and a request type with $p_{\max}$ is linked to all the resources (with I arcs) and the rest of the request types are only linked to a single resource (with $J-1$ arcs).

Proposition 1 also implies that adding one arc to an existing connected network with $I+J-1$ arcs can potentially decrease $d(j)$ and achieve a higher GCG. For example, a long chain with $p_j=\frac{1}{J}$ for all $j\in\mathcal{J}$ can achieve the highest GCG $\eta^*=\frac{1}{J}=p_{\min}$. Removing any arc does not affect the connectivity of the network, but it reduces the GCG by a half to $\eta^*=\min_{j\in\mathcal{J}}\frac{1/J}{|\mathcal{I}(j)|}\leq\frac{1}{2J}$. This further confirms the effectiveness of the long chains.

6.1.2. Beyond $I+J$ Arcs. With more arcs, a network may still be connected after removing a request type. Thus, we need to extend $d(j)$ and let $d(\mathcal{J}')$ be the number of connected subnetworks after removing $\mathcal{J}'\subset\mathcal{J}$ and all the arcs associated with it. The higher the $d(\mathcal{J}')$, the weaker $\mathcal{J}'$ is connected to the rest of the network. Let $\mathcal{J}^*=\arg\min_{\mathcal{J}'\subset\mathcal{J}}\frac{\sum_{j\in\mathcal{J}'}p_j}{d(\mathcal{J}')}$ (choose one arbitrarily if multiple minimizers exist). In all the subnetworks after removing $\mathcal{J}^*$, we add a single request node with demand $\frac{\sum_{j\in\mathcal{J}^*}p_j}{d(\mathcal{J}^*)}$ so that the total demand from $\mathcal{J}^*$ is evenly allocated to the $d(\mathcal{J}^*)$ subnetworks. Each link from $\mathcal{J}^*$ to a resource in $\mathcal{I}(\mathcal{J}^*)$ in the original network is represented by a link from the new request node in the same subnetwork to it. The process can be viewed as if we group the request types in $\mathcal{J}^*$ into a single request type with demand $\sum_{j\in\mathcal{J}^*}p_j$. We then split it into $d(\mathcal{J}^*)$ requests, each with demand $\frac{\sum_{j\in\mathcal{J}^*}p_j}{d(\mathcal{J}^*)}$ and in a unique subnetwork, while maintaining all the links from $\mathcal{I}(\mathcal{J}^*)$ to $\mathcal{J}^*$ in their respective subnetworks. As an illustration, consider the example in Figure 5(a), where $\mathcal{J}^*=\{3\}$ and without the dashed arc. Then, removing request type 3 and arcs associated with it will break the network into $d(\mathcal{J}^*)=2$. We split request type 3 into two, each with demand $p_3/2$ in its associated subnetwork as in Figure 5(b).

Of course, with a well-connected network, some of the $d(\mathcal{J}^*)$ subnetworks may still be well connected or even have full flexibility. In that case, we suggest following the same procedure in those subnetworks

Figure 5. A Network Structure with $I+J+1$ (Solid) Arcs



until all their subnetworks are of the structures in Section 6.1.1 when making the inventory allocation decision, although we are not able to prove that it will maximize the GCG.

Proposition 2. *If all the $d(\mathcal{J}^*)$ subnetworks are of the structures in Section 6.1.1, $\eta^* = \frac{\sum_{j\in\mathcal{J}^*} p_j}{d(\mathcal{J}^*)}$ if inventory is allocated according to their respected rules for the subnetworks.*

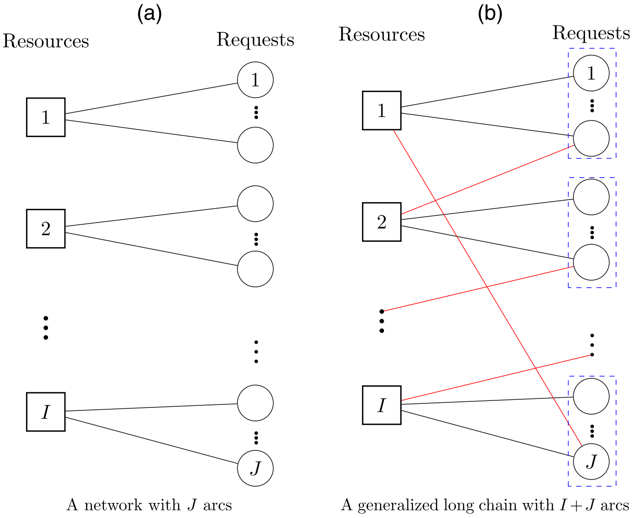
6.2. Network Structure Design

In Section 6.1, we provided the maximum achievable GCG, η^*, for given $\mathbf{p}$ and $\mathscr{A} = (\mathcal{I}, \mathcal{J}, E)$, and the inventory allocation that achieves η^*. We now determine the set E that maximizes η^* for networks with $I+J-1$ arcs in Section 6.2.1, $I+J$ arcs in Section 6.2.2 and more than $I+J$ arcs in Section 6.2.3.

Since we do not explicitly consider the storage capacity at the resources or geographical constraints, a network with a single resource and a total of J arcs would be sufficient to achieve the best performance, which is obviously not practical. In real networks, each distribution center is responsible for multiple demand locations due to their proximity to it, and each demand location is covered by at least one distribution center. Thus, to avoid the complications of explicit capacity and geographical constraints, we start with networks with J arcs that link each request type to a resource and each resource is linked to at least one request, as illustrated in Figure 6(a). That is, design of a network structure starts with an existing network with I groups of request types. The decision is to determine the rest of the arcs that maximize η^* under the inventory allocations described in Section 6.1.

6.2.1. Networks with $I+J-1$ Arcs. To construct networks with $I+J-1$ arcs that maximize η^*, we first find the optimal number of suppliers for each request type $|\mathcal{I}^*(j)|$, $j \in \mathcal{J}$. Insights from Section 6.1.1 suggest that request types with higher demand should be given higher supplier flexibility, that is, access to more resources. That is, suppose that $p_1 \le \cdots \le p_J$ without loss of generality. Then, $|\mathcal{I}^*(j)| \ge 1$ should be nondecreasing in j and can be obtained using Algorithm 1. By Proposition 3, any connected network with $|\mathcal{I}(j)| = |\mathcal{I}^*(j)|$, $j \in \mathcal{J}$, maximizes η^*. Note that for given $|\mathcal{I}(j)| = |\mathcal{I}^*(j)|$, $j \in \mathcal{J}$, and the existing J arcs, there may be multiple ways to form a network with $I+J-1$ arcs, which provides flexibility and opportunities to take capacity constraints into consideration in designing a network in practice.

