Working Paper

LP-Based Artificial Dependency for Probabilistic Etail Order Fulfillment

Stefanus Jasin
Stephen M. Ross School of Business
University of Michigan

Amitabh Sinha
Stephen M. Ross School of Business
University of Michigan

Ross School of Business Working Paper
Working Paper No. 1250
October 2014

LP-Based Artificial Dependency for Probabilistic Etail Order Fulfillment

Stefanus Jasin, Amitabh Sinha
Stephen M. Ross School of Business, University of Michigan, Ann Arbor, MI 48109, USA.
sjasin@umich.edu, amitabh@umich.edu

October 5, 2014

Abstract

We consider an online multi-item retailer with multiple fulfillment facilities and finite inventory, with the objective of minimizing the expected shipping cost of fulfilling customer orders over a finite horizon. We approximate the stochastic dynamic programming formulation of the problem with an equivalent deterministic linear program, which we use to develop a probabilistic fulfillment heuristic that is provably optimal in the asymptotic sense. This first heuristic, however, relies on solving an LP that is exponential in the size of the input. Therefore, we subsequently provide another heuristic which solves an LP that is polynomial in the size of the input, and prove an upper bound on its asymptotic competitive ratio. This heuristic works by modifying the LP solution with artificial dependencies, with the resulting fractional variables used to probabilistically fulfill orders. A hardness result shows that asymptotically optimal policies that are computationally efficient cannot exist. Finally, we conduct numerical experiments that show that our heuristic's performance is very close to optimal for a range of parameters.

1 Introduction

E-commerce retail sales in the US in the twelve months ending in September 2013 exceeded $250 billion (U.S. Department of Commerce 2013). Although this constitutes only around 6% of total retail sales in the US, the growth rates of e-commerce retail versus traditional retail (15% versus 4%) leave little doubt about how voluminuous this sector will be over the next few years—a fact that should be of little surprise to anyone engaged in retail purchases or sales. When one considers the distribution logistics of the e-commerce retail (henceforth called etail) industry, there is one significant (albeit obvious) difference from traditional brick-and-mortar retailers: *The etailer can choose where to fulfil the orders from.* This has several benefits. First, it enables the etailer to minimize total shipping costs. Some etailers offer membership schemes whereby, in exchange for an annual fee, customers never have to pay shipping cost (e.g. Amazon.com's Prime program), thus incenting the etailer to minimize shipping costs. Others follow standardized shipping costs that are displayed to customers, but that still leave an opportunity for the etailer to minimize its actually incurred costs. In addition to minimizing shipping

costs, the ability to decide where to serve the orders from allows an etailer to balance inventory, avoid congestion, and further optimize its stocking and supply decisions.

Despite the importance of the optimization of fulfillment decisions by etailers, in terms of actual practice and academic literature, very little progress has been made in this area. Xu et al. (2009) report that many retailers simply follow a "myopic" policy, where orders received only over the past few hours are considered and served in a cost-minimizing fashion, with no consideration on the impact to future costs. As we discuss in Section 1.2, academic research on this question also leaves several open questions as well as significant opportunity to find savings by making the fulfillment policies more efficient. This is in stark contrast to the distribution logistics of brick-and-mortar retailers, where several decades of research have given us a strong understanding on inventory policies, network design, transshipment policies, etc.

In order to better illustrate the opportunity to lower costs by making decisions in a forward-looking rather than myopic fashion, it helps to consider a simple example. Consider a firm that has a network of two distribution centers (henceforth abbreviated as DC) to serve various regions in the US: one in Georgia and one in California. We focus on two products, labeled A and B, and any given customer may demand one of the four combinations: $\{\}, \{A\}, \{B\}, \{A, B\}$. There is a positive probability for each of the four combinations. Also, suppose that each product has weight 1 pound. A customer from Miami, Florida, has just placed an order on the company's website for the combination $\{B\}$. The inventory position in the Georgia and California DCs are given by (S_{GA}, S_{GB}) and (S_{CA}, S_{CB}) respectively, where both vectors are currently strictly positive in both components. Should the customer's order be fulfilled by the Georgia DC? As one would expect, the fulfilment decision will be found to depend on shipping costs as well as the demand distribution of these products over the remaining time horizon. Suppose that the firm uses UPS' 3 day select service to ship products. From the UPS Standard Rate and Service Guide (UPS 2012), we find that the costs of shipping are as shown in Table 1.

Weight (lbs.)	From Georgia	From California
1	\$10.95	\$15.50
2	\$11.60	\$18.05

Table 1: UPS shipping rates to Miami, using 3 day select service, 2012.

Consider the case when $S_{GB} = 1$, that is, only 1 unit of inventory of B is available in Georgia, and a large number of customer orders are expected to arrive before replenishment. In contrast, the inventory of A is high enough that there is no possibility of a stockout before replenishment. If the order of $\{B\}$ is fulfilled from Georgia, the firm incurs a shipping cost of \$10.95. Suppose the firm had, instead, redirected the customer order to the California DC. Then, the shipping cost for this order would have been \$15.50. This would make sense only if the firm was saving the remaining unit of B for an order consisting of the combination $\{A, B\}$. By how much would the firm's shipping costs have changed if it had done this? The cost for the Miami customer goes up by $\$15.50 - \$10.95 = \$4.55$. However, when the order for $\{A, B\}$ arrived, the firm would have been able to serve it from Georgia instead of California, saving $\$18.05 - \$11.60 = \$6.45$ on that order, resulting in a net saving of \$1.90. All other orders would have been unaffected. Given our assumption of a large number of customers

remaining before the next replenishment and strictly positive probability of a customer order consisting of the combination $\{A, B\}$, this saving is virtually guaranteed. Although the example above features two products, it is in fact the case that even with a single product, a myopic policy (which assigns each arriving order to the least-cost facility that can serve it) is not optimal.

The example above brings us to our research question: How can firms implement policies that direct online orders to appropriate fulfillment centers in a way that uses information about future demand distribution and inventory positions in order to minimize expected total fulfillment cost over the entire horizon? Although the problem is fairly straightforward to identify and define (a formal definition appears in Section 2), it is not easily amenable to standard techniques. The main hurdle in applying standard inventory theory algorithms or policies is the strongly combinatorial nature of the problem: some orders contain more than one item, and the shipping costs scale in a way that splitting such orders (i.e. treating each multi-item order as multiple single-item orders) is very costly. In fact, *even the problem of figuring out the allocation of items to fulfillment centers for a single multi-item order is NP-hard*, as will be discussed in Section 7. Our approach, of using a linear program (LP) with asymptotic scaling, allows us to sidestep this difficulty while providing heuristics with provably good performance guarantees.

1.1 This paper and our contributions

Our approach is based on the idea of using the LP relaxation of the asymptotically scaled version of the problem as a lower bound. A brief description of what this means is as follows. First, we consider the problem over a fixed finite horizon T, which we will eventually scale to infinity. (We will discuss the appropriateness of this setting later.) Within this finite horizon, at time zero, we construct a deterministic linear program (DLP) by replacing the stochastic demand for each bundle of items with their expected value. We solve this DLP and treat the fractional solution as a probability distribution that determines how actual orders are assigned to warehouses when they begin arriving. Note that this approach fixes the assignment of orders to warehouses up front, and does not change them once the actual demand starts arriving. Of course, the actual demand sequence will not match the expected demand. So, our actual costs will be higher than the cost computed by our heuristic. In addition, our heuristic considers the fractional relaxation, causing a second source of gap between our actual costs and our lower bound. However, we are able to prove that, as the time horizon goes to infinity (with demand and inventory levels also scaling up at the same rate), the gap between the actual cost and our expected cost is bounded by an additive term that scales with the square-root of the time horizon.

One drawback of the above approach is that the lower bound is computed via an LP with exponential size. Consequently, solving such an LP may be difficult for some real instances. Therefore, we also construct an approximate LP that is polynomial in the size of the input. We first show that simply interpreting the fractional solution of the LP as probabilities to guide the fulfillment decision results in a heuristic with a bounded competitive ratio. But a careful examination of this heuristic suggests that independent probabilistic fulfillment can be inefficient, and there may be a way to further improve the competitive ratio. This brings us to the main contribution of this paper: *We develop a heuristic that*

uses the LP solution to construct a probabilistic fulfillment control that artificially injects dependencies into the fulfillment decisions, and prove a better upper bound on the asymptotic competitive ratio of this heuristic.

We supplement our work with numerical experiments, where we find that the competitive ratio of our algorithm is very close to 1 for a wide range of parameter values, with the heuristic running very fast particularly when compared with the DLP. Additionally, we discuss the hardness of this problem from two perspectives: we establish a lower bound on the competitive ratio for any heuristic for this problem via a reduction of the set cover problem, and we also show that the integrality gap of our LP formulation precludes the existence of a rounding scheme that is provably optimal.

The rest of this paper is organized as follows. We provide a brief survey of the literature below. In Section 2, we define our model and the notations used. Our DLP formulation, which we show to be asymptotically optimal if used as a probabilistic control, is developed in Section 3. We then develop an approximate LP that is polynomial in the size of the input in Section 4. We derive the competitive ratio of a heuristic that naively uses the LP solution as independent fulfillment probabilities. We then develop our dependent probabilistic fulfillment scheme in Section 5, and derive its competitive ratio. In Section 6, we show the effectiveness of our heuristics using numerical experiments. We discuss lower bounds on the competitive ratio for this problem in Section 7, before concluding with a brief discussion on future research in Section 8.

1.2 Literature Review

The first paper to explicitly model and propose a solution strategy for etail order fulfillment was Xu et al. (2009). They propose a heuristic which periodically re-evaluate all orders that have been assigned to warehouses but not yet picked and then re-assign orders with the goal of minimizing the total number of shipments. They numerically demonstrate that this approach reduces the number of orders that were initially split by about 50%. Our paper adds two additional layers of complexity to their model: We consider total shipping cost rather than number of shipments and we also incorporate demand forecasts into our model, both of which make the problem significantly harder.

More recently, Acimovic et al. (2012) considered the problem of minimizing total shipping costs for single-item orders by proposing a heuristic that assigns orders to warehouses based on dual values of an LP that incorporates the expected cost of fulfilling future orders. They report that, based on their data set, the total opportunity for saving on outbound transportation costs is of the order of 2%, and their heuristic saves approximately one-fourth of it. Two other papers that consider the benefit of assigning orders non-myopically are Mahar and Wright (2009) and Cattani and Souza (2002). Both consider firms that are dual-channel, i.e., they sell online as well as via brick-and-mortar stores. They show that rationing inventory for the online channel can be beneficial under certain circumstances, but they do not consider the multiproduct fulfillment decision that is the thrust of this paper.

Other than the papers cited above, there appear to be no studies of the etail order fulfillment problem that we are aware of. The larger area of studying various aspects of supply chains in an era of electronic communication has, of course, seen substantial research. We refer the reader to Agatz et al.

(2008) and Simchi-Levi et al. (2004) for some recent reviews.

In terms of methodology, our approach is perhaps closest to that in the revenue management (RM) literature. Similar to ours, in a typical RM setting, we are dealing with large-scale stochastic problems which cannot be exactly solved using the standard dynamic program (DP) formulation. This has motivated many researchers to develop heuristics which are easy to implement and yet at the same have a respectable performance. Among these is a class of heuristics constructed using the solution of the LP formulation of RM problem. See, for example, Liu and van Ryzin (2008), Reiman and Wang (2008), Ciocan and Farias (2012), and Jasin and Kumar (2012, 2013). Similar to these papers, we start with a deterministic formulation of the etail order fulfillment problem and then use its solution to construct a heuristic with a competitive performance guarantee.

On the surface, our model may appear to be similar to a minimum cost network flow problem. Indeed, if the shipping costs were directly proportional to the number of items (i.e., no economies of scale in shipping two or more items in a single package), the deterministic continuous relaxation of our problem is a simple multi-commodity network flow problem, which can be directly solved using standard network optimization techniques (Ahuja et al. 1993, Ch. 9–11,17), as well as standard linear or integer programming techniques (Simchi-Levi et al. 2004, Ch. 15,17). However, the fixed cost per shipment, discrete nature of the problem, and multi-period feature all add additional complications that prevent the direct applicability of these techniques in a way that is provably fast and with guarantees on the performance.

2 Model Description

2.1 Basic Setting

Let $\mathcal{S}_I$, $\mathcal{S}_K$, $\mathcal{S}_J$, and $\mathcal{S}_Q$ denote the set of items, facilities, regions, and order types, indexed by i, k, j and q respectively. Regions are customer locations from which orders arrive and order types are characterized by the unique composition of items contained in the order. For example, an order type $q = 1$ may correspond to a request for item $i = 1$ and 2 and an order type $q = 2$ may correspond to a request for item 1 and 3. In general, an order may contain more than one request for the same item. This suggests that a proper definition of order type must also account for the number of requests per item contained in the order. However, since such multi-request orders are quite rare in practice (Xu et al. 2009), in this paper, we only allow at most one request per item. Mathematically, we can treat multi-request orders as separate orders without significantly affecting expected total cost. We write $q \ni i$ (or $i \in q$) if order type q contains item i (similarly, item i is requested in order type q).

The selling horizon is divided into T periods and at most one order arrives during each period. This is without loss of generality since we can always slice the selling horizon fine enough to ensure that at most one customer arrives during each period. In addition, as in Xu et al. (2009) and Acimovic et al. (2012), we also assume that no inventory replenishment occurs during the selling horizon. This assumption is motivated by the fact that the number of orders arriving between two subsequent replenishment times is usually large, which provides ample opportunities for the etailer to implement

clever fulfillment heuristics in order to minimize expected total shipping cost. For example, according to Amazon.com's Press Releases (www.amazon.com), the etail giant sold 26.5 million items worldwide on the peak day of Nov. 26 during the holiday season 2012. This record-breaking number is equivalent to selling 306 items per second. That said, we do allow for a way to account the cost of stock-outs by routing excess requests to a designated artificial facility with appropriate costs (see discussion below). (The joint optimization of inventory replenishment and order fulfillment is an important research problem. We leave this for future research pursuit.) Additionally, we assume that each order must be fulfilled immediately, and the etailer cannot deliberately hold back an order for later shipping. In practice, etailers are moving towards faster and faster delivery of items, justifying this assumption.

