Optimal Online Assignment with Forecasts

Erik Vee
Yahoo! Research
Santa Clara, CA
erikvee@yahoo-inc.com

Sergei Vassilvitskii
Yahoo! Research
New York, NY
sergei@yahoo-inc.com

Jayavel Shanmugasundaram
Yahoo! Research
Santa Clara, CA
jaishan@yahoo-inc.com

ABSTRACT

Motivated by the allocation problem facing publishers in display advertising we formulate the *online assignment with forecast* problem, a version of the online allocation problem where the algorithm has access to random samples from the future set of arriving vertices. We provide a solution that allows us to serve Internet users in an online manner that is provably nearly optimal. Our technique applies to the forecast version of a large class of online assignment problems, such as online bipartite matching, allocation, and budgeted bidders, in which we wish to minimize the value of some convex objective function subject to a set of linear supply and demand constraints.

Our solution utilizes a particular *subspace* of the dual space, allowing us to describe the optimal primal solution implicitly in space proportional to the demand side of the input graph. More importantly, it allows us to prove that representing the primal solution using such a compact allocation plan yields a robust online algorithm which makes near-optimal online decisions. Furthermore, unlike the primal solution, we show that the compact allocation plan produced by considering only a sampled version of the original problem generalizes to produce a near optimal solution on the full problem instance.

Categories and Subject Descriptors

F.2.0 [**Analysis of Algorithms and problem complexity**]: General; G.2.3 [**Discrete Mathematics**]: Applications

General Terms

Algorithms, Theory

Keywords

Online Matching, Computational Advertising, Guaranteed Delivery

EC'10, June 7–11, 2010, Cambridge, Massachusetts, USA.

1. INTRODUCTION

Display advertising—the practice of showing graphical ads on webpages—is a multibillion dollar business. In the traditional display advertising setting a large publisher, such as Yahoo!, enters into contracts with individual advertisers guaranteeing to show their advertisements to a pre-specified number of users matching the advertiser's targeting constraints. For example, an advertiser may wish to target computer scientists in New York visiting Fashion websites, and ask for 50 Million such impressions. In guaranteed contracts the publisher takes on the risk of uncertainty in supply and guarantees displaying the ad 50M times to users matching the targeting constraints. As the publisher faces penalties for underdelivering, it is in his best interest to deliver on every contract it promises.

A single publisher simultaneously manages tens of thousands of contracts competing for overlapping inventory. For example, a different advertiser may be seeking to advertise to New Yorkers in the 18-26 age group. Yet another one may only be interested in the 18-26 demographic across all of the United States, and so on. Then, given a user visit by a 25-year old Manhattan computer scientist to a fashion website, the publisher must decide which of the possibly thousands of matching advertisements to show. The split second decision must be consistent—ensuring that at the end all of the guarantees are met.

The overall scenario can be modeled as an allocation problem. Consider a bipartite graph $G = (I \cup J, E)$. The set I represents the individual user visits, J represents the guaranteed contracts, and there is an edge $(i, j) \in E$ if user i matches the targeting constraints of contract j. Moreover, each advertiser has an overall demand d_j; while a user has a supply parameter s_i, representing how many times the user appears during the time period. The goal of the publisher is to find an allocation of users to advertisers so that all of the supply and demand constraints are satisfied: at most one ad is shown on each user visit (supply constraint), and each advertiser fulfills its demand (demand constraint). A toy example of such a graph is shown in Figure 1.

To further complicate the overall problem, publishers and advertisers are not happy simply to satisfy the guarantees. Indeed, how would an advertiser feel to only have his national ad shown in only one state, or to have all of the ads for a month-long campaign shown in one day? In fact, it has previously been argued [4] that it is in the long term interest of the publisher to strive for *fair* or representative allocations. Of course, even more complicated or detailed objectives are often desirable. Thus, the allocation problem

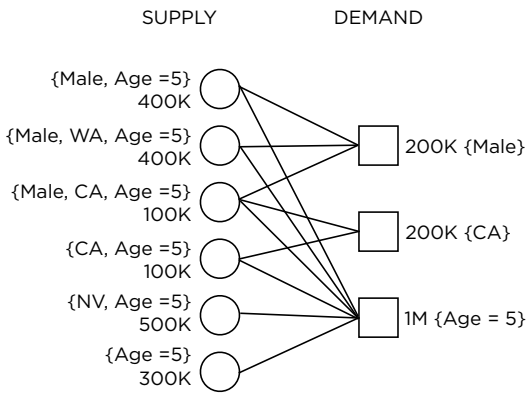

Figure 1: An example allocation problem with six types of different users and three guaranteed contracts.

typically has a complex objective function associated with it.