Figure 6. (Color online) Illustration of an Existing Network and a Generalized Long Chain



Proposition 3. *For given I, J, and demand vector $\mathbf{p}$, any connected network with $|\mathcal{I}^*(j)|$, $j \in \mathcal{J}$, obtained by Algorithm 1 achieves the same and highest possible GCG with $I+J-1$ arcs, if p_j amount of inventory is evenly allocated to the resources in $\mathcal{I}^*(j)$.*

Algorithm 1. Network Design with $I+J-1$ Arcs

```
Initialization: Find a solution $1 = |\mathcal{I}^*(1)| \le |\mathcal{I}^*(2)| \le \cdots \le |\mathcal{I}^*(J)|$ such that $\sum_{j=1}^{J} |\mathcal{I}^*(j)| = I+J-1$.
while 1 do
    Compute $\frac{p_j}{|\mathcal{I}^*(j)|}$ and $\frac{p_j}{|\mathcal{I}^*(j)|+1}$ for $j \in \{1,\ldots,J\}$. Let $\underline{j}$ be the smallest index such that $\frac{p_j}{|\mathcal{I}^*(j)|}$ is minimal and $\bar{j}$ the largest index such that $\frac{p_j}{|\mathcal{I}^*(j)|+1}$ is maximal. Let $\eta^* = \frac{p_{\underline{j}}}{|\mathcal{I}^*(\underline{j})|}$.
    if $|\mathcal{I}^*(\underline{j})| = 1$ then break;
    end if
    if $\eta^* < \frac{p_{\bar{j}}}{|\mathcal{I}^*(\bar{j})|+1}$ then,
        Update $|\mathcal{I}^*(\underline{j})| \leftarrow |\mathcal{I}^*(\underline{j})| - 1$,
            $|\mathcal{I}^*(\bar{j})| \leftarrow |\mathcal{I}^*(\bar{j})| + 1$;
    else break;
    end if
end while
Output: $\{|\mathcal{I}^*(1)|, \cdots, |\mathcal{I}^*(J)|\}$ and $\eta^*$.
```

6.2.2. Networks with $I+J$ Arcs. Note that, with only $I+J-1$ arcs, the highest GCG that can be achieved is likely to be below $p_{\min}$. If we have the freedom to design a network structure with $I+J$ arcs, we can form a GLC by linking the I resources and I groups of request types with I arcs through a request node in each request group (with the highest demand if possible), as illustrated in Figure 6(b). Since $d(j) = 1$ for all $j \in \mathcal{J}$, $\eta^* = p_{\min}$ by Proposition 1.

We note that Shi et al. (2019) provide an algorithm for designing a network with $I+J$ arcs for given $(\mathbf{p}, \mathbf{c})$ that achieves a GCG above a threshold, while we treat the inventory vector $\mathbf{c}$ as a decision and design networks that achieve the highest GCG, that is, $p_{\min}$. Indeed, with the flexibility in inventory allocation, one has more freedom in selecting a network structure that not only achieves the highest GCG, $p_{\min}$, but also takes other constraints (e.g., storage capacity and shipping distance) into consideration.

Proposition 1 also provides two insights if we have the option to add an additional arc to an existing connected network with $I+J-1$ arcs. First, the request type with the lowest $\frac{p_j}{|\mathcal{I}(j)|}$ should be in the cycle as $d(j) < |\mathcal{I}(j)|$ for all request type j in the cycle. Second, an additional arc should always be added to form as large a cycle as possible. This insight is consistent with the observations made by Jordan and Graves (1995, p. 586), who state that as one of the flexibility principals, "the right way to add flexibility is to create fewer, longer plant-product chains."

6.2.3. Networks Beyond $I+J$ Arcs.

Since $I+J$ arcs are sufficient to design a network that achieves the maximum GCG possible, $p_{\min}$, adding one more arc will not improve the GCG. However, if we have the option to add an additional arc to an existing network, Proposition 2 suggests that it should be added to strengthen the weakest link, that is, to reduce $d(\mathcal{J}^*)$. In the example in Figure 5, we can add the dashed link to reduce $d(\{3\})$ from 2 to 1.

7. Extensions

7.1. Random Batch Arrivals

So far we have assumed that each arrival only needs one unit of the product. In this section, we allow a random batch size of ℓ_j for request type j, in which case the total demand may not be equal to the total initial inventory K and the number of lost sales under full flexibility may not be zero. Suppose that the batch sizes are independent and identically distributed across different arrivals. Then, $p'_j = p_j\mathbb{E}(\ell_j)/D$ represents the expected demand rate for request type j, where $D = \sum_{r=1}^{J} p_r\mathbb{E}(\ell_r)$ is the expected batch size for each arrival. We say that the previous system has a positive GCG denoted as η' if the system defined in our main model, $(\mathcal{A}, \mathbf{c}, \mathbf{p}')$, is a GCG system.

With batch arrivals, we need to modify the load deviation policy. Let $\ell(k)$ be the batch size of the kth arrival, $L_i(k)$ be the number of units that have been assigned to resource i, the load of resource i, and $X_i(k) = L_i(k) - c_i\sum_{s=1}^{k}\ell(s)$ be the load deviation of resource i after the kth arrival. As the $(k+1)$th request, say of type j, arrives, it is assigned to a resource in $\mathcal{I}(j)$ that has the smallest load deviation regardless of its inventory status, denoted by $i^*(j)$, and the load at this resource is updated as $L_{i^*(j)}(k+1) = L_{i^*(j)}(k) + \ell(k+1)$ while $L_i(k+1) = L_i(k)$ for $i \neq i^*(j)$. We recognize that an order may be fulfilled by multiple resources in reality. For simplicity, we will assume that each order can only be fulfilled by inventory at one resource and may be partially fulfilled due to insufficient inventory at the resource. Next we establish a similar performance bound as in Theorem 1 under mild conditions.

Theorem 2. *Suppose that the random batch sizes* $\{\ell_j : j \in \mathcal{J}\}$ *have finite support, that is,* $\ell_j \le \bar{\ell}$ *for all* $j \in \mathcal{J}$. *The expected optimality gap between a GCG system and the system with full flexibility is bounded from above by* $\ln 64 \cdot \max\{\bar{\ell}, \frac{\max_{j\in\mathcal{J}}\frac{\mathbb{E}(\ell_j^2)}{\mathbb{E}(\ell_j)}}{\min\{c_{\min}, \eta'/I\}}\}$, *independent of the total initial inventory* K.

Here, $\max_{j\in\mathcal{J}}\frac{\mathbb{E}(\ell_j^2)}{\mathbb{E}(\ell_j)}$ measures the variability of the batch sizes. The bound reduces to that in Theorem 1 when $\ell_j \equiv 1$.