Let λ_j^q denote the arrival probability of order type q from region j during any period t and let λ_0 denote the probability of no arrival. By definition, we must have $\lambda_0 + \sum_j \sum_q \lambda_j^q = 1$. (We implicitly assume that demands are time-homogeneous. This is only for expositional simplicity since our results can also be easily extended to the case of non-homogeneous demands.) The initial inventory of item i at facility k is given by $S_{ki} > 0$. Per our discussion above, without loss of generality, we will assign facility 1 as the *back-up* facility by setting $S_{1i} = +\infty$ for all i. Facility 1 is therefore a fictitious facility, with transportation costs set by us, to model what the firm does if the item is not shipped from any of its "real" facilities. Generally, a shipment from facility 1 would indicate that the item is stocked-out at all real facilities. But, in principle, the firm is allowed to assign some items/orders to facility 1 even if inventory is available (e.g., if the costs and demand forecasts make it cost-efficient to do so). The cost of shipping from facility 1 models the real cost of whatever the firm does if it is unable to ship from its real facilities, which could include drop-shipping from suppliers, delaying the order until the next replenishment, or simply reneging on the order and paying a penalty for it. Facility 1 therefore guarantees that we always have feasible solutions to our problems. For most of the sequel we will not need to refer to facility 1's special role, other than in the numerical study where we will deliberately impose a higher cost on shipments from facility 1 in recognition of its special role.

Let $X_{kij}^{qt} \in \{0, 1\}$ be a random variable denoting the etailer's decision whether to ship item i in order type q coming from region j during period t from facility k. It is important to note that we allow different items in the same order to be fulfilled from different facilities. Indeed, this combinatorial aspect of the problem is one of the key features of etail order fufillment. Since all items must be fulfilled, in terms of the decision variable, the following must always be satisfied:

$$\sum_k X_{kij}^{qt} = 1 \qquad \forall i \in q.$$

Let $D_j^{qt} \in \{0, 1\}$ be a random variable denoting the realized demand for order type q from region j during period t, i.e. $D_j^{qt} = 1$ if an order type q arrives from region j during period t and 0 otherwise. For any sequence of realized demands $\{D_j^{qt}\}$, the following set of inventory constraints must be satisfied almost surely (or with probability one):

$$\sum_t \sum_j \sum_{q \ni i} D_j^{qt} X_{kij}^{qt} \leq S_{ki} \qquad \forall\ k, i.$$

To mimic the typical cost structures often found in practice, such as the UPS rates discussed in the introduction, we model the outbound shipping cost using two components: *variable cost* and *fixed cost*. The variable cost of shipping item i from facility k to region j is denoted by c_{kij} and the fixed cost of shipping from facility k to region j at all is denoted by b_{kj}. Thus, we are allowing different items to have different variable shipping costs (for instance, if they differ in size or weight). Using this notation, the total shipping cost for an order type q arriving from region j during period t can be written as:

$$\sum_k \left[\sum_{i \in q} c_{kij} \, X_{kij}^{qt} \; + \; b_{kj} \max_{i \in q} \{ X_{kij}^{qt} \} \right].$$

The optimal control formulation of Etail Order Fulfillment (EOF) is given by

$$\begin{aligned}
J^* \;:=\; \min \quad & \sum_t \sum_j \sum_q \sum_k \mathbf{E} \left[D_j^{qt} \left(\sum_{i \in q} c_{kij} \, X_{kij}^{qt} \; + \; b_{kj} \max_{i \in q} \{ X_{kij}^{qt} \} \right) \right] \\
\text{s.t.} \quad & \sum_t \sum_j \sum_{q \ni i} D_j^{qt} X_{kij}^{qt} \le S_{ki} \quad \forall k, i & (1) \\
& \sum_k X_{kij}^{qt} = 1 \quad \forall q, i, j, t & (2) \\
& X_{kij}^{qt} \in \{0, 1\} \quad \forall i, j, k, q, t & (3)
\end{aligned}$$

where the minimization is taken over the set of non-anticipating policies (i.e. the shipping decision during period t depends only on the accumulated information up to the beginning of period t).

2.2 Asymptotic Scaling and Performance Measure

Motivated by the large number of daily orders in etail industry, especially during high seasons, in this paper, we will consider a sequence of increasing problems where both the number of selling periods and the amount of initial inventories are scaled by a factor of $\theta > 0$. To be precise, in the θ^{th} problem, the number of selling periods is given by $T(\theta) = \theta T$ and the number of initial inventories by $\{S_{ki}(\theta) = \theta S_{ki}\}$. Since multiplying the number of selling periods by θ is equivalent to multiplying the number of average demands by θ, in the so-called *asymptotic* setting, we essentially increase both total demands and total inventories while preserving their relative proportion. (Under a proper scaling of T and $\{S_{ki}\}$, the factor θ can be interpreted as the size of the problem. For example, $\theta = 100$ may correspond to a problem instance with total demands, and total inventories, about 100 whereas $\theta = 1000$ may corresponds to a larger instance with total demands, and total inventories, about 1000.) The use of scaling factor in performance analysis is not new and has been a standard methodology in the study of queueing systems (Halfin and Whitt 1981, Harrison 1998, Maglaras 2000, Ata and Kumar 2005), revenue management and dynamic pricing (Gallego and van Ryzin 1994, 1997, Cooper 2002, Levi and Radovanović 2010, Jasin and Kumar 2012, 2013), and inventory control (Huh et al. 2009a,b, Plambeck and Ward 2006, Plambeck 2008). It is particularly useful to study the performance of a heuristic in the

setting of large demands and large inventories. Let $C_\pi(\theta)$ denote the total realized cost under heuristic π and let $J^*(\theta)$ denote the expected total cost under optimal control, both for a problem with scaling factor θ. We are interested in the following limit, which will be used as our performance measure:

$$\lim_{\theta\to\infty} \frac{\mathbf{E}[C_\pi(\theta)]}{J^*(\theta)}.$$

By definition of $J^*(\theta)$, the above ratio is always greater than or equal to 1. It captures the first-order magnitude of expected total cost under heuristic π. Although, in reality, demands and inventories are always finite, the limiting ratio as $\theta \to \infty$ still serves as a good proxy for the performance of a given heuristic in the setting of large demands and large inventories. Indeed, numerical results in Section 6 show that our proposed heuristic has a strong performance even when θ is relatively small. The following conventions will be used throughout the rest of this paper. A heuristic π is said to be *asymptotically optimal* if $\lim_{\theta\to\infty} \mathbf{E}[C_\pi(\theta)]/J^*(\theta) = 1$. In addition, it is said to be *asymptotically α-competitive* if $\lim_{\theta\to\infty} \mathbf{E}[C_\pi(\theta)]/J^*(\theta) \leq \alpha$; we sometimes refer to α as the *competitive ratio* of the heuristic π.

3 The Exact LP

In theory, the optimal control for EOF can be exactly solved using dynamic program (DP). Unfortunately, the well-known curse of dimensionality quickly kicks in even for problems with moderate size. Thus, despite being optimal, the DP approach is computationally intractable and simply practically infeasible. This motivates us to find an approximate solution which can be used to construct a near-optimal heuristic. In this section, we will consider a deterministic formulation of EOF, which we call the *exact* LP. (We call it "exact" because, as we will see shortly, it leads to an asymptotically optimal heuristic for EOF.) Let $\sigma_j^q : \mathcal{S}_I \mapsto \mathcal{S}_K$ denote the fulfillment assignment vector for order type q coming from region j, i.e., $\sigma_j^q(i) = k$ means we are shipping item $i \in q$ to region j from facility k. Also, let $G_j^{qt}(\sigma_j^q) = P(X_j^{qt} = \sigma_j^q \,|\, D_j^{qt} = 1)$ denote the probability of fulfilling order type q from region j during period t with σ_j^q. (The expression $X_j^{qt} = \sigma_j^q$ is shorthand for $X_{kij}^{qt} = \mathbf{1}\{\sigma_j^q(i) = k\}$ for all $i \in q$.) Given a demand realization $D_j^{qt} = 1$, we can write:

$$\begin{aligned}
\mathbf{E}\left[X_{kij}^{qt}\right] &= \sum_{\{\sigma_j^q:\, \sigma_j^q(i)=k\}} G_j^{qt}(\sigma_j^q) \quad \text{and} \\
\mathbf{E}\left[\max_{i\in q}\{X_{kij}^{qt}\}\right] &= \sum_{\{\sigma_j^q:\, \exists i\in q,\, \sigma_j^q(i)=k\}} G_j^{qt}(\sigma_j^q).
\end{aligned}$$

Thus, taking expectation over the constraints in J^* yields a lower bound:

$$
\begin{aligned}
J^* \;\geq\; J_{LP} \;:=\; \min_{u} \quad & \sum_{t,j,q,k} \lambda_j^q \left[\sum_{i \in q} \sum_{\{\sigma:\, \sigma(i)=k\}} c_{kij}\, u_{\sigma j}^{qt} \;+ \sum_{\{\sigma:\, \exists i \in q,\, \sigma(i)=k\}} b_{kj}\, u_{\sigma j}^{qt} \right] \\
\text{s.t.} \quad & \sum_{t} \sum_{j} \sum_{q \ni i} \lambda_j^q \left[\sum_{\{\sigma:\, \sigma(i)=k\}} u_{\sigma j}^{qt} \right] \leq S_{ki} \quad \forall k, i & (4) \\
& \sum_{k} \sum_{\{\sigma:\, \sigma(i)=k\}} u_{\sigma j}^{qt} = 1 \quad \forall q, t, j & (5) \\
& 0 \leq u_{\sigma j}^{qt} \leq 1 \quad \forall q, t, j, \sigma & (6)
\end{aligned}
$$

where for brevity we simply write σ instead of σ_j^q. (We will continue using this convention throughout the remainder of the paper provided there is no confusion on the meaning of σ.) Per our notations above, any policy essentially corresponds to a set of distributions $\{G_j^{qt}\}$ over the set of fulfillment assignment $\{\sigma_j^q\}$. Thus, we can immediately see that for any non-anticipating policy, constraints (1)-(3) imply (4)-(6) in expectation. Thus, proving J_{LP} is a lower bound of J^*.

Observe that (5) and (6) can be simplified to $\sum_\sigma u_{\sigma j}^{qt} = 1$ and $u_{\sigma j}^{qt} \geq 0$. This is so because, for each triplet (q, i, j) where $i \in q$, the set of fulfillment vectors $\{\sigma_j^q\}$ can be decomposed into $\cup_k \{\sigma_j^q : \sigma_j^q(i) = k\}$. If we now define $c_{\sigma j}^q := \sum_{i \in q} c_{\sigma(i) ij} + \sum_{\{k:\, \exists i \in q,\, \sigma(i)=k\}} b_{kj}$ (it can be interpreted as the cost of applying assignment σ to order type q from region j), we can rewrite J_{LP} in a more compact form as follows:

$$
\begin{aligned}
J_{LP} \;:=\; \min \quad & \sum_{t} \sum_{j} \sum_{q} \lambda_j^q \left[\sum_{\sigma} c_{\sigma j}^q\, u_{\sigma j}^{qt} \right] \\
\text{s.t.} \quad & \sum_{t} \sum_{j} \sum_{q \ni i} \lambda_j^q \left[\sum_{\{\sigma:\, \sigma(i)=k\}} u_{\sigma j}^{qt} \right] \leq S_{ki} \quad \forall k, i & (7) \\
& \sum_{\sigma} u_{\sigma j}^{qt} = 1 \quad \forall q, t, j & (8) \\
& u_{\sigma j}^{qt} \geq 0 \quad \forall q, t, j, \sigma & (9)
\end{aligned}
$$

The linear program J_{LP} has a natural interpretation. If demands are deterministic and arrive with rate $\{\lambda_j^q\}$, then the variable $u_{\sigma j}^{qt}$ can be interpreted as the probability of fulfilling demand type q from region j during period t according to assignment σ. Let $U_{\sigma j}^q$ denote the number of times order type q from region j are fulfilled using assignment σ during the selling horizon. We can formulate the time-aggregate version of J_{LP} as:

$$
\begin{aligned}
\tilde{J}_{LP} \;:=\; \min \quad & \sum_{j} \sum_{q} \sum_{\sigma} c_{\sigma j}^q\, U_{\sigma j}^q \\
\text{s.t.} \quad & \sum_{j} \sum_{q \ni i} \sum_{\{\sigma:\, \sigma(i)=k\}} U_{\sigma j}^q \leq S_{ki} \quad \forall k, i & (10) \\
& \sum_{\sigma} U_{\sigma j}^q = T \lambda_j^q \quad \forall q, j & (11)
\end{aligned}
$$

$$U^q_{\sigma j} \geq 0 \quad \forall j, q, \sigma \tag{12}$$

Let $\{\mathbf{u}^{qt}_{\sigma j}\}$ and $\{\mathbf{U}^q_{\sigma j}\}$ denote an optimal solution of J_{LP} and $\tilde{J}_{LP}$, respectively. (These solutions may *not* be unique.) It is not difficult to see that $J_{LP} = \tilde{J}_{LP}$. First, since $U^q_{\sigma j} = \sum_t \lambda^q_j \mathbf{u}^{qt}_{\sigma j}$ is a feasible solution to $\tilde{J}_{LP}$, we have $\tilde{J}_{LP} \leq J_{LP}$. To show the converse, simply note that $u^{qt}_{\sigma j} = \mathbf{U}^q_{\sigma j}/(T\lambda^q_j)$ is a feasible solution to J_{LP}. Indeed, it is also optimal. This proves our claim. (In the scaled problem, we have $\tilde{J}_{LP}(\theta) = \theta \tilde{J}_{LP}$ and $\mathbf{U}^q_{\sigma j}(\theta) = \theta \mathbf{U}^q_{\sigma j}$. So, this claim still holds.) All that we have done so far is showing that the large J_{LP} can be written in its most compact form as $\tilde{J}_{LP}$. We are now ready to introduce our first heuristic and derive its performance guarantee. The heuristic is very straightforward. At the beginning of the selling horizon, we first solve the linear program $\tilde{J}_{LP}$ and then use its (possibly fractional) optimal solution as probabilities to assign items to facilities. Formally, the heuristic is stated below, followed by a theorem stating its performance guarantee. The proof is provided in the appendix.

Probabilistic Fulfillment Control (PFC)

Input: $u^{qt}_{\sigma j} = \mathbf{U}^q_{\sigma j}/(T\lambda^q_j)$, where $\{\mathbf{U}^q_{\sigma j}\}$ is an optimal solution of $\tilde{J}_{LP}$

During period t, for an order type q from region j, do:

1. Sample σ with probability $u^{qt}_{\sigma j}$
2. For each $i \in q$, do:
   - If $S_{\sigma(i)i} \geq 1$, fulfill item i from facility $\sigma(i)$
   - Otherwise, fulfill item i from facility 1.

Theorem 1 *There exists a positive constant M independent of $\theta > 0$ such that for all $\theta > 0$ we have* $\mathbf{E}[C_{PFC}(\theta)] - J^*(\theta) \leq \mathbf{E}[C_{PFC}(\theta)] - J_{LP}(\theta) \leq M[1 + \sqrt{\theta}]$.