The one piece of good news is that, despite the pessimistic worst-cast view of many online algorithms, the landscape of future inventory is known to some degree. Thus, in theory, we could produce the bipartite graph G (together with the supply and demand constraints), and actually find the optimal solution for a given objective function. However, this too has many problems in practice.

The first is that G is many orders of magnitude too large. Of course, we could down-sample the graph and solve a smaller problem. But what would the solution on the down-sampled problem even tell us? For any user visit that was not sampled, we would have no information at all. And for user visits that were sampled, what guarantees do we have that the solution is meaningful for the actual graph?

A second issue has to do with serving. Suppose we actually found an optimal solution using the full graph G. There is no manageable way to have ad servers actually implement this solution. It would require identifying a specific, *predicted* user visit as it arrived, then finding what the optimal solution says to do. Recall that the full graph G has billions of user visits, with trillions of edges. The full optimal solution could not be communicated to hundreds (or more) servers and actually served.

A third issue is robustness. Although we could have good forecasts for G, they will never be perfect. We would still like to produce a solution that is good, even with imperfect forecasts. Even more importantly, we need to generalize to handle user visits that were never forecast; how do we serve to a user that we never optimized for?

Our solution addresses each of this issues. We demonstrate a method to produce a compact *allocation plan* using a sampled version of G. This allocation plan is small, using just $O(1)$ state per contract. Further, we guarantee that the serving decisions made using this allocation plan are nearly optimal (within $(1-\varepsilon)$), even when the allocation plan was computed on a sampled graph with imperfect forecasts. The ad server implementation of this plan is simple and requires no state, other than the allocation plan and contracts themselves; we do not need to remember how many times each ad has been served.

1.1 Online allocation with forecast

The problem facing the publisher is a variant of the online bipartite matching problem. This problem has a rich history, beginning with a celebrated result by Karp et al. [5] who provide a randomized online algorithm that can find a matching of size $n(1-\frac{1}{e})$. They further show that the result is tight: in the adversarial model no algorithm can achieve a better competitive ratio.

Recent work [2, 3] has argued that the input is typically far from worst case, and has resulted in algorithms that break the $1-1/e$ threshold. Feldman et al. [3] assume that the set of user visits is drawn i.i.d from a known distribution $\mathcal{D}$. Devanur and Hayes [2] propose using variations on the *random order model*. In this scenario, the main assumption is that the set of user visits, I, is fixed, but unknown, and a random permutation of I arrives in the online fashion.

We introduce the *online assignment with forecast* problem, a natural generalization of the bipartite matching problem that occurs in many online applications. Instead of assuming that the input is drawn i.i.d. from a known distribution or that arrives in a random order, we give the algorithm the ability to obtain a sample the set of vertices that will arrive online (i.e., a sample of the future). This aspect models common real world scenarios, where the input to the online algorithm is not entirely unknown, but can be modeled in some reasonable manner.

We propose a two-phase model for solving the online assignment with forecast problem. In the first (offline) phase, we use a black box to obtain a sample of vertices that will arrive online, and compute a *compact allocation plan*. In the second (online) phase, we use the allocation plan to compute assignments for vertices as they arrive. In this model we show that the number of samples required in the offline phase is relatively small, roughly quadratic in the number of the known vertices (and for the problems described above, independent of the number of unknown vertices that arrive online). Further, we show that the allocation plan is also small (linear in the size of the known set) and robust, leading to a near optimal solution with high probability.

More formally, given a graph on $|I|$ source nodes and $|J|$ destination nodes, we show that there exists a concise representation of the solution that takes only $O(|J|)$ space, *independent* of the number of source nodes, $|I|$. (Note that the simple description of the matching in terms of edges may take $O(|I||J|)$ space; for convenience, we assume each real number takes $O(1)$ space.). We call such a representation a *compact allocation plan.* We show that for a large set of objective functions, not only does such a short description exist, but it can be used to efficiently reconstruct the full solution in time $\tilde{O}(|\Gamma(j)|)$ per node (where following standard notation $\Gamma((j)$ is the neighborhood of node j in the allocation graph).