7.2. Time-Varying Demand Rates

In this section, we establish bounded performance to the case where the demand vector is time varying, that is, the demand vector of the kth arrival is $\mathbf{p_k} = \{p_{jk}, j = 1, \ldots, J\}$ with the corresponding GCG of the system $(\mathcal{A}, \mathbf{c}, \mathbf{p_k})$ as η_k and $\theta_k = \min\{\eta_k, Ic_{\min}\}$. We say that the system with time-varying demand vector $\{\mathbf{p_k}, k = 1, \ldots K\}$ is a GCG system if $\eta_k > 0$ for all $1 \le k \le K$. Next, we develop an upper bound of the expected number of lost sales for any GCG system, and then show bounded performance if η_k is not too small as k becomes large.

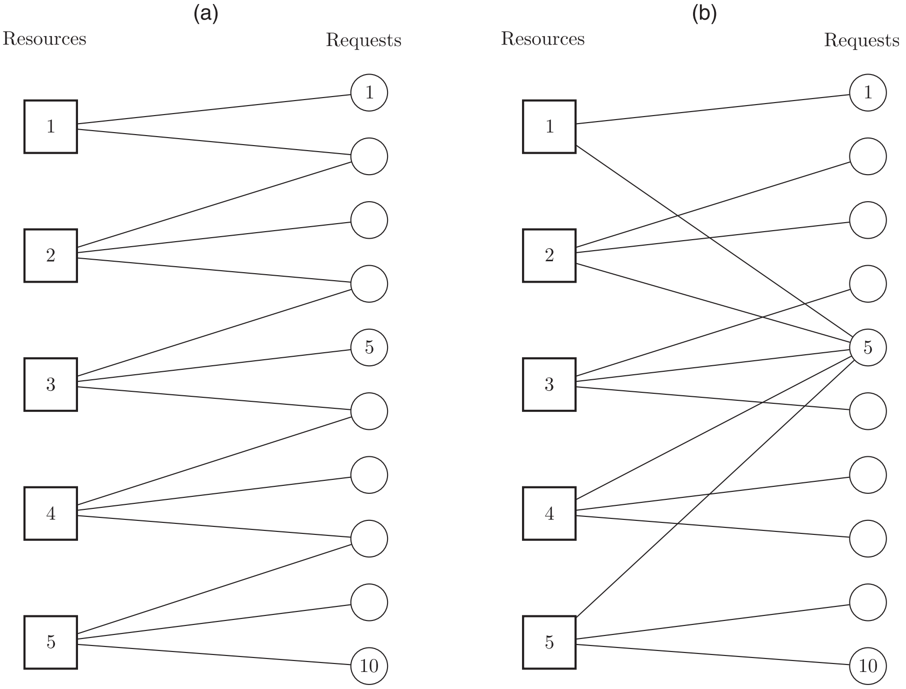
Theorem 3. *Let* $\bar{\theta}_k = \frac{\sum_{r=K-k+1}^{K}\theta_r}{k}$ *for* $1 \le k \le K$. *For any given* K, *the expected number of lost sales of a GCG system is bounded from above by* $\frac{I\ln 64}{\min_{1\le k\le K}\{\bar{\theta}_k\}}$.

Here, $\bar{\theta}_k$ is the average of θ_j's for the last k arrivals. For a special case of $\mathbf{p_k} \to \mathbf{p}$, the upper bound in Theorem 3 converges to $\ln 64 \cdot \max\{\frac{1}{c_{\min}}, \frac{I}{\eta}\}$ when $K \to \infty$, the upper bound in Theorem 1, and bounded performance is guaranteed. Although the bound is a function of K in general, as K becomes large, the impact of earlier arrivals on the performance should diminish. Bounded performance can be achieved as long as η_k is bounded away from zero for k large enough. This condition is similar to the ξ-Hall condition in Asadpour et al. (2019) that guarantees bounded performance when demand is time varying. When applied to long chains under the ξ-Hall condition, the upper bound in Theorem 3 reduces to $\ln 64 \cdot \max\{\frac{1}{c_{\min}}, \frac{I}{\xi}\}$, tighter than the upper bound $2I/\xi\ln(1+18I^2/\xi^2)$ in Asadpour et al. (2018).

8. Numerical Studies

We perform numerical experiments to verify the effectiveness of the GCG as a proxy for system performance under different network structures in Section 8.1, and test some design principles obtained in Section 6.2 for networks in Section 8.2. We show that

Figure 7. Network Structures with $I+J-1$ Arcs



networks with $I+J-1$ and $I+J$ arcs can achieve very good performance.

8.1. The Effectiveness of the GCG as a Proxy for Performance Measure

For a given network structure, different inventory allocation leads to different GCG. Thus, we first consider four different structures, (a) and (b) in Figure 7 with $I+J-1$ arcs, and two GLCs with $I = 5, 10$ and $J = 10$ as in Figure 6(b), where there are $I+J$ arcs and each request group has exactly J/I request types. We let $p_j = 1/J$ and find the inventory allocation that achieves the highest GCG η^* under each structure. We then move different amounts of inventory from the resource with the tightest supply to another resource to create different GCGs as η^*, $\eta^*/2$, $\eta^*/4$, and $\eta^*/8$. As one can see in Figure 8, the expected number of lost sales can increase much faster in K when the GCG is extremely small, while they seem to converge rather quickly or stay flat when the GCG is large. Thus, the GCG is indeed a good indicator of the system performance.