Theorem 1 tells us that PFC is *asymptotically optimal*. (Although the constant M is independent of $\theta > 0$, its magnitude is possibly exponential in the problem size.) In fact, the relative difference between the expected total cost under PFC and that under the optimal control is of order $\sqrt{\theta}/\theta = \theta^{-1/2}$, which is negligible for large θ. (To illustrate, if $\theta = 100$, then the expected total loss of PFC is about 10%.) The good news is that PFC only requires solving a linear program instead of a dynamic program. Moreover, this linear program only needs to be solved once at the beginning of the selling horizon. (Although re-solving can potentially reduce the bound in Theorem 1, we do not analyze it here. See Jasin and Kumar (2012) for a discussion on related literature.) The bad news is that the size of $\tilde{J}_{LP}$ can still be prohibitively large, making it infeasible for practical implementation. This is so because, for each pair (q, j), we have one decision variable for each fulfillment vector σ^q_j, whose number is exponential in the size of the input. So, we have a control which is asymptotically optimal but can be difficult to implement. This raises an important question whether it is possible to construct a heuristic which is

not exponential in the size of the input and, if yes, what performance guarantee can be obtained for such a heuristic.

4 An Approximate LP

In this section, we consider a relaxation of the exact LP which disentangles the dependency among items contained in the same order. This makes the LP smaller (polynomial in the size of the input). We then use its solution to construct a simple fulfillment heuristic, and bound its competitive ratio by the expected order size. More importantly, the structure of the LP and our heuristic enables us to develop a new heuristic in Section 5 with a tighter competitive ratio, which is our main result. Here, we will bound the term $\mathbf{E}[\max_{i\in q} X^{qt}_{kij}]$ in J^* with $\max_{i\in q} \mathbf{E}[X^{qt}_{kij}]$ and consider the following LP formulation:

$$
\begin{aligned}
J^M_{LP} \;=\; \min \quad & \sum_t \sum_j \sum_q \sum_k \lambda^q_j \left[\sum_{i\in q} c_{kij}\, u^{qt}_{kij} \;+\; b_{kj}\, y^{qt}_{kj} \right] \\
\text{s.t.} \quad & \sum_t \sum_j \sum_{q\ni i} \lambda^q_j\, u^{qt}_{kij} \le S_{ki} \quad \forall k, i && (13)\\
& \sum_k u^{qt}_{kij} = 1 \quad \forall q, t, j, i \in q && (14)\\
& y^{qt}_{kj} \ge u^{qt}_{kij} \quad \forall q, t, k, j, i \in q && (15)\\
& u^{qt}_{kij} \ge 0 \quad \forall q, t, k, i, j && (16)
\end{aligned}
$$

(The variable y^{qt}_{kj} essentially equals $\max_{i\in q} u^{qt}_{kij}$.) Here, we use a somewhat more traditional approach with a pair of variables (u, y) for each order, where $y^{qt}_{kj} = 1$ indicates that some, possibly all, items in order q are being fulfilled by facility k. This forces us to incur the fixed cost component of shipping from facility k to region j. The variable cost component depends on the specific items being shipped and is accounted for using the u variables in the first term in the objective function. Since we are considering a linear relaxation of the original problem, by Jensen's inequality, it follows that $J^* \ge J^M_{LP}$ (because $\mathbf{E}[\max_{i\in q} X^{qt}_{kij}] \ge \max_{i\in q} \mathbf{E}[X^{qt}_{kij}]$). So, J^M_{LP} provides another lower bound for J^* in addition to J_{LP}.

Let U^q_{kij} denote the number of times item i in order type q from region j are fulfilled from facility k during the selling horizon and let Y^q_{kj} denote the number of times order type q from region j are fulfilled from facility k at all during the selling horizon. The time-aggregate formulation of J^M_{LP} is

$$
\begin{aligned}
\tilde{J}^M_{LP} \;=\; \min \quad & \sum_j \sum_q \sum_k \left[\sum_{i\in q} c_{kij}\, U^q_{kij} \;+\; b_{kj}\, Y^q_{kj} \right] \\
\text{s.t.} \quad & \sum_j \sum_{q\ni i} U^q_{kij} \le S_{ki} \quad \forall k, i && (17)
\end{aligned}
$$

$$\sum_k U^q_{kij} = T\lambda^q_j \quad \forall q, j, i \in q \tag{18}$$

$$Y^q_{kj} \geq U^q_{kij} \quad \forall q, k, j, i \in q \tag{19}$$

$$U^q_{kij} \geq 0 \quad \forall q, k, i, j \tag{20}$$

Let $\{\mathbf{u}^{qt}_{kij}, \mathbf{y}^{qt}_{kj}\}$ and $\{\mathbf{U}^q_{kij}, \mathbf{Y}^q_{kj}\}$ denote an optimal solution of J^M_{LP} and $\tilde{J}^M_{LP}$, respectively. As in Section 3, it can be argued that $J^M_{LP} = \tilde{J}^M_{LP}$. In fact, an optimal solution of J^M_{LP} can be recovered via $u^{qt}_{kij} = \mathbf{U}^q_{kij}/(T\lambda^q_j)$ and $y^{qt}_{kj} = \mathbf{Y}^q_{kj}/(T\lambda^q_j)$. Let $|q|$ denote the number of items contained in order type q, i.e., $|q| = \sum_{i\in I} \mathbf{1}\{i \in q\}$. Observe that, for each pair (q, j), $\tilde{J}^M_{LP}$ only has $|\mathcal{S}_K|.|q|+|\mathcal{S}_K|$ variables whereas $\tilde{J}_{LP}$ has $|\mathcal{S}_K|^{|q|}$ variables. So, we have just reduced the size of the LP formulation from exponential to linear in the size of the input, which is good. But, is the new LP a good approximation of the original one? In particular, can we construct a heuristic using the solution of $\tilde{J}^M_{LP}$ which still maintains the asymptotic optimality of PFC? To answer this, we first propose a new fulfillment heuristic, which we call the *Modified PFC* (MPFC).

Modified Probabilistic Fulfillment Control (MPFC)

Input: $u^{qt}_{kij} = \mathbf{U}^q_{kij}/(T\lambda^q_j)$, where $\{\mathbf{U}^q_{kij}\}$ is an optimal solution of $\tilde{J}^M_{LP}$

During period t, for an order type q from region j, and for each $i \in q$, do:

- Fulfill item i from facility k with probability u^{qt}_{kij}
- If the sampled facility is out of stock, fulfill item i from facility 1.

For each triplet (q, k, j), define $F(q, k, j)$ as follows:

$$F(q,k,j) \;=\; \frac{b_{kj}\,\mathbf{Y}^q_{kj}}{\sum_{q',k',j'}\; b_{k'j'}\;\mathbf{Y}^{q'}_{k'j'}} \;=\; \frac{\lambda^q_j\, b_{kj}\,\mathbf{y}^{q1}_{kj}}{\sum_{q',k',j'}\; \lambda^{q'}_{j'}\; b_{k'j'}\;\mathbf{y}^{q'1}_{k'j'}}.$$

By definition, we always have $\sum_{q,k,j} F(q, k, j) = 1$. So, $F(.,.,.)$ can be interpreted as a probability distribution on the set $\{(q, k, j)\}$. Below, we state the performance of MPFC.

Theorem 2 *Let Q be a random variable denoting the order type. Then,*

$$\lim_{\theta\to\infty} \frac{\mathbf{E}[C_{MPFC}(\theta)]}{J^*(\theta)} \;\leq\; \lim_{\theta\to\infty} \frac{\mathbf{E}[C_{MPFC}(\theta)]}{J^M_{LP}(\theta)} \;\leq\; \sum_{q,k,j} |q|\, F(q,k,j) \;:=\; \mathbf{E}_F[|Q|].$$

Theorem 2 tells us that the performance of MPFC depends on the typical size of $|q|$ under F. If $|q|$ is typically small, then MPFC is near-optimal. If, on the other hand, $|q|$ is typically large, then MPFC may not provide a very satisfactory performance. Since the distribution F depends explicitly

on the optimal solution of $\tilde{J}^M_{LP}$ it may not be possible in general to know the performance of MPFC before solving the LP. It is, however, possible to get a more intuitive bound for the special case where the etailer is primarily interested in minimizing total shipments instead of total shipping costs. (Xu et al. 2009) argue that the former is sometimes a good proxy for the later. For all q, define P_q to be the probability that an arriving order is of type q, i.e. $P_q := P(Q = q) = (1 - \lambda_0)^{-1} \sum_j \lambda^q_j$. We state a lemma.

Lemma 1 *Suppose that we set all variable costs equal to 0. In addition, we also set $b_{kj} = M > 1$ for $k = 1$ and $b_{kj} = 1$ otherwise. If $\mathbf{U}^q_{1ij} = 0$ for all q, i, j, then*

$$\frac{P_q}{\mathbf{E}[|Q|]} \;\leq\; \sum_{k,j} F(q,k,j) \;\leq\; \min\{|\mathcal{S}_K|, |q|\}\, P_q.$$

The condition $\mathbf{U}^q_{1ij} = 0$ for all q, i, j simply says that all orders can be completely satisfied by the available inventories in non-virtual facilities. So, we do not have to incur stock-out costs, at least deterministically. Since the number of facilities $|\mathcal{S}_K|$ is fixed, Lemma 1 tells us that the sum $\sum_{k,j} F(q,k,j)$ is roughly proportional to P_q for all q, especially so for large $|q|$. Put together Lemma 1 with Theorem 2, for the setting described in Lemma 1, we immediately have

$$\lim_{\theta\to\infty} \frac{\mathbf{E}[C_{MPFC}(\theta)]}{J^*(\theta)} \;\leq\; \min\left\{|\mathcal{S}_K|\,\mathbf{E}[|Q|],\, \mathbf{E}[|Q|^2]\right\}.$$

Admittedly, this is a rather weak bound. However, it is still useful to give a sense on the potential performance of MPFC before solving $\tilde{J}^M_{LP}$. In particular, since $\mathbf{E}[|Q|^2] = \mathbf{E}[|Q|]^2 + \mathrm{var}(|Q|)$, one expects that MPFC should perform reasonably well if both $\mathbf{E}[|Q|]$ and $\mathrm{var}(|Q|)$ are small. In simpler language, this means that if (1) the average order size is small and (2) most customers only purchase a few items at a time, then MPFC is a good candidate for practical implementation. But, is this the case? Fortunately, the answer is yes. Xu et al. (2009) report that based on their analysis of data from a major online retailer, approximately 65% of orders during the non-peak season consist of single items. During the peak season, this drops somewhat to approximately 56%. Still, most multi-item orders (close to 100%) are fulfilled with two or three shipments. In addition to the work of Xu et al. (2009), a recent press release by eDataSource.com also reveals that shoppers at Amazon.com only purchase on average 1.5 items per order while shoppers at Walmart.com only purchase on average 2.3 items per order (accessed online at www.edatasource.com on Nov 28, 2013.). This provides more evidence for the fact that most customers only purchase very few items at a time. The numbers seem to suggest that MPFC may be appropriate after all. Indeed, our numerical results in Section 6 show a reasonably strong performance of MPFC. And yet, it is sometimes desirable to have a stronger performance guarantee than that provided by Theorem 2. For example, if all orders contain exactly 2 items, then simply having an asymptotically 2-competitive performance guarantee is hardly satisfactory. This gives rise to an important question whether we can improve the performance of MPFC by constructing a different

heuristic which still uses the solution of $\tilde{J}^M_{LP}$, albeit in a more sophisticated manner. In particular, we ask: Can we obtain a heuristic whose asymptotic competitive ratio is *strictly less* than $\mathbf{E}_F[|Q|]$? It turns out that this is possible. In fact, the new improved heuristic recovers the asymptotic optimality of PFC (i.e., it is 1-competitive) if all orders contain at most 2 items and is asymptotically 3-competitive if all orders contain at most 10 items (in constrast, MPFC is only 10-competitive). We discuss this next.

5 Improving the Bound

The key to the strong performance of PFC lies in the explicit inclusion of dependency factor (via the assignment term σ) in the exact LP formulation. MPFC, on the other hand, attempts to completely decouple this dependency by assuming that the fulfillment decision for each item can be made independently of the others. As the bound in Theorem 2 suggests, this may not yield a satisfactory performance. Obviously, dependency is an important factor and should not be ignored. And yet, implementation challenge arises precisely because an explicit inclusion of dependency, even if only partially (e.g., via partial decomposition instead of complete decomposition into independent items), would require us to introduce another assignment term, which can still be exponential in the size of the input. Motivated by this practical concern, instead of creating a new large LP, we will propose a new heuristic which still uses the solution of J^M_{LP} and, given this solution, automatically constructs an *artificial dependency* among the items contained in the same order. We first illustrate the idea using a simple example and then we discuss its extension to the general setting.

5.1 A Simple Example

Suppose that we only have one selling period (i.e., $T = 1$), two items, two facilities, one region, and one order type containing both items. All variable costs are equal to 0 and all fixed costs are equal to 1. The approximate LP formulation for our problem is given by:

$$\begin{aligned}
J^M_{LP} \;=\; \min \quad & \lambda \sum_k y_k \\
\text{s.t.} \quad & \lambda\, u_{ki} \le S_{ki} \quad i = 1,2; k = 1,2 && (21)\\
& \sum_k u_{ki} = 1 \quad i = 1,2 && (22)\\
& y_k \ge u_{ki} \quad i = 1,2; k = 1,2 && (23)\\
& u_{ki} \ge 0 \quad i = 1,2; k = 1,2 && (24)
\end{aligned}$$

where, for simplicity, we suppress notational dependency on q, t, and j. For illustration purpose, suppose that an optimal solution to the above LP is given by $\mathbf{u}_{11} = \frac{1}{4}$, $\mathbf{u}_{21} = \frac{3}{4}$, and $\mathbf{u}_{12} = \mathbf{u}_{22} = \frac{1}{2}$. Let $X_{ki} \sim \text{Bernoulli}(\mathbf{u}_{ki})$. If we ignore the capacity constraint for the moment, implementing MPFC

yields the per-request total expected cost equal to

$$\begin{aligned}\mathbf{E}\left[\max_i\{X_{1i}\}\right] \;+\; \mathbf{E}\left[\max_i\{X_{2i}\}\right] &= \mathbf{E}\left[1-\prod_i(1-X_{1i})\right] \;+\; \mathbf{E}\left[1-\prod_i(1-X_{2i})\right] \\ &= 1-\prod_i(1-\mathbf{u}_{1i})+1-\prod_i(1-\mathbf{u}_{2i}) \\ &= \frac{3}{2}.\end{aligned}$$