We then turn to solving the optimization problem at hand. We show that taking a judicious sample of nodes, $\hat{I} \subseteq I$, and then solving the assignment problem on the graph induced by $\hat{I}$ and J allows us to recover a near-optimal solution to the full problem. We note here that the sample $\hat{I}$ is not a uniform sample of all nodes in I. Rather, for each $j \in J$ we use a variant of importance sampling to sample nodes from $\Gamma(j) \subseteq I$. This restriction seems necessary, for consider a

j with a small $\Gamma(j)$. Unless some of the nodes $i \in \Gamma(j)$ are selected, we have no information to guide our allocation choices for j. As before, the online algorithm needs only the $O(|J|)$-space compact allocation plan.

1.2 Related work

As we mentioned above the online allocation problem is a variant of the online bipartite matching problem. Karp et al. [5] proved a lower bound of $1 - 1/e$ on the competitive ratio and gave an algorithm matching that bound. Recently, Feldman et al. [3] broke through that barrier and gave a 0.67 competitive algorithm relying on the power of two choices. While their algorithm achieves a better competitive ratio, it is not applicable in our setting, in addition to assuming that the online arrivals are independent, it requires one to maintain $O(|I|)$ state, which is prohibitively large.

A closely related problem is the AdWords problem from sponsored search advertising, introduced by Mehta et al.[6]. In this problem at every point in time a set of bidders submit bids on an item up for sale and the objective is to select the winning bidders in such a way that by the end every bidder's budget is (nearly) exhausted. Here again, the exact set of items (e.g., keywords in search advertising) is very large and can only be approximately predicted based on historical logs, and the goal is to optimally select the winning bidders at every point in time.

Mehta et al. [6] show a $1 - 1/e$ competitive ratio for this problem under worst case assumptions. In recent work Devanur and Hayes [2] show that in the random order model, a $1-\epsilon$ competitive algorithm exists for finding a revenue maximizing matching, provided that each bid is at most a ϵ^3/n^2 fraction of the overall revenue. Their algorithm is tuned for the specific case of revenue-maximization, and they leave the problem of finding fair allocation as an open problem.

On a high level we employ techniques similar to those of Devanur and Hayes [2], namely looking at the dual space on a sampled version of the problem. There are, however, two key differences. The first lies in the ability to handle non-linear convex objectives, which can be used to encode sophisticated fairness constraints [4]. The other is in the nature of the sampling: a uniform sample is not sufficient because in that case, small contracts may not have any matching user visits in the sample. In the work of [2], this does not cause a problem because bidders with small budgets can simply be ignored. In our work, however, these are hard constraints that cannot be violated; further, small contracts may also contribute in a non-linear way to the overall objective.

2. BACKGROUND AND DEFINITIONS

This paper is concerned with the problem of online assignment with forecast, which extends to a broad range of optimization problems expressible with a set of linear constraints. However, for the purposes of exposition, we will focus on the problem of *online allocation with forecast*. All of our results will be stated and proven for this more specific problem. The more general results appear in the technical report for this paper [8].

2.1 General framework

The *online assignment with forecast* problem is a variant of the classic online bipartite matching problem, in which we are allowed to specify additional linear constraints for each supply and demand node.

There is an underlying bipartite graph, with the right-hand side nodes are known apriori. The left-hand side nodes arrive one at a time, together with a specified amount of supply and the graph neighborhood of the node. In an online manner, the algorithm must decide how much supply to assign to each right-hand side node in the neighborhood, subject to the additional linear constraints specified; once an assignment is made, the algorithm may not change the decision. (In the classic matching problem, we have no additional constraints.)

The ultimate goal in this setting is to minimize some objective function, over the set of solutions satisfying the additionally specified linear constraints. We generally relax the problem so that assignments may be fractional, rather than the 0-1 solutions required in online matching. (For the types of problems we are most interested in, the number of left-hand nodes is huge– on the order of billions or more—therefore converting from a fractional solution to a probabilistically served one has negligible impact.) We do not assume the objective function to be linear, in fact one motivation comes from the work of Ghosh et al [4]. There the authors argue that publishers should optimize for fair, or maximally representative allocations, to maximize long term revenue. Due to competing contracts, such allocations may be impossible, in which case the objective is to get as close as possible (e.g. under an ℓ_2^2 norm) to the fairest allocation.

The key difference in the online assignment with forecast problem is that we are given a *forecast*. Prior to the online portion of the algorithm, we are given a *forecast graph*, which is meant to represent the true underlying graph. (The results we show in this paper will focus on the cases when (1) the forecast graph is precisely the true underlying graph, and (2) the forecast graph is generated by sampling nodes from the underlying graph.) We may preprocess as we like, but we are only allowed to store a small amount of information; in our case, we keep just O(1) additional numbers per right-hand node. We call this additional information the *allocation plan*. The online algorithm then proceeds as before; however, the algorithm now has access to the allocation plan in order to make better decisions.