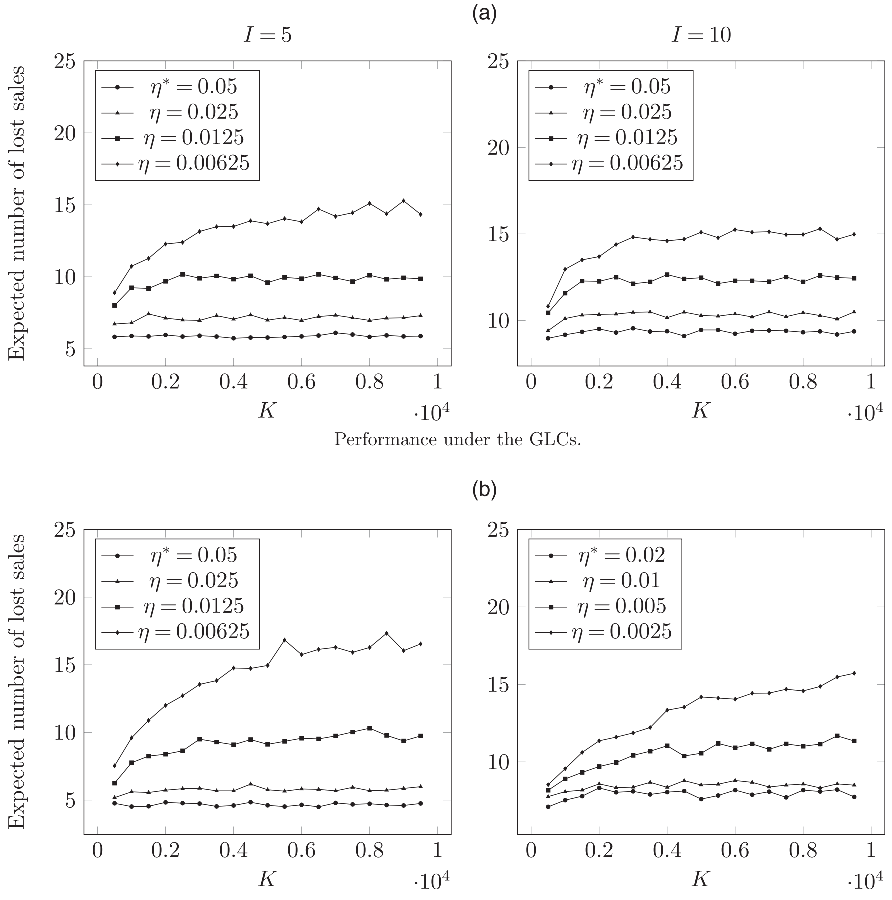
Since network structure also affects the GCG, we now compare the performance of some well-known network structures with $I+J-1$ arcs that maximize the GCG. For $J = I+1$ and $I = 5, 10, 15, 20$, we consider (1) network structures formed according to the outputs from Algorithm 1 that maximize the GCG, and (2) open chains formed by removing arc $(1, I)$ from GLCs. If the demand for all request types are fairly balanced, then the open chains will perform well under the inventory allocation in Proposition 1. So we set $p_j = 1/(2I)$ for $j = 1, \cdots, J-1$ and $p_J = 1/2$. We allocate p_j amount of inventory evenly to all the resources in $\mathcal{I}(j)$ under all network structures and report their performance for different K in Figure 9. As we can see, all the GCG networks with $I+J-1$ arcs perform quite well. Thus, if it is too expensive to add an arc, a GCG network with the minimum number of arcs can be an excellent option. If one is allowed to redesign a network structure with $I+J-1$ arcs rather than removing one arc from a long chain, one may be able to achieve much better performance with the help of Algorithm 1, especially facing asymmetric demand.

8.2. The Impact of Chaining

Note that there is no cycle in any connected network with $I+J-1$ arcs, whereas there is exactly one cycle in networks with $I+J$ arcs. We conjectured in Section 6.2.2 using the GCG as an indication of system performance that, if we are allowed to add one more arc to a connected network with $I+J-1$ arcs, we should do so to form a large cycle rather than a small one. To confirm this, we consider a network with $I+J-1$ arcs obtained by removing arc $(1, J)$ from the corresponding GLC where the J request types are divided evenly into I groups, referred to as an open chain. We then compare its performance with that of networks with one more arc and hence cycles of different sizes, referred to as short chains [with arc $(1, 40/I)$], middle chains [with arc $(1, 10)$], and long chains [with arc $(1, 20)$], as shown by the dotted arcs in Figure 10 where $J = 20$.

We consider $I = 5, 10$. For $I = 5$, there are four requests in each group, and we let $p_j = \frac{1}{25}$ for all j except

Figure 8. Expected Number of Lost Sales as Functions of the GCG



Performance under the GLCs.

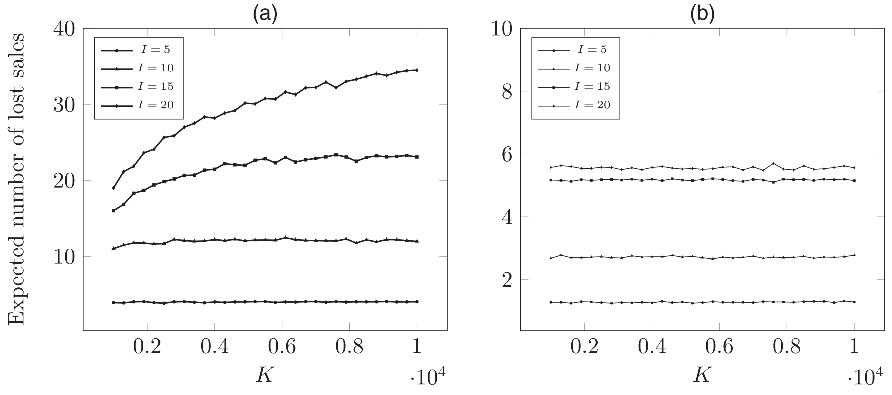
Performance under structures (a) and (b) in Figure 7.

the last request type in each group and $p_{4i} = \frac{2}{25}$, $i = 1, \dots, 5$. For $I = 10$, there are only two request types in each group and we let $p_{2i-1} = \frac{1}{30}$ and $p_{2i} = \frac{1}{15}$ for $i = 1, \dots, 10$. We allocate the inventory so that the GCGs are maximized. Figure 11 provides the expected number of lost sales under these network structures for $K \in [1{,}000, 10{,}000]$ and the performance increases in the size of the cycle in general.

9. A Numerical Study on Amazon China

Both our theoretical bounds and numerical results demonstrate that GCG networks perform very well compared with fully flexible systems with respect to the expected number of lost sales. However, GCG networks may fulfill an order through a farther resource when the closest one has a large load deviation, which may lead to higher shipping costs. In this section, we

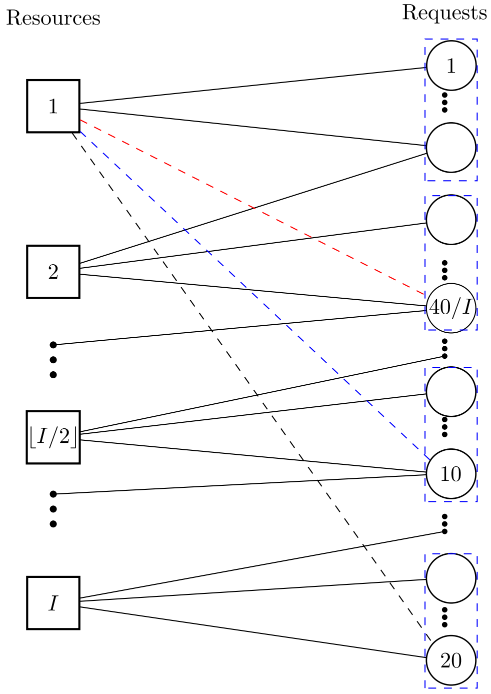
Figure 9. Comparisons Between Open Chains and Structures Generated by Algorithm 1



The open chains

Network structure following Algorithm 1

Figure 10. (Color online) Network Structures Used in Section 8.2



examine both the expected number of lost sales and outbound shipping costs under several flexible GCG networks using real data.