In contrast, $\mathbf{y}_1+\mathbf{y}_2 = \max_i\{\mathbf{u}_{1i}\}+\max_i\{\mathbf{u}_{2i}\} = \frac{1}{2}+\frac{3}{4} = \frac{5}{4}$. So, the competitive ratio of MPFC with respect to J^M_{LP} is $\frac{3/2}{5/4} = \frac{6}{5}$. We now show how to reduce this competitive ratio to 1 by constructing an artificial dependency among $\{X_{ki}\}$. For each item i, we first construct a line partition on a unit interval and designate each partition to a unique facility. The length of the union of partitions designated to facility k for each item i must equal to $\mathbf{u}_{ki}$ for all k and i. After all partitions have been constructed, we perform a uniform sampling on $[0,1)$. The outcome of this sampling completely determines the shipping decision for each item in the order. For example, consider the following partitions for our problem above: For item 1, we designate interval $[0,\frac{1}{4})$ to facility 1 and interval $[\frac{1}{4},1)$ to facility 2 whereas, for item 2, we designate interval $[0,\frac{1}{2})$ to facility 1 and interval $[\frac{1}{2},1)$ to facility 2. Suppose that we sample $\chi \sim \text{Uniform}[0,1]$ and get $\chi = \frac{1}{3}$. Since $\frac{1}{3} \in [\frac{1}{4},1)$ and $\frac{1}{3} \in [0,\frac{1}{2})$, we ship item 1 from facility 2 and item 2 from facility 1. Using these partitions, we can re-calculate the total expected cost to be:

$$\begin{aligned}\mathbf{E}\left[\max_i\{X_{1i}\}\right] \;+\; \mathbf{E}\left[\max_i\{X_{2i}\}\right] &= P\left(\chi \in \left[0,\frac{1}{4}\right) \cup \left[0,\frac{1}{2}\right)\right) \;+\; P\left(\chi \in \left[\frac{1}{4},1\right) \cup \left[\frac{1}{2},1\right)\right) \\ &= \frac{1}{2} \;+\; \frac{3}{4} \;=\; \frac{5}{4}.\end{aligned}$$

Since $\mathbf{y}_1+\mathbf{y}_2 = \frac{5}{4}$, surprisingly, we have just reduced the competitive ratio of MPFC from $\frac{6}{5}$ to 1! In fact it can be shown that for an optimal LP solution for this example with all $S_{ki} = 1$, the competitive ratio of MPFC is actually $3/2$, which also gets reduced to 1 with the injection of artificial dependencies. The example highights the power of injecting artificial dependency into fulfillment decisions. It is important to note here that there are many partitions that can be used to still guarantee the 1-competitive ratio in the above example. For example, we can use the following: For item 1, we still designate $[0,\frac{1}{4})$ to facility 1 and $[\frac{1}{4},1)$ to facility 2 whereas, for item 2, we now designate $[0,\frac{1}{4}) \cup [\frac{3}{4},1)$ to facility 1 and $[\frac{1}{4},\frac{3}{4})$ to facility 2. It is not difficult to check that $\mathbf{E}\left[\max_i\{X_{1i}\}\right] + \mathbf{E}\left[\max_i\{X_{2i}\}\right]$ still equals to $\frac{5}{4}$. Thus, the new partitions still yield the same 1-competitive ratio as the old one. It is also important to note that in either partition, the old or the new, we always have $\mathbf{E}[\max_i\{X_{ki}\}] = \max_i \mathbf{u}_{ki}$. This observation will play an important role in our analysis later. In particular, in the general setting, we want to construct a dependency such that the expectation $\mathbf{E}[\max_{i\in q}\{X^{q1}_{kij}\}]$ is as close as possible to $\max_{i\in q}\mathbf{u}^{q1}_{kij}$ for all q, k, and j.

5.2 General Setting

Let $\mathbf{u}^{qt}_{kij} = \mathbf{U}^q_{kij}/(T\lambda^q_j)$ and $\mathbf{y}^{qt}_{kj} = \mathbf{Y}^q_{kj}/(T\lambda^q_j)$, where $\{\mathbf{U}^q_{kij}, \mathbf{Y}^q_{kj}\}$ are the optimal solution of $\tilde{J}^M_{LP}$. The combined task of constructing line partitions and using a uniform random number to generate fulfillment decisions is essentially equivalent to constructing a joint probability distribution g^q_j on the assignment vector $\{\sigma\}$ satisfying $\mathbf{E}[X^{q1}_{kij}] = \mathbf{u}^{q1}_{kij}$, or equivalently $\sum_{\{\sigma:\sigma(i)=k\}} g^q_j(\sigma) = \mathbf{u}^{q1}_{kij}$ for all $i \in q$, and to fulfill order type q coming from region j according to assignment σ with probability $g^q_j(\sigma)$. Ideally, the distribution g^q_j must be constructed in a way that minimizes total expected shipping costs (i.e., the objective function of J^*). Although an optimal g^q_j can be exactly computed using an optimization approach (see Section 7), the resulting optimization can be large if $|q|$ is large and it may not be practically convenient to solve g^q_j for each pair (q, j) if the number of possible order types is large. Thus, in order to maintain simplicity, we resort to an optimization-free approach in constructing an approximate g^q_j. We call the resulting heuristic *Improved PFC* (IPFC). We will discuss how to construct g^q_j shortly. For now, we first discuss its performance. Define $B(n)$ as follows: $B(n) = \frac{n+2}{4}$ if n is even and $B(n) = \frac{(n+1)^2}{4n}$ if n is odd. We state our result below.

Theorem 3 *There exist joint distributions $\{g^q_j\}$ such that*

$$\lim_{\theta\to\infty} \frac{\mathbf{E}[C_{IPFC}(\theta)]}{J^*(\theta)} \;\le\; \sum_{q,k,j} B(|q|)F(q,k,j) \;=\; \mathbf{E}_F[B(|Q|)],$$

where F is as defined in Theorem 2.

Table 2 provides the values of $B(n)$ for selected values of n. Two comments are in order. First, if $|q| \le 2$ for all q, which means that all orders contain at most two items, then IPFC is asymptotically optimal. So, IPFC recovers the asymptotic optimality of PFC, in some cases. Second, since $\lim_{n\to\infty} \frac{B(n)}{n} = \frac{1}{4}$, the expected total cost under IPFC can be up to *four* times smaller than that of MPFC. This tells us that IPFC yields a significant improvement over MPFC.

n	1	2	3	4	5	6	10	20	50	100
$B(n)$	1	1	1.33	1.5	1.8	2	3	5.5	13	25.5

Table 2: The values of $B(n)$.

Constructing Line Partitions. We now show how to construct the joint distribution g^q_j which satisfies $\mathbf{E}\left[\max_{i\in q}\{X^{q1}_{kij}\}\right] \le B(|q|)\left(\max_{i\in q}\mathbf{u}^{q1}_{kij}\right)$ for all pairs (q, j), from which the result of Theorem 3 immediately follows. For the rest of discussion, we will fix (q, j) and suppress notational dependencies on q and j for easier readability. We first describe the big picture. For each item $i \in q$, we partition the unit interval $\mathcal{I}_i = [0, 1]$ into small segments $\{\mathcal{I}^v_i\}$ with $\sum_v |\mathcal{I}^v_i| = 1$. We define a mapping $h_i : \{v\} \mapsto \mathcal{S}_K$, so that each segment is mapped/designated to a certain facility. Although two different segments may be mapped to the same facility, we will ensure that the mapping preserves the J^M_{LP} solution $\mathbf{u}$. That

is, $\sum_{\{v:h_i(v)=k\}} |\mathcal{I}_i^v| = \mathbf{u}_{ki}$. For each $\chi \in [0,1]$, let $v_i(\chi)$ denote the index of segment $\mathcal{I}_i^v$ that contains χ. So, by definition, $\chi \in \mathcal{I}_i^v$ if and only if $v = v_i(\chi)$. As illustrated in Section 5.1, our idea of creating dependency among items $\{i \in q\}$ is via uniform sampling on a unit interval. Suppose that we sample $\chi \in [0,1]$ and ship item i from facility $h_i(v_i(\chi))$ if it is available, and from facility 1 otherwise. (Since $\sum_{\{v:h_i(v)=k\}} |\mathcal{I}_i^v| = \mathbf{u}_{ki}$, if there is no stock-out, any such scenario will always yield $\mathbf{E}[X_{ki}] = \mathbf{u}_{ki}$. So, the marginal distribution of x_{ki} is preserved.) If we can construct $\{\mathcal{I}_i^v\}$ and $\{h_i(.)\}$ in a way that reduces the number of distinct facilities used by *lining up* segments of the interval that assign the item to the same facility, then we can avoid having to pay the fixed cost of shipment $|q|$ times, and obtain a better competitive ratio. In what follows, we first describe how to construct the partitions $\{\mathcal{I}_i^v\}$ for all $i \in q$. Subsequently, we will discuss an example and prove that $\mathbf{E}\left[\max_{i\in q}\{X_{ki}^q\}\right] \leq B(|q|)\left(\max_{i\in q}\mathbf{u}_{ki}^q\right)$.

The construction of the line partition $\{\mathcal{I}_i^v\}$ proceeds in several steps.

STEP 1. Let $|q| = n$. For each k, write $\{\mathbf{u}_{ki}\} = \mathbf{u}_k$ as a column vector where its i^{th} element is given by $\mathbf{u}_{ki}$. Our first step is to decompose $\mathbf{u}_k$ as the sum $\sum_{m=1}^n \tilde{u}_k^m$ where $\tilde{u}_k^m$ either has exactly m non-zero elements and they are all the same, or is a zero vector. Formally, we generate the $\tilde{u}$ vectors using the following algorithm:

DECOMPOSE. Input: $\mathbf{u}_k$. Initialize: $v = \mathbf{u}_k$. For $m = n:1$ (counting backwards from n to 1), do: Let r denote the number of non-zero elements of v. If $r < m$, set $\tilde{u}_{ki}^m = 0$ for all $i \in q$. If, on the other hand, $r = m$, set $\tilde{u}_{ki}^m = 0$ if $v_i = 0$ and $\tilde{u}_{ki}^m =$ the smallest non-zero elements of v otherwise for all $i \in q$. Recompute $v = v - \tilde{u}_k^m$. Set $m = m - 1$ and redo all the steps.

STEP 2. Let $M_k^m = \max_{i\in q} \tilde{u}_{ki}^m$. Since $\sum_k \mathbf{u}_{ki} = 1$ for all $i \in q$, we have

$$n \;=\; \sum_{i\in q}\sum_{k} \mathbf{u}_{ki} \;=\; \sum_{k}\sum_{i\in q} \mathbf{u}_{ki} \;=\; \sum_{k}\sum_{m}\sum_{i\in q} \tilde{u}_{ki}^m \;=\; \sum_{k}\sum_{m} m\, M_k^m.$$

(By construction, either $\tilde{u}_{ki}^m = 0$ or M_k^m. So, either $\tilde{u}_k^m$ is a zero vector or it has exactly m non-zero components.) This implies

$$\sum_k M_k^n \;+\; \frac{n-1}{n}\sum_k M_k^{n-1} \;+\; ... \;+\; \frac{2}{n}\sum_k M_k^2 \;+\; \frac{1}{n}\sum_k M_k^1 \;=\; 1.$$

Define $L_0 = 0$ and $L_k \;=\; L_{k-1} + M_k^n \;+\; \frac{n-1}{n} M_k^{n-1} \;+\; ... \;+\; \frac{2}{n} M_k^2 \;+\; \frac{1}{n} M_k^1$ for $k = 1, ..., K$. (Note that $L_K = 1$.) Also, define the sequence $\{H_{km}\}$ as follows: $H_{k0} = L_{k-1}$ and $H_{km} = H_{k,m-1} + \frac{m}{n} M_k^m$ for $m = 1, 2, ..., n$. (By construction, we have $H_{kn} = L_k$.) Let $\tilde{\mathcal{I}}_{km} = [H_{k,m-1}, H_{k,m})$. The intervals $\{\tilde{\mathcal{I}}_{km}\}$ form a partition for the unit interval $[0,1)$ (see Figure 1).

STEP 3. We are now ready to construct our line partition. For each $i \in q$, the partition $\{\mathcal{I}_i^v\}$ is constructed as follows. Let $\mathcal{I}_i^{km} = \tilde{\mathcal{I}}_{km}$. If $\tilde{u}_{ki}^m > 0$, map interval $\mathcal{I}_i^{km}$ to facility k. That is, set $h_i(k,m) = k$. If $\tilde{u}_{ki}^m = 0$, mark interval $\mathcal{I}_i^{km}$ as "unassigned." Let $\mathcal{I}_i^A$ be the union of all unassigned

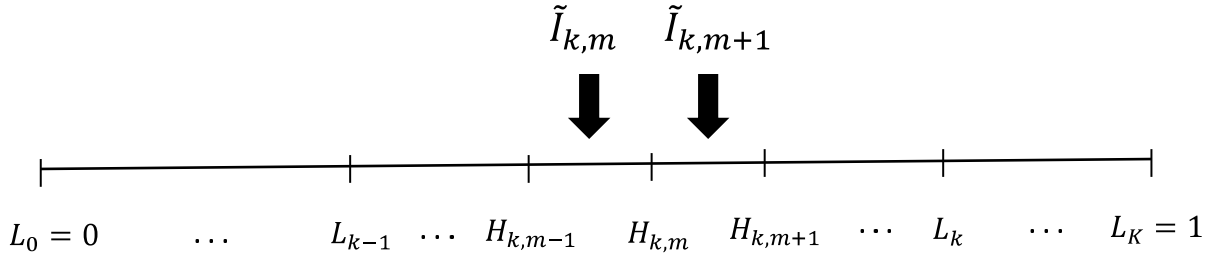

Figure 1: Constructed Line Partition

intervals. Arbitrarily partition $\mathcal{I}_i^A$ into small sub-intervals $\{\mathcal{I}_i^{Akm}\}$ such that $\bigcup_{k,m} \mathcal{I}_i^{Akm} = \mathcal{I}_i^A$ and $|\mathcal{I}_i^{km}| + |\mathcal{I}_i^{Akm}| = \tilde{u}_{ki}^m$ for all k and m. (This is always possible.) Finally, map $\mathcal{I}_i^{Akm}$ to facility k. This completes the construction of $\{\mathcal{I}_i^v\}$.