2.2 Online allocation with forecast

Let $G = (I \cup J, E)$ be a bipartite graph. We call the nodes of I *supply nodes* and the nodes of J *demand nodes*. Our job will be to find an assignment $x \in [0,1]^E$ satisfying both the *demand constraints* and the *supply constraints* (expressed below). For an edge (i, j) we think of x_{ij} as the percentage of supply from node i going to node j. In this problem, supply nodes represent Internet website visits, while demand nodes represent advertisers. So G is extremely unbalanced, with $|J|$ in the tens of thousands and $|I|$ numbering in the billions or more.

In the allocation problem, every demand node $j \in J$ requests a certain amount of supply, thus generating a single demand constraint per demand node. The supply constraints are implied from the structure of graph G and the supply vector $s \in [0, \infty)^I$. In particular, we say that x is a *feasible* solution of the online allocation with forecast prob-

lem, denoted $\mathcal{P} = \langle G, s, d\rangle$, if it satisfies

$$\begin{aligned}
&\forall_j \quad \textstyle\sum_{i\in\Gamma(j)} s_i x_{ij} \geq d_j && \textit{demand constraints}\\
&\forall_i, \quad \textstyle\sum_{j\in\Gamma(i)} s_i x_{ij} \leq s_i && \textit{supply constraints}\\
&\forall_{(i,j)\in E}, \quad x_{ij} \geq 0 && \textit{non-negativity constraints}
\end{aligned}$$

where we use $\Gamma(i)$ to denote the neighborhood of i (i.e. the set of demand nodes adjacent to supply node i in graph G), and likewise for $\Gamma(j)$. We write $x \in \mathcal{P}$ if x is a feasible solution to $\mathcal{P}$, and $x \notin \mathcal{P}$ otherwise.

In the *pre-processing phase*, the algorithm is given a problem instance $\mathcal{P}' = \langle G', s', d\rangle$ and an objective function $F(s, x)$, where G' and s' are forecasts of G and s, but may or may not actually be G and s.[1] The output of the pre-processing phase is an *allocation plan*, which may be represented as a set of real numbers associated with each demand node.

In the *online phase*, the algorithm is given the nodes J, together with d and the allocation plan generated during the pre-processing phase. Nodes of $i \in I$ arrive one at a time, along with the value of s_i and $\Gamma(i)$ (the set of nodes in J that are adjacent to i). The algorithm must decide the allocation x_{ij} for every $j \in \Gamma(i)$ so that the supply and demand constraints are satisfied. Once the value of x_{ij} has been decided, it cannot be changed. The goal of the online algorithm is to produce an $x \in [0, 1]^E$ so that $F(s, x)$ is minimized (as much as possible), subject to $x \in \mathcal{P}$. Note that x_{ij} may be fractional.

Although it may not be apparent at first blush, finding the optimum in terms of x, a percentage-wise solution, is crucial for solving the problem using a sampled graph G'. Indeed, a single node for the sampled graph may have an associated supply that is orders of magnitude larger than in the original underlying graph. So while the percentage-wise optimum translates well to the original graph, an optimum expressed in terms of the absolute magnitudes would be meaningless. We will see later that this percentage-wise solution is also quite robust to forecast and sampling errors.

2.3 Online budgeted-bidders with forecast

Although the results in this paper will be stated in terms of the online allocation problem, we briefly describe the setting for the *online budgeted-bidders with forecast* problem as well. Here, a problem instance consists of bipartite graph $G = (I \cup J, E)$, a supply s for the nodes of I, a *budget* B_j for each node $j \in J$, and a *cost* c_{ij} for every $(i, j) \in E$. An allocation x is feasible iff it satisfies the following:

$$\begin{aligned}
&\forall_j \quad \textstyle\sum_{i\in\Gamma(j)} s_i c_{ij} x_{ij} \leq B_j && \textit{demand constraints}\\
&\forall_i, \quad \textstyle\sum_{j\in\Gamma(i)} s_i x_{ij} \leq s_i && \textit{supply constraints}\\
&\forall_{(i,j)\in E}, \quad x_{ij} \geq 0 && \textit{non-negativity constraints}
\end{aligned}$$

Note that in this problem as well, the number of demand constraints is quite small, one for each demand node (corresponding to advertisers), while the number of supply constraints is huge— one for each Internet visit.