Amazon China has a total of 12 fulfillment centers in 10 regions across China. For simplicity, we combine the centers in the same region into one, which gives $I = 10$ in the network. We select a total of 44 main cities with a population above three million as the demand centers, that is, $J = 44$, and normalize the total retail sales of consumer goods in those cities in 2015 from the National Bureau of Statistics of China[2] to form the nominal demand vector $\mathbf{p}$. The 10 regions, the fulfillment centers in each region, and the cities in each region are shown on the map in Figure 12. Table 1 provides the names of the fulfillment centers, the demand centers in each region, and the demand vector. The number of demand centers in a region ranges from two to nine, and each demand center has access to the fulfillment center in the same region, which requires $J = 44$ arcs.

9.1. Performance of Network Structures with Different Flexibilities

We first construct five different network structures with increasing numbers of arcs, as illustrated in Figure 13(a). Although structures 2–4 are connected networks, structures 0 and 1 are not.

- Structure 0: Each fulfillment center only serves its own region and there is a total of $J = 44$ arcs (arcs with no number). This is similar to the current practice of Amazon China, where each region has its primary fulfillment center.
- Structure 1: Add $I - 2$ arcs (marked as 1) to structure 0 to form two connected networks, the south (regions 1–3) and north (regions 4–10). There is a total of $I + J - 2$ arcs.
- Structure 2: Add 1 arc (marked as 2) linking Beijing (region 3) and Xuzhou (in region 5) to structure 1 to form a connected network with $I + J - 1$ arcs.
- Structure 3: Add 1 arc (marked as 3) linking Xian (region 10) and Jinan (in region 5) with a total of $I + J$ arcs in the network. There is a cycle connecting all the fulfillment centers south of Beijing.
- Structure 4: Add 1 arc (marked as 4) linking Shanghai (region 5) and Qingdao (in region 3) with a total of $I + J + 1$ arcs in the network.

For all the structures, we allocate $p_j K$ amount of inventory evenly to the resources in $\mathcal{I}(j)$ for all $j = 1, \cdots, 44$ and fulfill each arriving request following the load deviation policy. Since structures 0 and 1 are not connected, the systems are not GCG systems, whereas the systems under structures 2–4 are GCG ones. The expected number of lost sales and total shipping distance under the five structures for various K are presented in Figure 14. As we add more arcs, the networks become more flexible and the expected number of lost

Figure 11. Performance for Different Cycle Sizes with $J = 20$

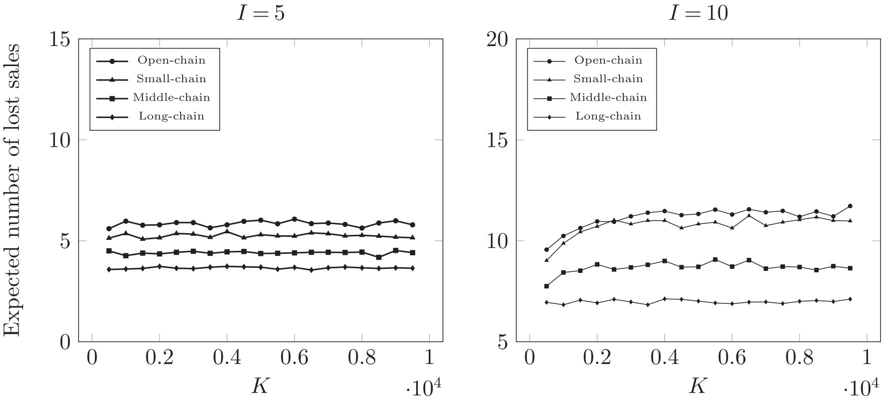
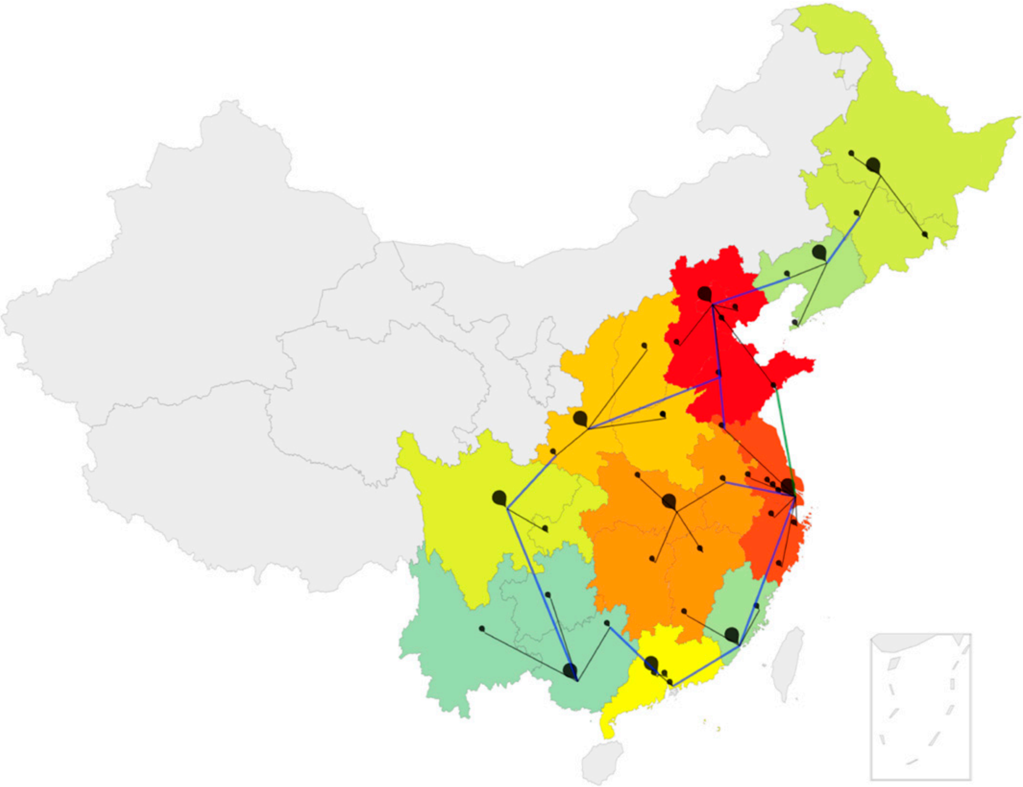


Figure 12. (Color online) Regions, Fulfillment Centers, and Demand Centers of Amazon China



Note. Regions are shown in different colors, fulfillment centers are shown as large black dots, and demand centers are shown as small black dots.

sales decreases rapidly at the (albeit small) expense of shipping distance. The GCG systems work very well and structure 4 represents only a slight improvement over structure 3 with one less arc. Although there is a significant increase in performance when cross-region fulfillments are allowed, the total shipping distance does not increase further as more flexibility is introduced. This suggests that a network with $I + J$ arcs provides enough flexibility to accommodate cross-region shipments.