An Example. To illustrate the above construction, we will now consider a simple 4-item and 3-facility example. Suppose that the optimal solution $\mathbf{u}_{ki}$ is given by the following matrix, where each row represents an item and each column represents a facility:

$$\mathbf{u} = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.0 & 1.0 & 0.0 \\ 0.4 & 0.5 & 0.1 \\ 0.0 & 0.3 & 0.7 \end{bmatrix}.$$

For our first step, the decomposition for $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u_3}$ yield

$$\mathbf{u_1} = \tilde{u}_1^4 + \tilde{u}_1^3 + \tilde{u}_1^2 + \tilde{u}_1^1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0.4 \\ 0 \\ 0.4 \\ 0 \end{bmatrix} + \begin{bmatrix} 0.2 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

$$\mathbf{u_2} = \tilde{u}_2^4 + \tilde{u}_2^3 + \tilde{u}_2^2 + \tilde{u}_2^1 = \begin{bmatrix} 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0.2 \\ 0.2 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0.5 \\ 0 \\ 0 \end{bmatrix},$$

$$\mathbf{u_3} = \tilde{u}_3^4 + \tilde{u}_3^3 + \tilde{u}_3^2 + \tilde{u}_3^1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0 \\ 0.1 \\ 0.1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0.6 \end{bmatrix}.$$

The resulting $\{M_k^m\}$, $\{L_k\}$, $\{H_{km}\}$, and $\{\tilde{\mathcal{I}}_{km}\}$ in the second step are given below (the columns are indexed by k and the rows by m):

$$M = \begin{bmatrix} 0.2 & 0.5 & 0.6 \\ 0.4 & 0.2 & 0 \\ 0 & 0 & 0.1 \\ 0 & 0.3 & 0 \end{bmatrix}, \quad L = [0.25 \quad 0.775 \quad 1], \quad H = \begin{bmatrix} 0.05 & 0.375 & 0.925 \\ 0.25 & 0.475 & 0.925 \\ 0.25 & 0.475 & 1 \\ 0.25 & 0.775 & 1 \end{bmatrix},$$

$$\text{and} \quad \tilde{\mathcal{I}} = \begin{bmatrix} [0,\, 0.05) & [0.25,\, 0.375) & [0.775,\, 0.925) \\ [0.05,\, 0.25) & [0.375,\, 0.475) & \emptyset \\ \emptyset & \emptyset & [0.925,\, 1) \\ \emptyset & [0.475,\, 0.775) & \emptyset \end{bmatrix}.$$

The final partitions for our example are shown in Figure 2. It is easy to verify that, for each item i, the total length of partitions designated to each facility k equals $\mathbf{u}_{ki}$. The point to note here is that the partitions have been constructed in a way to *increase overlaps*, so that if a particular facility is chosen for an item, it is more likely than in the case of independent probabilistic fulfillment that the same facility is chosen for other items as well. This *dependent probalistic fulfillment* thus reduces the total number of shipments. For instance, the partition [0.475, 0.775] is designated to facility 2 for every item. Similarly, although only two items are fractionally served by facility 1 (these are items 1 and 3), the partition designated for facility 1 in $\mathcal{I}_3$ is a subset of that for facility 1 in $\mathcal{I}_1$. The effect of this manifests itself when we use the partitions to assign the items to facilities. Recall the IPFC assignment rule: We sample χ uniformly from the interval $[0, 1]$, and assign each item i to the facility mapped to by the number χ in $\mathcal{I}_i$. So, in our example, if we obtain $\chi = 0.95$, then item 2 is assigned to facility 2 and all other items are assigned to facility 3. By averaging the number of facilities used as χ goes from 0 to 1, it is easy to verify that for this example, the expected number of facilities used is 2.3. In contrast, if we just used the $\mathbf{u}$ solution and assigned items to facilities *independently*, the expected number of facilities is 2.517.

The Proof. We now provide the proof of Theorem 3. The key is to show that $\mathbf{E}\left[\max_{i\in q}\{X_{kij}^{qt}\}\right] \leq B(|q|)\left(\max_{i\in q}\mathbf{u}_{kij}^{qt}\right)$ for all q, k, j. Fix (q, j) and let $|q| = n$ (whenever possible, we will suppress notational dependency on (t, q, j)). By our construction, for each k, we have

$$\mathbf{E}\left[\max_{i\in q}\{X_{ki}\}\right] \;\leq\; \sum_{m=1}^{n}\left[\frac{m}{n}M_k^m \;+\; m\left(1-\frac{m}{n}\right)M_k^m\right],$$

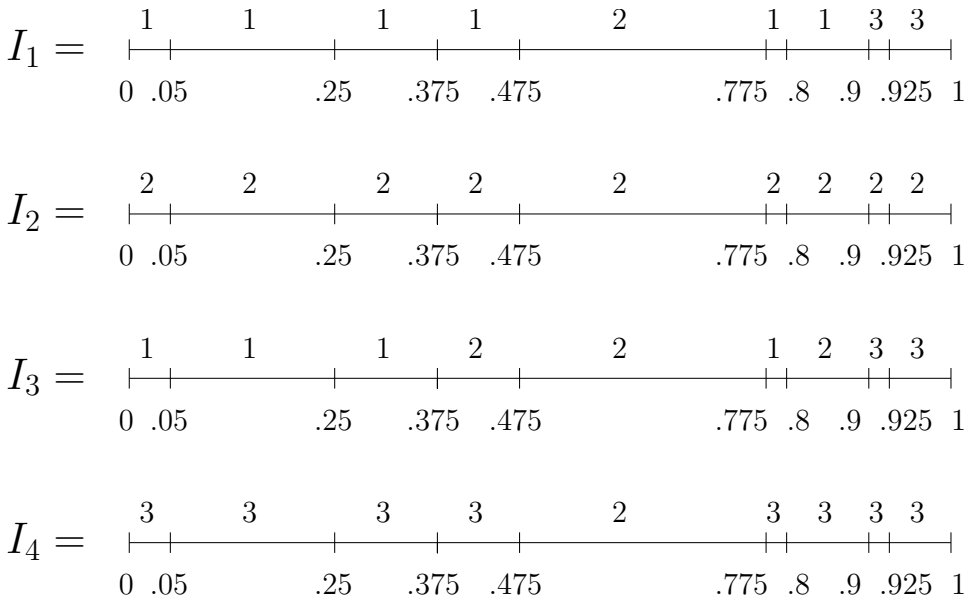

Figure 2: Intervals for 4-item 3-facility example.

where the expectation is taken with respect to the induced joint distribution. The first term in the summation follows because $\mathcal{I}_i^{km} = \mathcal{I}_{i'}^{km}$ for all i, $i' \in q$ with $\tilde{u}_{ki}^m = \tilde{u}_{ki'}^m = M_k^m > 0$, and $|\mathcal{I}_i^{km}| = \frac{m}{n} M_k^m$. The last term follows because, under the worst case scenario, the intervals $\mathcal{I}_i^{Akm}$ and $\mathcal{I}_{i'}^{Akm}$ may not intersect at all. Since there can be at most m such intervals (by definition, $\tilde{u}_k^m$ contains at most m non-zero elements), we have a multiplicative factor m. We divide our analysis into two cases. If n is even, then $\max_{1 \leq m \leq n} \left[\frac{m}{n} + m\left(1 - \frac{m}{n}\right)\right] = \frac{n+2}{4}$. If n is odd, $\max_{1 \leq m \leq n} \left[\frac{m}{n} + m\left(1 - \frac{m}{n}\right)\right] = \frac{(n+1)^2}{4n}$. So, by definition of $B(.)$, $\mathbf{E}\left[\max_{i \in q} \{X_{ki}\}\right] \leq B(n) \sum_m M_k^m = B(|q|) \sum_m M_k^m$. But, $\max_{i \in q} \mathbf{u}_{ki} = \max_{i \in q} \sum_m \tilde{u}_{ki}^m = \sum_m \max_{i \in q} \tilde{u}_{ki}^m = \sum_m M_k^m$, where the second equality follows because, by construction, $M_k^m > 0$ and $\tilde{u}_{ki}^m = 0$ imply $\tilde{u}_{ki}^{m'} = 0$ for all $m' < m$. We conclude that $\mathbf{E}\left[\max_{i \in q} \{X_{ki}\}\right] \leq B(|q|)\ (\max_{i \in q}\ \mathbf{u}_{ki})$. The theorem now follows by the same argument as in the proof of Theorem 2.

6 Numerical Experiments

We now demonstrate the efficacy of our algorithms via numerical simulations. The numerical simulations were constructed so as to model a real business environment as closely as possible (within the abstraction considered in this paper). Broadly speaking, our results demonstrate that the IPFC algorithm performs exceedingly well, obtaining an observed competitive ratio that is very close to 1 under a wide variety of settings. It always performs better than a simple myopic strategy, often much better. Additionally, even within our limited computing resources (relative to the resources that may be at the disposal of a large corporation), we are able to solve problems of a fairly large scale in very small amounts of time. These details are explained below. Complete details of the simulation are provided in the appendix, so as to not distract too much from the flow of the paper.

6.1 Setup of numerical experiments

We model a firm located in the continental United States (that is, the US minus Alaska and Hawaii), with customer demand arriving from any of the 99 largest cities (we actually took the 100 largest cities, but excluded Honolulu)(U.S. Census Bureau 2014). Possible locations of the fulfillment centers are those that have been determined to be optimal by another study**?**. We assume each item is exactly 1 pound in weight. We model shipping costs by obtaining UPS rates for 891 combinations of facilities, customer locations, and package weights, and using linear regression to estimate the shape of the cost function. The resulting shipping cost function, for shipping an order q from facility k to customer j is the following, where d_{kj} denotes the distance in miles from k to j:

$$cost(q,k,j) = 8.759 + 0.423|q| + 0.000541|q|d_{kj}$$

This estimation has an R^2 of 94.5%, and all three coefficients are significant with p-values of the order of 10^{-15}. The coefficient for d_{kj} in the absence of $|q|$ was insignificant and hence dropped. To translate this in to the terminology of the rest of the paper, we find that $b_{kj} = 8.759$ and $c_{kij} = 0.423 + 0.000541 d_{kj}$. Note that the distances d_{kj} are often in the hundreds or even thousands of miles, so it does have a non-negligible effect on shipping costs. We use thse shipping costs for this numerical study, except for a few experiments where we simply minimize the number of packages ($b_{kj} = 1$ and $c_{kij} = 0$).

The number of items $|I|$ we model ranges from 10 to 500. Given a set I of items, we construct the set Q randomly; the precise process is described in the appendix. Typically, an order $q \in Q$ will have between 1 and 10 items, while the size of the set Q ranges from 10 to 50. We then generate demand rates λ_j^q, such that λ_j is proportional to the actual metropolitan population of city j (so, for example, demand from New York will be approximately 10 times as high as demand from Las Vegas, because their metropolitan area populations are roughly 20 million and 2 million respectively). We compute $E[B(|q|)]$ based on this construction and compare our results against it.

As described in Section 2, there is an additional fictitious facility, labeled facility 1, with infinite inventory of all products, to model what the firm may do in case of stockout. The shipping cost parameters for this facility are $b_{1j} = 2 \times 8.759$ and $c_{1ij} = 2 \times (0.423 + 0.000541 \max_{k,j} d_{kj})$. In general, not all facilities stock all products. As described further in the appendix, we define a parameter p_{stock}, and each facility $k \neq 1$ stocks item i with probability p_{stock}, independently for all k, i. For each facility k that stocks item i, we then set its initial inventory to equal the expected demand from all cities whose nearest facility that stocks item i is facility k. We test for sensitivity with respect to both the p_{stock} parameter and the initial inventory level later in this section.

Our numerical experiments test the performance of both our algorithms $IPFC$ and $MPFC$. In addition, we test a Myopic algorithm, which simply sends each item in an order from the nearest facility that stocks it. We attempted to implement a perfect hindsight algorithm, but that requires solving an integer program, which took too long even on problem sizes approximately one-thousandth the size of the typical simulations we report below.

Our experiments were implemented on our institution's scientific computing platform, using 2.27 GHz Intel Xeon E7-4860 processors with 200 GB of RAM. The typical run time of a single simulation trial (generating demands, solving LP, and simulating IPFC, MPFC, and Myopic) takes less than one minute, with even the largest instances taking no more than 5 minutes. This allowed us to run over 10,000 simulation trials over various ranges of the input parameters, some of which are reported below.

6.2 Base case simulation

We begin with a detailed look at a single set of simulation trials with no parameters varying. We consider the following case: $|I| = 20$ items, $|J| = 10$ customer locations, $|K| = 5$ facilities, and $|Q| = 26$ order types. Using $T = 100$ and $\theta = 100$, the total time horizon in our simulation is given by $\theta T = 10,000$. Given our formulation, it does not matter what the precise values of θ and T are; only the product θT matters for the purpose of numerical simulation. Therefore, for the simulation results below, we generally report θT. Complete details of our numerical experiment set-up are provided in Appendix B.

We report the results of two sets of experiments below. In the first, we **fix the demand rates** λ, so the only source of variation is the actual sequence of demand arrivals. This allows us to get a sense of the variation caused by only the stochasticity in demand arrivals. In the second set, **we also vary the demand rates** λ, while keeping all problem parameters ($|I|, |J|, |K|, |Q|$ as well as inventory control parameters described in Appendix B) fixed. For each set, we report results over 30 simulation trials.

	Fixed λ, varying demand					Varying λ and demand				
	Myopic	MPFC	IPFC	$E[B(\|q\|)]$	Impr.	Myopic	MPFC	IPFC	$E[B(\|q\|)]$	Impr.
Mean	1.056	1.083	1.028	1.230	0.028	1.082	1.156	1.042	1.313	1.040
Stdev	0.016	0.016	0.014	0	0.005	0.027	0.078	0.018	0.144	0.022
LCL	1.050	1.077	1.023	1.230	0.026	1.072	1.127	1.035	1.259	0.031
UCL	1.062	1.088	1.033	1.230	0.030	1.092	1.185	1.049	1.367	0.048

Table 3: Base case simulation data. First four colums in each sub-table are competitive ratios with respect to LP lower bound. Impr. (Improvement) is defined as competitive ratio of Myopic minus competitive ratio of IPFC. LCL and UCL are lower and upper confidence limits respectively, at the 95% level.

There are several points worth noting. First, note that the competitive ratio of IPFC is less than 5%: and this is against the LP lower bound, not against the optimal solution. This suggests that IPFC's performance is indeed very good in our numerical results. Next, observe that IPFC dominates Myopic in both situations. In fact, we find strong statistical evidence (based on the confidence intervals) that IPFC is able to recover approximately half the optimality gap of Myopic. Also note that $E[B(|q|)]$ is a very loose upper bound, so even though our theoretical guarantee of $E[B(|q|)]$ may not appear very encouraging, IPFC's actual performance may be much better.

MPFC is dominated by IPFC here (to be expected) and also by Myopic (perhaps unexpected). The relationship between MPFC and Myopic is not universal: when the objective function is to minimize the

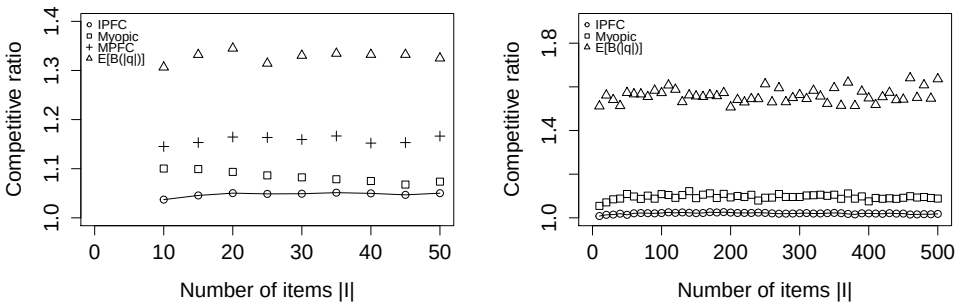

Figure 3: Performance with respect to number of items $|I|$. Left panel: minimize shipping cost in dollars. Right panel: minimize number of packages.

number of packages instead of the shipping cost, in most cases MPFC dominates Myopic. Nevertheless, given that MPFC is always worse than IPFC, for most of the rest of this section we focus on the comparison between IPFC and Myopic.