[1]Throughout this paper, there are two ways of thinking about G'. From the perspective of the pre-processing algorithm, G' will be created from G by sampling some small set of vertices from I, and setting G' to be the induced subgraph on the sampled vertices of I together with the vertices of J. However, it is much cleaner mathematically to view G' as precisely G, but with $s'_i = 0$ for any vertex $i \in I$ that has not been sampled. We will take this latter view throughout.

In the pre-processing phase, the algorithm is given a forecast graph G' and forecast supply s', as well as an objective function $F(s, x)$, the true budgets B and the true costs c_{ij} for all (i, j) in the edge-set of G'. It must produce an allocation plan. In the online phase, the online algorithm has access to the allocation plan, the nodes $j \in J$, and the budgets B. Supply nodes $i \in I$ arrive online, together with s_i, $\Gamma(i)$, and c_{ij} for each $j \in \Gamma(i)$. The algorithm must decide x_{ij} for each $j \in \Gamma(i)$ in an online fashion while respecting the constraints, and attempting to minimize $F(s, x)$.

2.4 Objective function

Instead of simply finding a feasible assignment, our goal is to find one minimizing a particular objective function. Let $F(s, x) : \mathbb{R}^I \times \mathbb{R}^E \to \mathbb{R}$ be a convex function. We say that $F(s, x)$ is *well-conditioned* if $\frac{\partial^2}{\partial x_{ij}^2} F(s, x)$ exists and is strictly positive for all $(i, j) \in E$. We say $F(s, x)$ is *separable* if it can be written as $F(s, x) = \sum_{(i,j)\in E} F_{ij}(s_i, x_{ij})$ for some set of functions $\{F_{ij}\}_{(i,j)\in E}$.

Note that linear objective functions are not well-conditioned, since the second derivative in each variable is 0. When faced with a linear objective function, we first find the optimal value for the linear objective, and then add a linear constraint to guarantee the objective is optimal. For the rest of this paper, we restrict our attention to well-conditioned objective functions.

We require an additional property of F in order for sampling to be applicable. Let $F(s, x) : \mathbb{R}^I \times \mathbb{R}^E \to \mathbb{R}$ be an objective function, differentiable in each x_{ij} for $(i, j) \in E$. We say $F(s, x)$ is *scale-free* if for each $(i, j) \in E$,

$$\frac{\partial}{\partial x_{ij}} F(s, x) = s_i f_{ij}(x),$$

where each f_{ij} is independent of s. Intuitively, scale-freeness says that we would get the same (percentage-wise) solution even if all supply was uniformly scaled by the same factor. Without this, it would be surprising if a sampled version of the problem (in which we have, say, half the nodes but with twice the supply for each) were guaranteed to give a good approximately optimal solution in general, even on the original underlying problem.

DEFINITION 1. *An objective function is* well-structured *if it is convex, well-conditioned, separable, and scale-free.*

For any well-structured function, $F(s, x)$, there is a unique minimal solution subject to $x \in \mathcal{P}$, since $\mathcal{P}$ itself is a convex space. Thus, we will often refer to the optimal solution,

$$x^* = \arg\min_{x\in\mathcal{P}} F(s, x) \ .$$

Note that although linear objective functions are not technically well-structured, the techniques we describe here may be extended to linear objectives, albeit with additional computational effort.

2.5 Robustness

Finally, we address the issue of robustness. In general, we will not be able to prove that the allocation plan obtained using forecast graph and supply yields an online solution that is both optimal and feasible. Instead, we define the notion of ε-*goodness*. Given the problem $\mathcal{P} = \langle G, s, d\rangle$, let $\mathcal{P}' = \langle G, s, (1 + \varepsilon)d\rangle$. That is, $\mathcal{P}'$ is the problem $\mathcal{P}$ in which

every demand d_j has been increased to become $(1+\varepsilon)d_j$, so that $\mathcal{P}'$ is harder to satisfy. If $\mathcal{P}'$ is feasible, then we say $\mathcal{P}$ is ε-feasible, meaning intuitively that there is some slack in the requirements for $\mathcal{P}$.

If $\mathcal{P}$ is ε-feasible, then we say a solution x is *ε-good* for $\mathcal{P}$ if $x \in \mathcal{P}$, and further, x is at least as good as any feasible solution to $\mathcal{P}'$, i.e. $F(s,x) \leq \min_{x' \in \mathcal{P}'} F(s,x')$. (Recall we are looking for x to minimize F.)