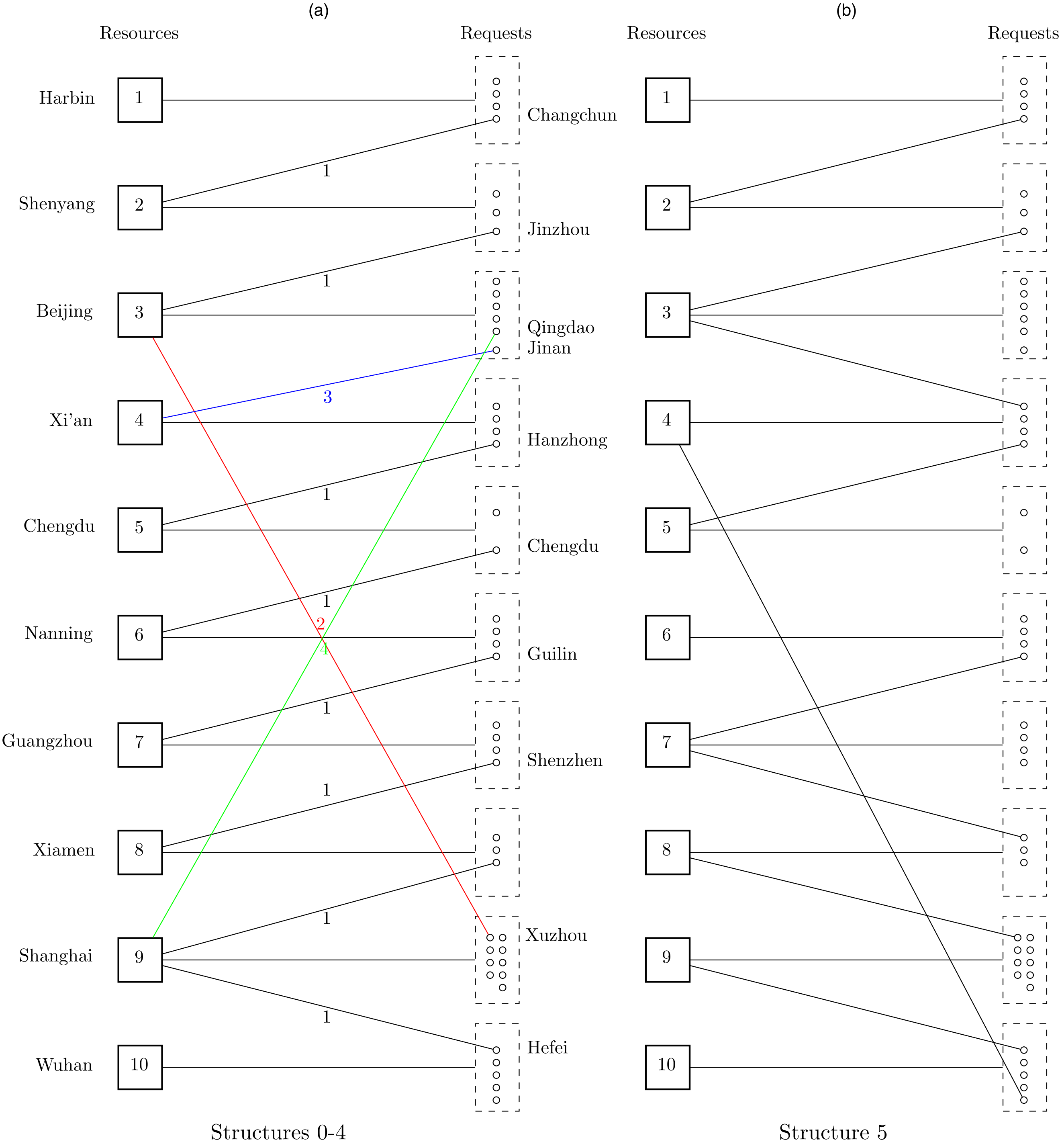
9.2. Comparisons Between Our Policy and a Myopic Policy Under Different Network Structures

We introduce a myopic order fulfillment and inventory allocation policy that is similar to Amazon China's practice and also mentioned as a common practice in industry by Acimovic and Graves (2014). Under the myopic policy, one allocates to each resource the amount of inventory equal to the average demand of the region and fulfills a request by its primary resource if there is inventory left or the closest resource

Table 1. Amazon China's Fulfillment Centers and Their Regional Demand Centers with the Demand Vector

Resources	Requests	p
Harbin	Harbin, Daqing, Changchun, Yanbian	(0.019, 0.019, 0.013, 0.013)
Shenyang	Shenyang, Jinzhou, Dalian	(0.017, 0.022, 0.022)
Beijing	Beijing, Tianjin, Tangshan, Shijiazhuang, Jinan, Qingdao	(0.057, 0.015, 0.015, 0.029, 0.020, 0.019)
Wuhan	Wuhan, Nanchang, Changsha, Hefei, Xiangyang	(0.029, 0.029, 0.021, 0.009, 0.012)
Shanghai	Shanghai, Suzhou, Hangzhou, Ningbo, Wenzhou, Changzhou, Wuxi, Nanjing, Xuzhou	(0.026, 0.026, 0.026, 0.026, 0.026, 0.019, 0.026, 0.026, 0.057)
Xiamen	Xiamen, Fuzhou, Ganzhou	(0.009, 0.007, 0.019)
Guangzhou	Guangzhou, Foshan, Shenzhen, Dongguan	(0.045, 0.045, 0.045, 0.028)
Nanning	Nanning, Kunming, Guilin, Guiyang	(0.006, 0.010, 0.010, 0.011)
Chengdu	Chengdu, Chongqing	(0.028, 0.036)
Xian	Xian, Hanzhong, Zhengzhou, Taiyuan	(0.009, 0.018, 0.019, 0.019)

Figure 13. (Color online) Network Structures Considered for Amazon China



allowed by the network structure. Such an inventory allocation makes sense as each region has a primary fulfillment center at Amazon China.

Under structures 3 and 4 with a total number of $I + J$ and $I + J + 1$ arcs, we plot the expected number of lost sales and average shipping distance under our inventory allocation and demand fulfillment policy with that of the myopic one in Figure 15. As we can see, our policy outperforms the myopic one with lower lost sales and slightly higher shipping distances. This is expected as our policy aims to fulfill more demand and ignores the fact that the arcs have different distances.