Also notice the difference between the tables on the left and the right. As is to be expected, when λ varies in addition to the demand varying, the overall competitive ratios are worse. However, IPFC still recovers approximately half the optimality gap of Myopic.

Remark on scaling simulation size. The numerical study in this section includes tests where the number of items $|I|$ grows to as much as 500. For the other quantities, the maximum numbers in our simulations are 95 customer regions, 9 facilities, and 100 order sizes (although we did not do a simulation where all these parameters were at their maximum values). In our computing environment, the main constraint appears to be that if $|I|.|J|.|K|.|Q|$ exceeds around 250,000, we run out of memory. This is not an insurmountable constraint: simply using file storage for memory would allow us to increase further the size of problems we can solve.

A real firm is likely to have access to significantly greater computing resources, particularly with the growth of on-demand elastic cloud computing technology. Additionally, our simulations generally ran within 1 minute: a real firm is likely to be willing to spend several hours or more if the algorithms generate an inventory policy that is expected to run for several days or weeks. With just those two extensions, we conjecture that scaling up the algorithms to thousands of items and packages will be fairly straightforward.

Additionally, the main algorithmic bottleneck in our work is the linear program. We use a naive LP solver in Matlab. It is possible that with techniques such as decomposition and exploitation of the sparsity of the linear programs, the size of the LP that can be solved can be much larger. Therefore, we believe that our numerical results provide strong support for the scalability of our algorithms to real-world sizes.

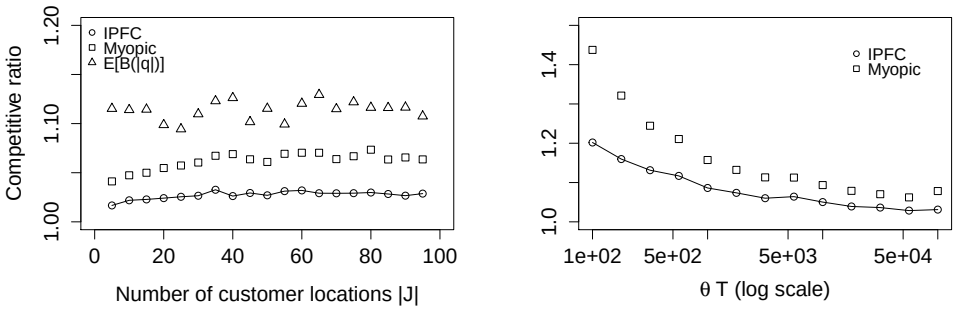

Figure 4: Performance with respect to number of customer locations $|J|$ (left panel) and length of time horizon θT (right panel).

6.3 Scaling with number of items

The left panel of Figure 3 shows the four output metrics (competitive ratios of IPFC, MPFC, and Myopic, and $E[B(|q|)]$) as the number of items $|I|$ increases. First, observe that *IPFC is always better than Myopic, and much lower than* $E[B(|q|)]$. For small number of items, the competitive ratios of IPFC and Myopic are 1.038 and 1.101 respectively; with 50 items these ratios are 1.050 and 1.074. This suggests that Myopic can incur 6% to 10% extra costs compared to the LP lower bound, while IPFC is able to recover 30% to 70% of this extra cost. This is encouraging; although some of our other results will show an even stronger performance of IPFC.

It is also worth noting that the upper bound provided by $E[B(|q|)]$ is between 1.30 and 1.35, substantially worse than the observed performance of IPFC. In fact, for all of our numerical simulations, the actual performance of IPFC was much better than the upper bound $E[B(|q|)]$. Given this, in the following sections we won't even report $E[B(|q|)]$, to allow for a clearer comparison of IPFC and Myopic. We also note in passing that as expected, MPFC performs worse than IPFC. This is also generally true in all our experiments, and we will ignore MPFC also from here on.

The right panel of Figure 3 shows the competitive ratios of IPFC and Myopic, and the upper bound $E[B(|q|)]$ for the case of minimizing the number of shipments. In these experiments, the setting is exactly as defined above except that instead of minimizing total shipping cost we simply minimize number of shipments. Given that shipping distances do not matter, we can simply aggregate all customers to a single location (so $|J| = 1$), and this allows us to scale the number of items $|I|$ to 500.

In this case, we find that the performance of IPFC is even stronger. The competitive ratio of IPFC with respect to the LP lower bound is approximately 1.02, while for Myopic it is approximately 1.10. That is, a Myopic algorithm will incur approximately 5 times excess cost compared to IPFC!

6.4 Sensitivity to number of locations and time horizon

The left panel of Figure 4 shows the competitive ratios as the number of possible customer locations changes. We observe that the competitive ratio stays somewhat stationary for both Myopic and IPFC,

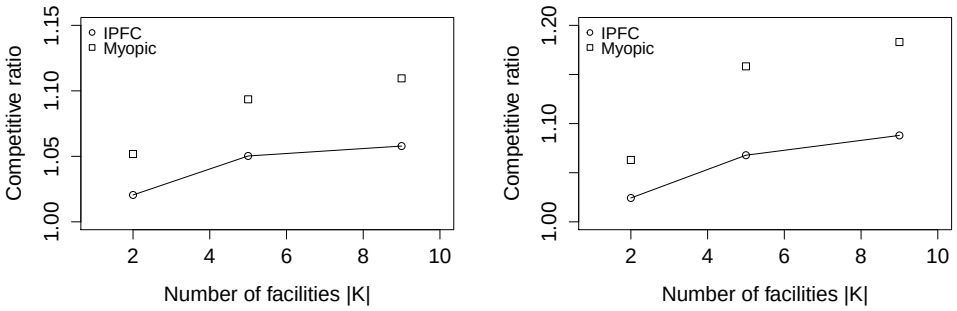

Figure 5: Performance with respect to number of fulfillment centers $|K|$, with shipping costs (left panel) and number of packages (right panel) as objective.

particularly after 35 cities or so. This is not entirely surprising given the structure of the shipping cost. Furthermore, if one were to use actual UPS rates, we would observe this stationarity with an even smaller number of customer locations because UPS charges according to a zone system: from a given origin, the rest of the country is divided into at most 9 zones, and all destinations within a single zone incur the same shipping cost.

The right panel shows the competitive ratios as the time horizon θT increases, from 100 to 100,000. The main observation is that although the competitive ratios decrease sharply with θT initially, they stabilize once θT is 10,000 or more. This is to be expected. At the lower extreme of $\theta T = 100$, with $|I| = 20$, $|J| = 10$ and $|K| = 5$, the time horizon is so small that there is a lot of statistical variation in the observed demands compared to the expected demands; our theoretical guarantee holds only as $\theta \to \infty$, and it stands to reason that for smaller θ the observed competitive ratio may be poor. Additionally, with small θT the integrality gap of the LP is much higher. Larger values of θT allow for enough time for observed demands to approach (statistically) their expected values, in which case the observed competitive ratios are truer measures of the actual performances of the algorithms. Note that given the time-homogenous nature of the problem, changing θT *does not change the size of the LP: so the time to implement MPFC and IPFC do not change.* What does change is the length of time our simulation runs, but even that is insignificant. We continue with $\theta T = 10,000$ for the rest of this section.

6.5 Sensitivity to number of facilities

Figure 5 shows the competitive ratios as the number of facilities changes. As is to be expected, with more facilities the competitive ratio is worse. However, even at the high end of 9 facilities, the competitive ratios when minimizing shipping costs of IPFC and Myopic are 1.05 and 1.10 respectively; when minimizing the number of packages these are 1.07 and 1.15. Again, it appears that IPFC is able to recover at least half of the excess cost compared to the LP lower bound, in a way that scales well with the number of facilities.

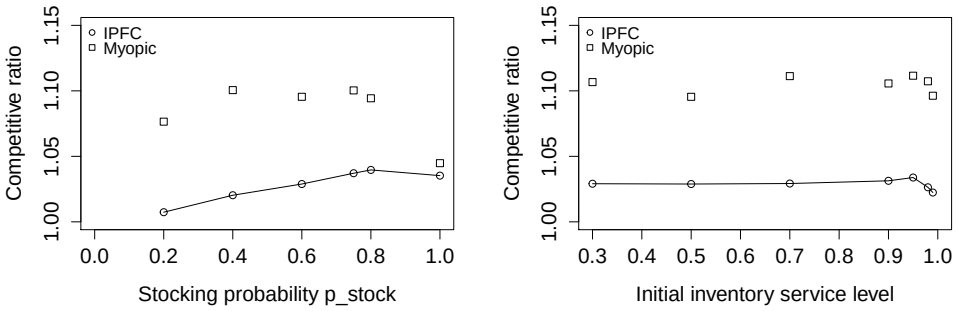

Figure 6: Performance with respect to initial inventory placement. Left panel: with $|K| = 5$ facilities, the probability for each item being in a given facility p_{stock} changes, with CSL=0.5. Right panel: Initial inventory service level changes, for each item at each facility, with $p_{stock} = 0.6$.

6.6 Sensitivity to initial inventory placement

Recall that our initial inventory placement uses two parameters, p_{stock} and CSL. For each item i and facility $k \neq 1$, the parameter p_{stock} is the probability that facility k even stocks item i, decided in an i.i.d. fashion for all item-facility pairs. Given this assignment, we then define the "service area" for each facility-item pair as the set of all customer regions for whom this facility is the nearest that stocks item i (details in appendix). Then, we set the initial inventory level S_{ki} as the quantity such that the probability that demand for item i from the service area exceeds S_{ki} is CSL.

Naturally, as $p_{stock} \to 1$ and $CSL \to 1$, both Myopic and IPFC (and in fact any reasonable algorithm) will trivially perform optimally, because there is large amount of initial inventory of every item at every facility. However, in practice, neither of these conditions hold. Each facility may stock only a subset of the items, because of considerations such as handling equipment, capacity, supplier locations, physical characteristics of items, etc.; more such constraints would make for a lower value of p_{stock}. For CSL, the obvious trade-off is that a higher CSL results in higher inventory holding costs. So, a firm would want to carefully balance both the assignment of items to facilities and the initial inventory levels so as to keep overall costs low and service levels high.

As is to be expected, lower levels of p_{stock} and lower levels of CSL result in higher competitive ratios. Once again, in all cases observed, IPFC completely dominates Myopic, with some convergence seen only when $p_{stock} = 1$. Although we show results above only for $CSL = 0.5$ when we vary p_{stock} and $p_{stock} = 0.6$ when we vary CSL, our experiments confirm these findings for a wide range of parameters. This strongly suggests that if a firm is facing real-world constraints that prevent it from keeping high levels of inventory of every item everywhere, the value of an algorithm like ours can be fairly substantial.

6.7 Extreme case: all packages of maximum size

One possible extreme case for our algorithm is when all customer demands are for packages of the same size q_0. That is, for a fixed q_0, the set of possible packages Q is the set of all $\binom{|I|}{q_0}$ subsets of I that

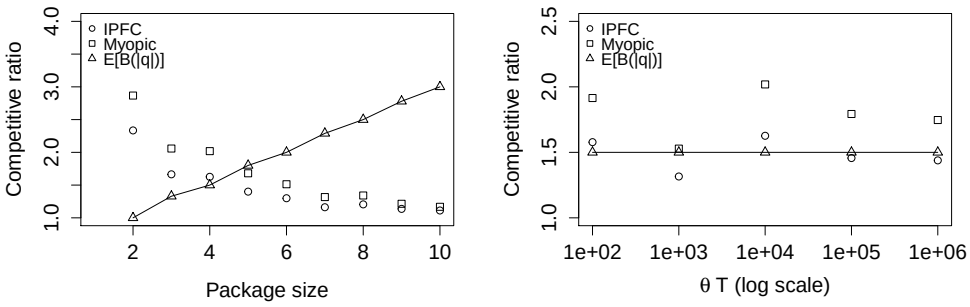

Figure 7: Performance with respect to package size q_0 when Q consists of all packages of size q_0. Left panel: Performance as q_0 changes. Right panel: q_0 fixed at 4, θT changes.

have exactly q_0 distinct items in them. In this case, we have $E[B(|q|)] = B(q_0)$.

Figure 7 shows the performance of IPFC as the package size increases (with $|I| = 10$ and $\theta T = 10,000$). We find that as q_0 (and therefore $E[B(q_0)]$ increases, the competitive ratio of IPFC (and Myopic) decreases. This suggests that our bound of $E[B(|q|)]$ is a weak bound, particularly if $|q|$ is large. Also note that although the competitive ratio of IPFC appears to exceed the $E[B(q_0)]$ bound for $q_0 \leq 4$ in the figure on the left, we believe this is because the theoretical bound holds only asymptotically. To buttress this argument, we show in the right panel of Figure 7 that as the time horizon (θT) increases, the performance of IPFC indeed drops below the $E[B(q_0)]$ bound.

7 Limits on Competitive Ratio

While the numerical studies are extremely encouraging in terms of the observed competitive ratio, the question remains open whether the provable bound on the competitive ratio can be improved. In this section, we attempt to answer that question partially. We prove by a simple reduction from the Set Cover problem that not only is it impossible to construct an algorithm to solve the problem optimally in polynomial time, but that the competitive ratio of any algorithm has to be at least $\Omega(\ln |Q|)$. We also show via example that it is impossible to construct an optimal joint distribution as in the IPFC approach.

Set Cover reduction. The Set Cover problem, at its simplest, comprises a universe U of elements, and a collection $\mathcal{S}$ of subsets of U. The objective is to select a sub-collection $\mathcal{C} \subseteq \mathcal{S}$ such that $\cup\mathcal{C} = U$ and $|\mathcal{C}|$ is minimized. The Set Cover problem is known to be not only NP-Hard, but also hard to approximate to a competitive ratio better than $O(\log |U|)$ in polynomial time (Feige 1998).

Consider the problem of fulfilling just a single order (with many items but each item demanded at most once), with the objective of minimizing the number of facilities used (equivalently, $c = 0$ and $b = 1$ in the cost function). We will reduce the Set Cover problem to this, as follows. Given an instance

$(U, \mathcal{S})$ of the Set Cover problem, map each element of U to an item i, and each set S in $\mathcal{S}$ to a unique facility $k(S)$. At facility $k(S)$, set the inventory as follows: $S_{k(S)i} = 1$ if and only if $i \in S$, otherwise $S_{k(S)i} = 0$. Now, consider the EOF problem, given these inventory levels and one single order defined by U. The optimal solution to EOF will minimize the number of facilities used, because that is all that counts in the objective function, and hence will also find the optimal set cover. So, the $\ln n$ threshold of approximation of Set Cover extends to approximating EOF even when we have just one single order.