3. COMPACT ALLOCATION PLAN

In this section, we show the existence of a compact allocation plan, and prove several of its key properties. The main theorem of this section holds for well-structured objective functions, using perfect forecasts. In Section 4, we will explore the effects of sampling the input graph G in producing the compact allocation plan.

One can easily represent the optimum solution x^* by describing the fractional assignment on each edge. Our main result in this section shows that the same solution has a smaller implicit representation. In particular, there is a function $\tilde{x}$, which, given the dual values *only* for the demand constraints, can reconstruct the optimum solution. Later, in Theorem 2 we will show that the reconstruction function $\tilde{x}$ remains optimal even for slightly perturbed versions of the problem.

THEOREM 1. *Let $G = (I \cup J, E)$ be a bipartite graph, and let $F(s,x) : \mathbb{R}^I \times [0,1]^E \to \mathbb{R}$ be a well-structured function. There is a continuous function, $\tilde{x}(\alpha) : \mathbb{R}^J \to \mathbb{R}^E$ with the property that for any feasible problem, $\mathcal{P} = \langle G, s, d \rangle$, there exists $\alpha^* \in \mathbb{R}^J$ such that $\tilde{x}(\alpha^*) = \arg\min_{x \in \mathcal{P}}\{F(s,x)\}$.*

Theorem 1 is somewhat surprising, since it claims, in some rough sense, that the optimal solution from $[0,1]^E$ is expressible in the space $\mathbb{R}^J$, independent of the size of I or E. (And the continuity of $\tilde{x}(\alpha)$ tells us that this is a real phenomenon, not simply an artificial packing of information into real numbers.) This compact representation is key to the utility of our solution.

The proof of the theorem stems from the Karush-Kuhn-Tucker (KKT) conditions of the optimal solution to our problem. We first phrase our problem in terms of the Lagrangian. Let α_j be the Lagrangian multiplier for the j-th demand constraint, let β_i be the Lagrangian multiplier for the i-th supply constraint, let γ_{ij} be the Lagrangian multiplier for the ij-th non-negativity constraint. The Lagrangian of our problem is then

$$F(s,x) - \sum_{j \in J} \alpha_j \Big(\sum_{i \in \Gamma(j)} s_i x_{ij} - d_j \Big) + \sum_{i \in I} \beta_i \Big(\sum_{j \in \Gamma(i)} s_i x_{ij} - s_i \Big) - \sum_{(i,j) \in E} \gamma_{ij} s_i x_{ij}$$

Let $x^* = \arg\min_{x \in \mathcal{P}}\{F(s,x)\}$. Since each of the constraints and the objective function are all continuously differentiable at $x^* \in [0,1]$, there are necessarily α_j,β_i, γ_{ij}, satisfying the KKT conditions:

$$\text{For all } (i,j) \in E, \quad s_i f_{ij}(x^*) - s_i \alpha_j + s_i \beta_i - s_i \gamma_{ij} = 0 \tag{1}$$

$$\text{For all } j, \quad \alpha_k \Big(\sum_{i \in \Gamma(j)} s_i x^*_{ij} - d_j \Big) = 0 \tag{2}$$

$$\text{For all } i, \quad \beta_i \Big(\sum_{j \in \Gamma(i)} s_i x^*_{ij} - s_i \Big) = 0 \tag{3}$$

$$\text{For all } (i,j) \in E, \quad \gamma_{ij} s_i x^*_{ij} = 0 \tag{4}$$

$$\text{For all } (i,j) \in E, \quad \alpha_k \geq 0, \beta_i \geq 0, \gamma_{ij} \geq 0 \tag{5}$$

Condition 1 is referred to as *stationarity*, Conditions 2, 3, 4 as *complementary slackness*, and Condition 5 as *dual feasibility*. Rewriting these conditions somewhat, we have the following:

$$\text{For all } (i,j) \in E, \quad (\text{such that } s_i \neq 0) \quad f_{ij}(x^*) - \alpha_j + \beta_i \geq 0 \ \text{ with equality unless } x^*_{ij} = 0. \tag{6}$$

$$\text{For all } j, \quad \alpha_j \geq 0, \ \text{with equality unless } \textstyle\sum_{i \in \Gamma(j)} s_i x^*_{ij} = d_j \tag{7}$$

$$\text{For all } i, \quad \beta_i \geq 0, \ \text{with equality unless } \textstyle\sum_{j \in \Gamma(i)} x^*_{ij} = 1 \ \text{ or } s_i = 0 \tag{8}$$

As is often the case with primal-dual methods, we can reconstruct the primal solution given only the dual solution. For each $(i,j) \in E$, let g_{ij} be the inverse of f_{ij}.