To balance the GCG and the total shipping distance, we also construct network structures that take into account the shipping distance between the resources and demand centers explicitly. We first link each request node to its nearest fulfillment center to form I demand regions and their primary fulfillment centers.

Figure 14. Performance Under Different Network Structures and for Various K for Amazon China

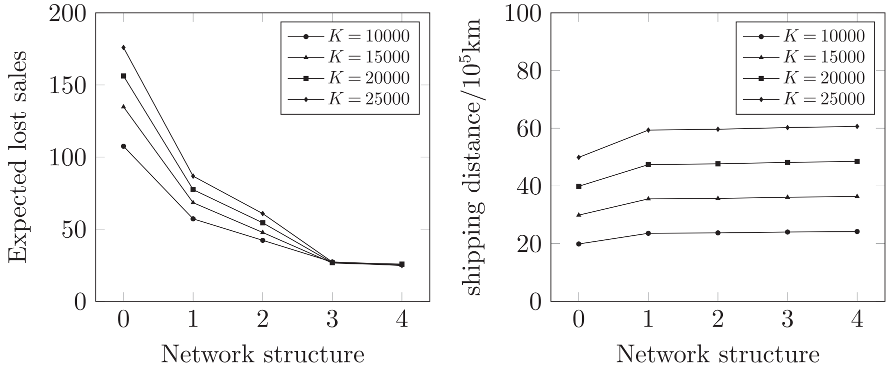
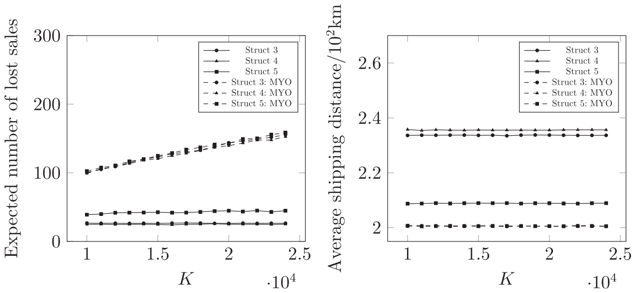


Figure 15. Performance Comparisons Between Our Policy and the Myopic Policy Under Structures 3–5 for Amazon China.



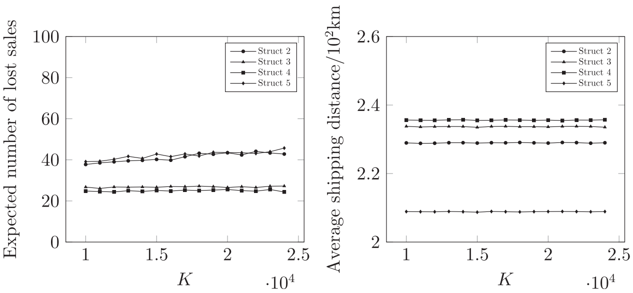
Note. Solid lines represent our policy; dashed lines represent the myopic policy.

We then solve the following integer program to determine the remaining cross-region arcs:

$$\begin{aligned}\min \; & \sum_{i,j=1,i\neq j}^{I} d_{ij}x_{ij}\\ \text{s.t.} \; & \sum_{i,j=1,i\neq j}^{I} x_{ij} = \text{the total number of arcs } -J,\\ & x_{ij} \in \{0,1\}, \text{the structure } \mathscr{A} \text{ is connected.}\end{aligned}$$

Here $x_{ij} = 1$ if fulfillment center i is linked to a closest request node in region j and d_{ij} is the distance between them. That is, the regions are connected by a fulfillment center in one region to the closest request node in another region. This optimization problem is not trivial in general. However, for network structures with a total of $I + J - 1$ arcs, the problem reduces to finding a minimum spanning tree on an edge-weighted graph, and we can apply Kruskal's algorithm to obtain an optimal solution for the integer programming. The resulting network structure, referred to as structure 5, is presented in Figure 13(b). In Figure 15, we also plotted the performance of our policy and the myopic policy under structure 5. As one can see, structure 5 achieves the lowest average shipping distance under each policy without losing much sales.

Figure 16. Performance Under Different Network Structures and Random **p** for Amazon China



9.3. Robustness of Our Policy With Uncertain Demand Rates

We test the robustness of our policy with uncertain demand rates in this section. Following Asadpour et al. (2019), we randomly generate the demand vector for the kth arrival from the uncertainty set $U_p = \{\hat{\mathbf{p}} | \hat{p}_i \in [p_i - \epsilon, p_i + \epsilon]$ for each $1 \le i \le I$ and $\sum_{i=1}^{I} \hat{p}_i = 1\}$ for a given ϵ. We vary $\epsilon = p_{\min}^2, p_{\min}^{2.5}, p_{\min}^3$, reflecting a decreasing level of uncertainty of the demand vector. Since the results are similar, we only report the simulation results for $\epsilon = p_{\min}^2$ in Figure 16. As one can see, the performance is quite stable.

10. Conclusions

Process flexibility has been studied extensively under offline fulfillment but not under online fulfillment except for the work of Asadpour et al. (2019), who focus on the long chain structure, which is a balanced system. They establish bounded performance of the long chain structure under a specific inventory allocation as the market size increases. In this paper, we extend their greedy fulfillment policy to a class of unbalanced systems called GCG systems under online fulfillment, where the number of request types can be arbitrarily larger than the number of resources, and establish bounded performance. We further extend bounded performance to systems with random batch arrivals and time-varying demand rates.

The upper bound on system performance also reveals that the GCG is an important indicator of system performance, which leads to simple inventory allocation decisions for any connected network structure with as few as $I + J - 1$ arcs that guarantee a positive GCG and achieves bounded performance. We also provide principles for the design of network structures that achieve bounded performance. For networks with $I + J$ arcs, we extend the long chain concept to unbalanced networks, referred to as generalized long chains (GLCs), by dividing the request types into I groups and forming a network structure with I resources and request groups. Numerical studies including one using some data from Amazon China are conducted to verify our findings.

Acknowledgments

The authors thank the department editor, associate editor, and three anonymous reviewers for their careful reading of the paper, and for providing constructive feedback.

Endnotes

[1] See https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/.

[2] See http://data.stats.gov.cn/english/easyquery.htm?cn=E0103.

References

Acimovic J, Graves SC (2014) Making better fulfillment decisions on the fly in an online retail environment. *Manufacturing Service Oper. Management* 17(1):34–51.

Asadpour A, Wang X, Zhang J (2019) Online resource allocation with limited flexibility. *Management Sci.*, ePub ahead of print September 5, https://pubsonline.informs.org/doi/10.1287/mnsc.2018.3220.