To extend the hardness result to our problem, it helps to consider a multi-period, capacitated, version of the set cover problem (abbreviated MCSC here). In MCSC, each set $S \in \mathcal{S}$ has a (non-negative integral) capacity $cap(S)$, and in each time period a new $q \subseteq U$ arrives. Each q represents an order, and must be fulfilled by choosing a cover from $\mathcal{S}$. The objective would be to minimize the total number of sets used, ensuring that no set is used more times than $cap(S)$. It can be shown that by simply scaling up the instance of the standard set cover problem that gives a $\ln n$ lower bound, one can construct an instance of MCSC where it would be NP-hard to obtain a solution that has competitive ratio better than $\Omega(\max_q \ln |q|)$. Using the correspondence between MCSC and our problem above, this implies an $\Omega(\ln |Q|)$ lower bound on the competitive ratio of EOF. As far as we are aware, there is no literature on any problem similar to MCSC. There is literature on a problem known as "capacitated set cover" (Chuzhoy and Naor 2006), but that is a different problem: there is still only a single universe that needs to be covered, and the capacity constraints only limit how many items each set can be used to cover.

This still leaves open a gap between our competitive ratio of $E[B(|Q|)$ and the inapproximability threshold of $\Omega(\ln |Q|)$. The standard techniques used to approximate the Set Cover problem do not directly extend to our problem; the main difficulty lies in the capacity constraints. Reducing this gap remains an open question.

Non-existence of asymptotically optimal distribution for IPFC. Is it theoretically possible to construct joint distributions $\{g_j^q\}$ such that IPFC is still asymptotically optimal even if $|q| > 2$ for some q? Unfortunately, the answer is no. In principle, the *optimal* joint distributions $\{g_j^q\}$ can be computed by solving a sequence of independent LPs, one for each pair (q, j), as follows:

$$\begin{aligned} V_j^q \;=\; \min \quad & \sum_k b_{kj} \left[1 - \sum_{\sigma:\, \sigma(i) \neq k \; \forall i} g_j^q(\sigma) \right] \\ \text{s.t.} \quad & \sum_{\sigma: \sigma(i)=k} g_j^q(\sigma) = \mathbf{u}_{kij}^{q1} \;\text{ and }\; g_j^q(\sigma) \geq 0. \end{aligned}$$

(The term inside the bracket [.] is equal to the expectation $\mathbf{E}\left[\max_{i \in q} \{X_{kij}^{qt}\}\right]$ under g_j^q.) It can be argued that if $|q| \leq 2$, then $V_j^q = \sum_k b_{kj} \left(\max_{i \in q} u_{kij}^{q1}\right)$ for all pairs (q, j). In fact, one can use the joint distributions hinted in Theorem 3 as a feasible solution to V_j^q. Since the same summations also show up in J_{LP}^M (because $\mathbf{y}_{kj}^{qt} = \max_{i \in q} \mathbf{u}_{kij}^{qt}$), this suggests the asymptotic optimality of the optimal joint distributions.

This observation, however, does *not* hold if $|q| = 3$ for some q. We give a counter-example. Consider an instance of EOF where there is only one order type q_1 with exactly 3 items (i_1, i_2, i_3), 3 facilities (k_1, k_2, k_3), and 1 region (j_1). (For brevity, we will suppress notational dependency on q_1 and j_1.) For simplicity, we assume that $S_{ki} = \infty$ for all (k, i), $b_k = \$1$ for all k, $c_{k_1 i_3} = c_{k_2 i_2} = c_{k_3 i_1} = \$10$, and $c_{ki} = 0$ otherwise. An optimal solution to J_{LP}^M is given by

$$\mathbf{u}_{k_1 i_3}^t = \mathbf{u}_{k_2 i_2}^t = \mathbf{u}_{k_3 i_1}^t = 0 \quad \text{and} \quad \mathbf{u}_{kj}^t = \frac{1}{2} \quad \text{otherwise.}$$

Substituting these into the above LP, we get $V_j^q = 2$. Since $\sum_k b_k (\max_i \mathbf{u}_{ki}^t) = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = \frac{3}{2}$, this tells us that the optimal joint distribution is asymptotically $\frac{4}{3}$-competitive. (As can be seen from Table 2, our constructed heuristic also achieves this exact performance for the case $|q| \leq 3$.) We want to stress, the above LP requires one decision variable for each permutation σ. In addition, we also have to solve this LP for each pair (q, j), which may not be the most efficient pursuit especially if $|q|$ is typically large. In contrast to this, our construction of $\{g_j^q\}$ does *not* require *any* optimization at all. So, it is relatively easy to implement.

8 Conclusion

The significant growth in online retail, and the availability of very precise data about consumer preferences, has resulted in the emergence of several new practices in the delivery of goods to consumers. Despite this, academic research in this area is relatively sparse. Our work addresses one piece of this larger research landscape: optimizing the fulfillment of multi-item orders in the presence of inventory constraints.

Several promising directions of research remain. With respect to the problem addressed in this paper, finding a heuristic with a better competitive ratio is an open question, although our hardness result indicates that an optimal algorithm that is computationally tractable, under the general framework of our paper, cannot exist. It also remains open whether in fact a stronger lower bound exists for the competitive ratio for this problem. The most direct natural extension involves delivery time windows: when each order also specifies a deadline for delivery. Acimovic et al. (2012) address a version of this problem with single-item orders, and we conjecture that our approach can also be extended to handle time windows by adding one more set of variables in the problem. Other promising directions of future research are the incorporation of inventory management policies into the fulfillment problem and the incorporation of non-stationary demand rates that must be learned over time. We expect to see substantial research in this overall area in the near future.

9 Acknowledgments

The numerical study in Section 6 was conducted with the assistance of Manqi Li and Jianyu Liu, undergraduate students at the University of Michigan—Shanghai Jiao Tong University Joint Institute.

The authors thank them for their help.

The paper has also benefited significantly from the input of anonymous referees and an editorial review team. We sincerely thank them for their feedback and suggestions.

APPENDIX A: Proofs

Proof of Theorem 1. Let $\mathbf{u}^{qt}_{\sigma j} = \mathbf{u}^{q1}_{\sigma j} = \mathbf{U}^q_{\sigma j}/(T\lambda^q_j)$. We consider a variant of PFC (VPFC) which works as follows: during period t, fulfill order type q from region j according to σ with probability $\mathbf{u}^{q1}_{\sigma j}$ *regardless* of availability. So, in contrast to PFC, at the end of selling horizon, VPFC incurs a large penalty for each violation of inventory constraints. Let $X^q_{\sigma j}(\theta)$ denote the number of times order type q from region j are fulfilled according to σ throughout the selling horizon. Total cost under VPFC is given by

$$C_{VPFC}(\theta) = \sum_j \sum_q \sum_\sigma c^q_{\sigma j} X^q_{\sigma j}(\theta) \; + \; c_p \sum_k \sum_i \left[\sum_j \sum_{q\in i} \sum_{\sigma:\sigma(i)=k} X^q_{\sigma j}(\theta) - S_{ki}(\theta) \right]^+,$$

where $c_p = \sum_j \sum_q \sum_\sigma c^q_{\sigma j}$. Obviously, $C_{PFC}(\theta) \le C_{VPFC}(\theta)$. So, we can bound $\mathbf{E}[C_{PFC}(\theta)] - J_{LP}(\theta)$ with $\mathbf{E}[C_{VPFC}(\theta)] - J_{LP}(\theta)$. Since $\mathbf{E}[X^q_{\sigma j}(\theta)] = \mathbf{U}^q_{\sigma j}(\theta)$ and there exists a positive constant M independent of $\theta > 0$, k, and i such that

$$\begin{aligned}
\mathbf{E}\left[\left(\sum_j \sum_{q\in i} \sum_{\sigma:\sigma(i)=k} X^q_{\sigma j}(\theta) - S_{ki}(\theta) \right)^+ \right] \;&\le\; \mathbf{E}\left[\left(\sum_j \sum_{q\in i} \sum_{\sigma:\sigma(i)=k} (X^q_{\sigma j}(\theta) - \mathbf{U}^q_{\sigma j}(\theta)) \right)^+ \right] \\
&\quad + \; \mathbf{E}\left[\left(\sum_j \sum_{q\in i} \sum_{\sigma:\sigma(i)=k} \mathbf{U}^q_{\sigma j}(\theta) - S_{ki}(\theta) \right)^+ \right] \\
&\le \; \sum_j \sum_{q\in i} \sum_{\sigma:\sigma(i)=k} \mathbf{E}\left[(X^q_{\sigma j}(\theta) - \mathbf{U}^q_{\sigma j}(\theta))^+ \right] \\
&\le \; \sum_j \sum_{q\in i} \sum_{\sigma:\sigma(i)=k} \sqrt{\mathrm{VAR}(X^q_{\sigma j}(\theta))} \;\le\; M\sqrt{\theta},
\end{aligned}$$

(the second inequality follows because $\sum_j \sum_{q\in i} \sum_{\sigma:\sigma(i)=k} \mathbf{U}^q_{\sigma j}(\theta) \le S_{ki}(\theta)$ and the last inequality follows because, by Binomial formula, $\mathrm{VAR}(X^q_{\sigma j}(\theta)) \le T\theta$) we conclude that $\mathbf{E}[C_{PFC}(\theta)] - J_{LP}(\theta) \le M\sqrt{\theta}$. This completes the proof. ■

Proof of Theorem 2. Similar to the proof of Theorem 1, we consider a variant of MPFC which works as follows: During period t, we fulfill item $i \in q$ from region j from facility k with probability $\mathbf{u}^{qt}_{kij} = \mathbf{U}^q_{kij}/(T\lambda^q_j)$ *regardless* of availability. We denote this heuristic by VPFC, where the "V" stands for Violated. Since VPFC ignores the inventory constraints, at the end of selling horizon, VPFC incurs a large penalty c_p for each violation of inventory constraints. Let D^{qt}_j be a binary random variable, $D^{qt}_j = 1$ if an order type q arrives from region j during period t and 0 otherwise. Total cost under

VPFC is given by

$$C_{VPFC}(\theta) \;=\; \sum_{t,j,q,k} D_j^{qt} \left[\sum_{i\in q} c_{kij}\, X_{kij}^{qt}(\theta) \;+\; b_{kj}\, \max_{i\in q}\, X_{kij}^{qt}(\theta) \right] + \; c_p \sum_k \sum_i \left[\sum_j \sum_{q\in i} D_j^{qt} X_{kij}^{qt}(\theta) - S_{ki}(\theta) \right]^+,$$

where $c_p = \sum_j \sum_q \sum_k \left[\sum_{i\in q} c_{kij} + b_{kj} \right]$ is the stock-out penalty. Obviously, $C_{MPFC}(\theta) \le C_{VPFC}(\theta)$. This allows us to bound $\mathbf{E}[C_{MPFC}(\theta)]/J^*(\theta)$ with $\mathbf{E}[C_{VPFC}(\theta)]/J^*(\theta)$. Since $J^*(\theta) \ge J_{LP}^M(\theta)$, we can further bound $\mathbf{E}[C_{VPFC}(\theta)]/J^*(\theta)$ with $\mathbf{E}[C_{VPFC}(\theta)]/J_{LP}^M(\theta)$. By the same argument as in the proof of Theorem 1, the penalty cost in $C_{VPFC}(\theta)$ (the term with $[.]^+$) is of order $O(\sqrt{\theta})$. Since $J_{LP}^M(\theta) = \theta\, J_{LP}^M$, the $O(\sqrt{\theta})$ term vanishes asymptotically as $\theta \to \infty$. So, we can focus on $\sum_{t,j,q,k} D_j^{qt} \left[\sum_{i\in q} c_{kij}\, X_{kij}^{qt}(\theta) \;+\; b_{kj}\, \max_{i\in q}\, X_{kij}^{qt}(\theta) \right]$. Observe that

$$\mathbf{E}\left[\max_{i\in q}\, X_{kij}^{qt}(\theta) \right] \;=\; \mathbf{E}\left[1 - \prod_{i\in q} (1 - X_{kij}^{qt}(\theta)) \right] \;=\; 1 - \prod_{i\in q} \mathbf{E}\left[1 - X_{kij}^{qt}(\theta) \right].$$

Since $\mathbf{E}[X_{kij}^{qt}(\theta)] = \mathbf{u}_{kij}^{qt}$, by Bernoulli's inequality, $\mathbf{E}\left[\max_{i\in q}\, X_{kij}^{qt}(\theta) \right] \le \sum_{i\in q} \mathbf{u}_{kij}^{qt} \le |q| \max_{i\in q} \mathbf{u}_{kij}^{qt}$. So, we can bound:

$$\mathbf{E}\left[D_j^{qt} \left(\sum_{i\in q} c_{kij}\, X_{kij}^{qt}(\theta) \;+\; b_{kj}\, \max_{i\in q}\, X_{kij}^{qt}(\theta) \right) \right] \;\le\; \lambda_j^q \left(\sum_{i\in q} c_{kij} \mathbf{u}_{kij}^{qt} + b_{kj} |q| \max_{i\in q} \mathbf{u}_{kij}^{qt} \right).$$

Using the fact that $\mathbf{u}_{kij}^{qt} = \mathbf{u}_{kij}^{q1}$ for all $t \ge 1$ and the inequality $\frac{\sum_i (a_i + b_i c_i)}{\sum_i (a_i + b_i)} \le \frac{\sum_i b_i c_i}{\sum_i b_i}$ for all $a_i > 0$, $b_i > 0$, and $c_i \ge 1$, we have:

$$\begin{aligned} \lim_{\theta\to\infty} \frac{\mathbf{E}[C_{VPFC}(\theta)]}{J_{LP}^M(\theta)} \;&\le\; \frac{\sum_{q,j,k} \lambda_j^q \left(\sum_{i\in q} c_{kij} \mathbf{u}_{kij}^{qt} + b_{kj} |q| \max_{i\in q} \mathbf{u}_{kij}^{qt} \right)}{\sum_{q,j,k} \lambda_j^q \left(\sum_{i\in q} c_{kij} \mathbf{u}_{kij}^{qt} + b_{kj} \max_{i\in q} \mathbf{u}_{kij}^{qt} \right)} \\ &\le\; \frac{\sum_{q,j,k} \lambda_j^q b_{kj} |q| \max_{i\in q} \mathbf{u}_{kij}^{qt}}{\sum_{q,j,k} \lambda_j^q b_{kj} \max_{i\in q} \mathbf{u}_{kij}^{qt}} \;=\; \sum_{q,k,j} |q|\, F(q,k,j). \end{aligned}$$

This completes the proof. ■

Proof of Lemma 1. Arguing as in Section 5, we can write:

$$\lim_{\theta\to\infty} \frac{\mathbf{E}[C_{MPFC}(\theta)]}{J^*(\theta)} \;\le\; \frac{\sum_{j,q,k} \lambda_j^q\, \mathbf{y}_{kj}^{q1}\, |q|}{\sum_{j,q,k} \lambda_j^q\, \mathbf{y}_{kj}^{q1}}.$$