PROPOSITION 1. *Let x^*, g, α, and β_i be defined as above, and suppose $s_i > 0$. Then*

$$x^*_{ij} = \max\{0, g_{ij}(\alpha_j - \beta_i)\}$$

PROOF. The proof follows from Equation 6 of the KKT conditions above. First, suppose that $x^*_{ij} > 0$. Then we see immediately that $x^*_{ij} = g_{ij}(\alpha_j - \beta_i)$. On the other hand, suppose $x^*_{ij} = 0$, but $g_{ij}(\alpha_j - \beta_i) > 0 = x^*_{ij}$. Since $F(s,x)$ is convex in x, we see that f_{ij} is increasing. Thus, $\alpha_j - \beta_i > f_{ij}(x^*_{ij})$, which implies that $f_{ij}(x^*) - \alpha_j + \beta_i < 0$, a contradiction. Thus, $g_{ij}(\alpha_j - \beta_i) \leq 0$, and the proof follows. □

For convenience, define $\hat{g}_{ij}(z) = \max\{0, g_{ij}(z)\}$. From the above, we see $x^*_{ij} = \hat{g}_{ij}(\alpha_j - \beta_i)$ for all i such that $s_i > 0$. In the case that $s_i = 0$, we may simply set $x^*_{ij} = \hat{g}_{ij}(\alpha_j - \beta_i)$. Note that when $s_i = 0$, any relative allocation results in a 0 allocation, so this is fine to do. It also allows us to "interpolate" when the supply is 0, in a natural way. However, we still need to produce a function purely of the α values. The following key insight allows us to do just that.

LEMMA 1. *Let $\hat{g}$, α_j, and β_i be defined as above, and suppose $s_i > 0$. If $\sum_{j \in \Gamma(i)} \hat{g}_{ij}(\alpha_j) < 1$, then $\beta_i = 0$. If not, then β_i is the unique value satisfying the following equality:*

$$\textstyle\sum_{j \in \Gamma(i)} \hat{g}_{ij}(\alpha_j - \beta_i) = 1 \ .$$

PROOF. We use Equation 8 of the KKT conditions above. First, consider $\sum_{j \in \Gamma(i)} \hat{g}_{ij}(\alpha_j) < 1$. Then, if $\beta_i > 0$, we would have

$$\sum_{j \in \Gamma(i)} \tilde{x}_{ij} = \sum_{j \in \Gamma(i)} \hat{g}_{ij}(\alpha_j - \beta_i) < 1 \ ,$$

a contradiction. On the other hand, if $\sum_{j\in\Gamma(i)} \hat{g}_{ij}(\alpha_j-\beta_i) \neq 1$ and $\beta_i > 0$, we again have a contradiction. Thus, the claim follows.

The uniqueness follows from the fact that $g_{ij}(z)$ is strictly increasing for all $(i,j) \in E$. Thus, $\sum_{j\in\Gamma(i)} \max\{0, g_{ij}(\alpha_j - z)\}$ is strictly decreasing in z unless the sum is 0. So there is a unique z making the sum equal to 1. □

PROOF OF THEOREM 1. Even in the case that $s_i = 0$, we may still define β_i as in Lemma 1; note that this is consistent with the KKT conditions. We then see that β is really a [continuous] function of α. Thus, we will occasionally write $\hat{\beta}(\alpha)$ to denote this function. Given this, we define

$$\tilde{x}(\alpha) = \max\{0, g_{ij}(\alpha_j - \hat{\beta}(\alpha))\} .$$

Defining $\tilde{x}$ as above and setting α^* to α satisfying the KKT conditions proves the theorem. □

We will refer to the setting of $\tilde{x}$ and α^* as the *compact allocation plan*. In Section 5 we fully specify the details of the reconstruction algorithm.

3.1 Key property: Robustness of $\tilde{x}$

In this subsection, we describe a key property of the function $\tilde{x}$, which will allow us to extend our results to forecast graphs G' that are different than the true underlying graph G. At a high level, it says the following: Suppose that we find the optimum for a graph G with supply s and demand d, i.e. for $\langle G, s, d\rangle$. Then if we perturb s a little to become s', there is a way to "tweak" d by a little to get d' so that the optimum for $\langle G, s', d'\rangle$ is *precisely the same!* In other words, if our forecast supply is a little bit off, and we had the omniscience to adjust the demand a little, then the solution we obtain would be the same as the optimum solution for the realized problem. (Note that this is not the same as saying that the optimum for $\langle G, s', d\rangle$ is the same as the optimum for $\langle G, s, d'\rangle$, where d' is close to d whenever s' is close to s. This is a much more technically difficult proof, and is the subject of Section 4.)