Bassamboo A, Mieghem JAV, Randhawa RS (2010) Optimal flexibility configurations in newsvendor networks: Going beyond chaining and pairing. *Management Sci.* 56(8):1285–1303.

Bušić A, Meyn S (2015) Approximate optimality with bounded regret in dynamic matching models. *ACM SIGMETRICS Performance Evaluation Rev.* 43(2):75–77.

Bušić A, Gupta V, Mairesse J (2013) Stability of the bipartite matching model. *Adv. Appl. Probab.* 45(2):351–378.

Chen X, Zhang J, Zhou Y (2015) Optimal sparse designs for process flexibility via probabilistic expanders. *Oper. Res.* 63(5):1159–1176.

Chen X, Ma T, Zhang J, Zhou Y (2019) Optimal design of process flexibility for general production systems. *Oper. Res.* 67(2):516–531.

Chou MC, Teo C-P, Zheng H (2008) Process flexibility: Design, evaluation, and applications. *Flexible Services Manufacturing J.* 20(1–2):59–94.

Chou MC, Chua GA, Teo C-P, Zheng H (2010) Design for process flexibility: Efficiency of the long chain and sparse structure. *Oper. Res.* 58(1):43–58.

Chou MC, Chua GA, Teo C-P, Zheng H (2011) Process flexibility revisited: The graph expander and its applications. *Oper. Res.* 59(5):1090–1105.

Deng T (2013) Process flexibility design in unbalanced and asymmetric networks. PhD thesis, University of California, Berkeley.

Désir A, Goyal V, Wei Y, Zhang J (2016) Sparse process flexibility designs: Is the long chain really optimal? *Oper. Res.* 64(2):416–431.

Ding Y, McCormick S, Nagarajan M (2018) A fluid model for an overloaded bipartite queueing system with heterogeneous matching utility. Working paper, University of British Columbia, Vancouver, BC, Canada.

Feldman J, Mehta A, Mirrokni V, Muthukrishnan S (2009) Online stochastic matching: Beating 1-1/e. *Proc. 50th Annual IEEE Symp. Foundations Comput. Sci. (FOCS),* (IEEE, Piscataway, NJ), 117–126.

Graves SC, Tomlin BT (2003) Process flexibility in supply chains. *Management Sci.* 49(7):907–919.

Gurumurthi S, Benjaafar S (2004) Modeling and analysis of flexible queueing systems. *Naval Res. Logist.* 51(5):755–782.

Iravani SMR, Kolfal B, Van Oyen MP (2007) Call-center labor cross-training: It's a small world after all. *Management Sci.* 53(7):1102–1112.

Iravani SMR, Van Oyen MP, Sims KT (2005) Structural flexibility: A new perspective on the design of manufacturing and service operations. *Management Sci.* 51(2):151–166.

Jaillet P, Lu X (2013) Online stochastic matching: New algorithms with better bounds. *Math. Oper. Res.* 39(3):624–646.

Jasin S, Sinha A (2015) An LP-based correlated rounding scheme for multi-item ecommerce order fulfillment. *Oper. Res.* 63(6):1336–1351.

Jordan WC, Graves SC (1995) Principles on the benefits of manufacturing process flexibility. *Management Sci.* 41(4):577–594.

Mandelbaum A, Stolyar AL (2004) Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized c\mu-rule. *Oper. Res.* 52(6):836–855.

Manshadi VH, Gharan SO, Saberi A (2012) Online stochastic matching: Online actions based on offline statistics. *Math. Oper. Res.* 37(4):559–573.

Shen Z-JM, Deng T (2013) Process flexibility design in unbalanced networks. *Manufacturing Service Oper. Management* 15(1):24–32.

Sheng L, Zheng H, Rong Y, Huh WT (2015) Flexible system design: A perspective from service levels. *Oper. Res. Lett.* 43(3):219–225.

Shi C, Wei Y, Zhong Y (2019) Process flexibility for multiperiod production systems. *Oper. Res.*, ePub ahead of print August 6, https://doi.org/10.1287/opre.2018.1810.

Simchi-Levi D, Wei Y (2012) Understanding the performance of the long chain and sparse designs in process flexibility. *Oper. Res.* 60(5):1125–1141.

Simchi-Levi D, Wei Y (2015) Worst-case analysis of process flexibility designs. *Oper. Res.* 63(1):166–185.

Simchi-Levi D, Wang H, Wei Y (2018) Increasing supply chain robustness through process flexibility and inventory. *Production Oper. Management* 27(8):1476–1491.

Tanrisever F, Morrice D, Morton D (2012) Managing capacity flexibility in make-to-order production environments. *Eur. J. Oper. Res.* 216(2):334–345.

Tsitsiklis JN, Xu K (2017) Flexible queueing architectures. *Oper. Res.* 65(5):1398–1413.

Van Roy B, Bertsekas DP, Lee Y, Tsitsiklis JN (1997) A neuro-dynamic programming approach to retailer inventory management. *Proc. 36th IEEE Conf. Decision Control*, vol. 4 (IEEE, Piscataway, NJ), 4052–4057.

Wallace RB, Whitt W (2005) A staffing algorithm for call centers with skill-based routing. *Manufacturing Service Oper. Management* 7(4):276–294.

Wang X, Zhang J (2015) Process flexibility: A distribution-free bound on the performance of k-chain. *Oper. Res.* 63(3):555–571.

Xu PJ, Allgor R, Graves SC (2009) Benefits of reevaluating real-time order fulfillment decisions. *Manufacturing Service Oper. Management* 11(2):340–355.

Zhen Xu is currently working as an assistant professor in the School of Management at Fudan University. His research lies at the interface of operations research, applied probability, and machine learning, with a focus on stochastic process control, dynamic programming, supply chain management, and data-driven decision making for a variety of static and dynamic models.

Hailun Zhang is an assistant professor in data and decision analytics at the Chinese University of Hong Kong, Shenzhen. His research interests include data-driven queueing networks, online algorithm design, stochastic modeling, and optimization.

Jiheng Zhang is an associate professor in industrial engineering and decision analytics at the Hong Kong University of Science and Technology. His research interests are in applied probability, stochastic modeling and optimization, data analysis, numerical methods, and algorithms.

Rachel Q. Zhang is a chair professor in industrial engineering and decision analytics at the Hong Kong University of Science and Technology. Her research interests include supply chain and inventory management, stochastic analysis of service operations, and the interface of finance and operations.

CORRECTION

In this article, "Online Demand Fulfillment Under Limited Flexibility" by Zhen Xu, Hailun Zhang, Jiheng Zhang, and Rachel Q. Zhang (first published in *Articles in Advance*, June 12, 2020, *Management Science*, DOI:10.1287/mnsc.2019.3449), the author affiliation, email address, and author biography for Zhen Xu have been corrected to include his current affiliation.