Since $\{\mathbf{y}^{q1}_{kj}\}$ is an optimal solution to J^M_{LP}, it must satisfy $\mathbf{y}^{q1}_{kj} = \max_{i\in q} \mathbf{u}^{q1}_{kij}$. We will now provide a lower and an upper bound for the sum $\sum_k \max_{i\in q} \mathbf{u}^{q1}_{kij}$. The lower bound is straightforward:

$$\sum_k \max_{i\in q} \mathbf{u}^{q1}_{kij} \;\geq\; \sum_k \frac{\sum_{i\in q} \mathbf{u}^{q1}_{kij}}{|q|} \;=\; \frac{1}{|q|}\sum_{i\in q}\sum_k \mathbf{u}^{q1}_{kij} \;=\; 1,$$

where the last equality follows because $\sum_k \mathbf{u}^{q1}_{kij} = 1$. We now give an upper bound. Obviously, since $\mathbf{u}^{q1}_{kij} \leq 1$, we must have $\sum_k \max_{i\in q} \mathbf{u}^{q1}_{kij} \leq |S_K|$. But, also,

$$\sum_k \max_{i\in q} \mathbf{u}^{q1}_{kij} \;\leq\; \sum_k \sum_{i\in q} \mathbf{u}^{q1}_{kij} \;=\; \sum_{i\in q}\sum_k \mathbf{u}^{q1}_{kij} \;=\; |q|.$$

We conclude that $\sum_k \max_{i\in q} \mathbf{u}^{q1}_{kij} \leq \min\{|S_K|, |q|\}$. Now, let $F(q,k,j) = \lambda^q_j \mathbf{y}^{q1}_{kj} / \sum_{j',q',k'} \lambda^{q'}_{j'} \mathbf{y}^{q'1}_{k'j'}$. Applying the above lower and upper bounds to $F(q,k,j)$ immediately yields the result. This completes the proof. ■

APPENDIX B: Numerical study description

In this appendix, we provide a detailed description of our numerical study in Section 6. Our numerical study was performed entirely using publicly-available data, and we provide enough detail here to allow readers to completely replicate our study. We first describe our numerical study domain (locations and distances), followed by the initial inventory placement, followed by the actual simulation details.

Geographical domain. Our study is placed in the continental United States. For customer locations, we start with the 100 largest metropolitan statistical areas (MSAs) as estimated by the US Census Bureau(U.S. Census Bureau 2014) in the US, then remove Honolulu. We take into account the population of the MSAs in generating demand, so a more populous city generates proportionately more demand. Then, given the number of customer locations $|J|$ we are interested in, we simply select uniformly at random from this set of 99 cities.

For the list of potential facility locations, we use (Chicago Consulting 2013), who report the locations of the best n facilities for minimizing shipping cost in the US, for $|J| = 1, 2, \ldots, 10$. We remove the Puerto Rico locations from this list, and select networks with $|J| = 2, 5$, and 9. Thus, although we do not optimize the location of the facility ourselves (optimal location of fulfillment centers for ecommerce is a different research question), our chosen fulfillment center locations arguably are somewhat close to optimal.

We use UPS ground shipping rates to estimate our shipping cost function. With 99 destination cities and 9 potential facility locations, there are 891 possible origin-destination pairs. For each such pair, we get the shipping rate from UPS for a package of weight 1, 2, or 3 pounds, choosing package weight uniformly at random. We then first estimated the following linear shipping cost model, where d_{kj} is the distance in miles from facility i to customer region j and $|q| \in \{1,2,3\}$ is the number of items in the package, assuming each item weighs exactly one pound:

$$cost(q,k,j) = \beta_0 + \beta_1|q| + \beta_2 d_{kj} + \beta_3|q|d_{kj}$$

The estimate of coefficient β_2 was insignificant in the above estimation, so we removed it and re-estimated the parameters. This resulted in the following final estimate, with an R^2 of 94.5% and p-values of the order of 10^{-15}:

$$cost(q,k,j) = 8.759 + 0.423|q| + 0.000541|q|d_{kj}$$

In our entire numerical study, we either used this shipping cost function or minimized the number of packages. However, our methodology is such that any shipping cost function should be easily usable.

Note also that we use a fictitious facility indexed $j = 1$ with infinite supply to model the costs incurred due to stockouts. To make the problem reasonable, facility 1 should have higher costs than regular facilities. We implement this using a penalty factor of 2. That is, we set $b_{1j} = 2 \times 8.759$ and $c_{1ij} = 2 \times (0.423 + 0.00541 \max_{k,j} d_{kj})$.

Demand Forecasts and Initial inventory. First, we describe how we construct the demand rates λ in our simulation, given a set of items $I = \{1, 2, \ldots, |I|\}$. The main problem this paper solves is that of fulfillment when customers order baskets of more than one item. So, we need to generate demand rates for baskets of items. However, for both real-world reasons and analytical tractability reasons, we cannot consider all possible baskets in 2^I. So, we consider a smaller set of baskets, defined by two parameters: n_{max} denotes the maximum order/basket size, and n_0 denotes the number of baskets with positive demand for each size less than or equal to n_{max}. For the most part, we use $n_{max} = n_0 = 5$, although we test various other values of both parameters.

Given n_{max}, we first generate the total probability of all orders of sizes $0, 1, 2, \ldots, n_{max}$. We denote these $p(n)$, and $p(n)$ is chosen uniformly at random so that $p(n) \in [0,1]$ $\forall n$ and $\sum_0^{n_{max}} p(n) = 1$. Note that knowing $p(n)$ we can directly compute $E[B(|q|)]$ as $(\sum_{n=1}^{n_{max}} p(n).B(n))/(1-p(0))$; it is easy to verify that with $n_{max} = 5$ we have $E[B(|q|)] = 1.32$ which is what we observe in the left panel of Figure 3.

Let λ^q denote the total demand rate for order q from all regions; that is, $\lambda^q = \sum_{j \in J} \lambda_j^q$. Given $p(n)$, we first generate λ^q as follows. For each $n \in \{1, 2, \ldots, n_{max}\}$, we select uniformly at random $\min\{n_0, \binom{|I|}{n}\}$ subsets of I^n to have positive demand rates. Let $Q(n)$ denote this subset. We then choose λ^q to be uniformly at random in $[0, p(n)]$ such that $\sum_{q \in Q(n)} \lambda^q = p(n)$. Lastly, we generate λ_j^q by simply scaling λ^q to the population of each city $j \in J$. That is, $\lambda_j^q = \lambda^q . pop(j) / \sum_{j \in J} pop(j)$, where $pop(j)$ is the population of the metropolitan statistical area j.

We now describe the initial inventory placement. In practice, it is often the case that any given single facility stocks only a subset of all the items sold by the retailer; this may be because of supplier considerations, material handling requirements, equipment, capacity constraints, etc. To model this, we use a parameter $p_{stock} \in [0,1]$, such that for any given facility k and item i, the probability that k stocks i is p_{stock}. That is, $P(S_{ki} > 0) = p_{stock}$, i.i.d. for all k, i. In our numerical studies we do test the

sensitivity with respect to p_{stock}, including the case $p_{stock} = 1$ where every facility stocks every item. For most of our experiments, we use $p_{stock} = 0.75$.

Now consider a facility k and item i for which we have determined that $S_{ki} > 0$. How should we set the initial inventory level S_{ki}? It is reasonable to believe that the firm would compute the expected demand from all customers for whom this facility is the nearest that stocks item i, and keep inventory equal to that level plus some safety stock, in a newsvendor fashion. That is exactly what we do. Formally, for a given facility k and item i, we first find the set of customers $J(k,i)$ for whom it should stock inventory: $J(k,i) = \{j \in J : d_{kj} = \min_{k' \in K: S_{k'i} > 0} d_{k'j}\}$. Define $\lambda(k,i) = \sum_{j \in J(k,i), q \ni i} \lambda_j^q$; this is the total incoming demand to facility k for item i from all orders that contain i. We then compute the expected value and standard deviation of the demand for item i from $J(k,i)$, given the time horizon θT, as follows: $\mu(k,i) = \theta T \lambda(k,i)$ and $\sigma(k,i) = \sqrt{\theta T \lambda(k,i)(1 - \lambda(k,i))}$. Next, given another global parameter CSL (for cycle service level), we simply use the newsvendor fractile at that level to determine the starting inventory: $S_{ki} = \mu(k,i) + z_{CSL}\sigma(k,i)$, where z_{CSL} is the inverse of the standard normal distribution at probability CSL. Our default value for CSL in the numerical experiments is 0.5, but we test values ranging from 0.3 to 0.99 and report the results in Figure 6.

Simulation procedure. Given the setup above, our simulation process is fairly simple. Once the parameters $|I|, |J|, |K|, n_{max}, n_0, p_{stock}, CSL$, and θT are defined, we generate the sets I, J, K, and Q, and the matrices λ, c, b, and S. We then compute the values $\{u_{kij}^{qt}\}$ that define the MPFC algorithm, as well as the values $\{g_j^q\}$ which defines the IPFC algorithm. We also implement a myopic algorithm, which works as follows: given an order q at time t from customer region j, it simply fulfills every item in q from the facility nearest to j that has positive inventory of that item.

We then generate a single demand sequence based on λ and θT. All three algorithms are applied to *the same demand sequence*. Therefore, the variation in the demand affects all three algorithms equally, and this allows for a better comparison of the algorithms. This constitutes one simulation trial.

For each setting of the parameters, we run 30 simulation trials. This allows us to obtain statistical significance in our results, as detailed in Section 6.2. In total, we ran over 11,000 simulation trials with several different combinations of parameters. A selection of these that are particularly insightful are reported in Section 6.

In our computing environment, the *total run time* of these 11,000 trials in series was about 20 days, so that for a single simulation trial the total time taken is about 2 to 3 minutes. As mentioned in Section 6.2, this is highly encouraging in terms of the ability to scale to levels appropriate for large firms.

References

Acimovic, Jason, Stephen Graves, Anonymous Authors. 2012. Making better fulfillment decisions on the fly in an online retail environment. Working Paper, MIT.

Agatz, Niels A.H., Moritz Fleischmann, Jo A.E.E. van Nunen. 2008. E-fulfillment and multi-channel distribution—A review. *European Journal of Operational Research* **187** 339–356.

Ahuja, Ravindra, Thomas Magnanti, James Orlin. 1993. *Network Flows: Theory, Algorithms, and Applications*. Pearson.

Ata, Baris, Sunil Kumar. 2005. Heavy traffic analysis of open processing networks with complete resource pooling: Asymptotic optimality of discrete review policies. *The Annals of Applied Probability* **15**(1A) 331–391.

Cattani, Kyle D., Gilvan C. Souza. 2002. Inventory rationing and shipment flexibility alternatives for direct market firms. *Production and Operations Management* **11**(4) 441–457.

Chicago Consulting. 2013. 10 best warehouse networks 2013. http://www.chicago-consulting.com/10best.shtml.

Chuzhoy, Julia, Joseph Naor. 2006. Covering problems with hard capacities. *SIAM Journal on Computing* **36**(2) 498–515.

Ciocan, Dragos Florin, Vivek Farias. 2012. Model predictive control for dynamic resource allocation. *Mathematics of Operations Research* **37**(3) 501–525.

Cooper, William L. 2002. Asymptotic behavior of an allocation policy for revenue management. *Operations Research* **50**(4) 720–727.

Feige, Uriel. 1998. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM* **45**(4) 634–652.

Gallego, Guillermo, Garrett van Ryzin. 1994. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science* **40**(8) 999–1020.

Gallego, Guillermo, Garrett van Ryzin. 1997. A multiproduct dynamic pricing problem and its applications to network yield management. *Operations Research* **45**(1) 24–41.

Halfin, Shlomo, Ward Whitt. 1981. Heavy-traffic limits for queues with exponentially many servers. *Operations Research* **29**(3) 567–588.

Harrison, J. Michael. 1998. Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *The Annals of Applied Probability* **8**(3) 822–848.

Huh, Wonghee Tim, Ganesh Janakiraman, John Muckstadt, Paat Rusmevichientong. 2009a. An adaptive algorithm for finding the optimal base-stock policy in lost sales inventory systems with censored demand. *Mathematics of Operations Research* **34**(2) 397–416.

Huh, Wonghee Tim, Ganesh Janakiraman, John Muckstadt, Paat Rusmevichientong. 2009b. Asymptotic optimality of order-up-to policies in lost sales inventory systems. *Management Science* **55**(3) 404–420.

Jasin, Stefanus, Sunil Kumar. 2012. A re-solving heuristic with bounded revenue loss for network revenue management with customer choice. *Mathematics of Operations Research* **37**(2) 313–345.

Jasin, Stefanus, Sunil Kumar. 2013. Analysis of deterministic LP-based heuristics for revenue management. Working paper, University of Michigan.

Levi, Retsef, Ana Radovanović. 2010. Provably near-optimal LP-based policies for revenue management in systems with reusable resources. *Operations Research* **58**(2) 503–507.

Liu, Qian, Garrett van Ryzin. 2008. On the choice-based linear programming model for network revenue management. *Manufacturing and Service Operations Management* **10**(2) 288–310.

Maglaras, Constantinos. 2000. Discrete-review policies for scheduling stochastic networks: trajectory tracking and fluid-scale asymptotic optimality. *The Annals of Applied Probability* **10**(3) 897–929.

Mahar, Stephen, P. Daniel Wright. 2009. The value of postponing online fulfillment decisions in multi-channel retail/e-tail organizations. *Computers and Operations Research* **36** 3061–3072.

Plambeck, Erica. 2008. Asymptotically optimal control for an assemble-to-order system with capacitated component production and fixed transportation costs. *Operations Research* **56** 1158–1171.

Plambeck, Erica, Amy Ward. 2006. Optimal control of a high-volume assemble-to-order system. *Mathematics of Operations Research* **31** 453–477.

Reiman, Martin I., Qiong Wang. 2008. An asymptotically optimal policy for a quantity-based network revenue management problem. *Mathematics of Operations Research* **33**(2) 257–282.

Simchi-Levi, David, S. David Wu, Z.J. Max Shen. 2004. *Handbook of Quantitative Supply Chain Analysis: Modeling in the E-Business Era*. Kluwer.

UPS. 2012. UPS rate and service guide: 2012 standard list rates.

U.S. Census Bureau. 2014. Annual estimates of the resident population: April 1, 2010 to july 1, 2013 - united states – metropolitan and micropolitan statistical area.

U.S. Department of Commerce. 2013. Quarterly retail e-commerce sales, 3rd quarter 2013.

Xu, Ping Josephine, Russell Allgor, Stephen Graves. 2009. Benefits of reevaluating real-time order fulfillment decisions. *Manufacturing and Service Operations Management* **11**(2) 340–355.