In fact, this robustness property extends to interpolation as well, an essential aspect for sampling. In particular, suppose that $s'_i = 0$ for some i, while $s_i > 0$. (In a sampling scenario, this will happen for every node that is not included in the sample.) The function $\tilde{x}$ is still well defined, even at such points. The theorem below, in some sense, says that $\tilde{x}$ correctly interpolates, even at these zero-supply points.

THEOREM 2. *Let $\mathcal{P}, \tilde{x}$ and α be as in Theorem 1, and $s' \in \mathbb{R}^I$ be any supply such that $s'_i \geq 0$ for all $i \in I$. Let $x^* = \tilde{x}(\alpha^*)$, and let d' be an adjusted demand such that for all j, $d'_j \geq \sum_{i\in\Gamma(j)} s'_i x^*$, with equality whenever $d_j = \sum_{i\in\Gamma(j)} s_i x^*$. Then x^* is also an optimal solution of $F(s', x)$ for $x \in \mathcal{P}'$, with $\mathcal{P}' = \langle G, s', d'\rangle$.*

PROOF. To prove the theorem, we will show that the α, β used to satisfy the KKT conditions for F and $\mathcal{P} = \langle G, s, d\rangle$ also satisfy the KKT conditions for F and $\mathcal{P}' = \langle G, s', d'\rangle$. Since there are multiple choices for β_i and x^*_{ij} when $s_i = 0$, we use the values described above. Specifically, $\beta = \hat{\beta}(\alpha^*)$ and $x^* = \tilde{x}(\alpha^*)$. This is the key to interpolating to unseen values. Note that by construction, x^* is feasible for both problems.

For each $(i,j) \in E$, set $\gamma_{ij} = f_{ij}(x^*) - \alpha_j + \beta_i$. Since $x^* = \max\{0, g_{ij}(\alpha_j - \beta_i)\}$, we see that $\gamma_{ij} \geq 0$. Thus, dual feasibility holds— all $\alpha_j, \beta_i, \gamma_{ij} \geq 0$. Furthermore, the stationarity condition holds:

$$s'_i f_{ij}(x^*) - s'_i\alpha_j + s'_i\beta_i - s'_i\gamma_{ij} = 0 .$$

By our choice of β_i when $s_i = 0$, we see that complementary slackness holds for the β_i. Similarly, if $x^*_{ij} > 0$, then $x^*_{ij} = g_{ij}(\alpha_j - \beta_i)$. Thus, by our choice of γ_{ij}, we have $\gamma_{ij} = f_{ij}(x^*) - \alpha_j + \beta_i = 0$. So we only need to show that complementary slackness holds for the α_j.

By the condition on d'_j, we have $d'_j = \sum_{i\in\Gamma(j)} s'_i x^*_{ij}$ whenever $d_j = \sum_{i\in\Gamma(j)} s_i x^*_{ij}$. Thus, in this case, we see

$$\alpha_j\Big(\sum_{i\in\Gamma(j)} s'_i x^*_{ij} - d_j\Big) = 0.$$

However, when $d_j > \sum_{i\in\Gamma(j)} s_i x^*_{ij}$, we see $\alpha_j = 0$ (by complementary slackness for the original problem $\mathcal{P}$). So again, complementary slackness holds. Since F is well-structured, this shows that x^* is optimal for $\mathcal{P}'$. □

Note that Theorem 2 also suggests that our allocation plan is robust to small errors in the forecast. Suppose, for example, that every supply s_i was replaced by a new supply s'_i that was approximately s_i, say $s_i/(1+\varepsilon) \leq s'_i \leq s_i(1+\varepsilon)$. Then by tweaking each d_j by at most $(1+\varepsilon)$ factor to become d'_j, we see that the optimal solution to $\langle G, s', d'\rangle$ is the same as the optimal solution to $\langle G, s, d\rangle$. Of course, we are most interested in proving that a small tweak to s' results in only a minor change to the delivery. This will be implied by Theorem 3.

4. USING A SAMPLE OF $\mathcal{P}$

So far we have shown how to construct a compact allocation plan in the case where the whole input is known at the beginning. But we are motivated to solve the problem when the input is revealed in an online fashion, one vertex at a time. Fortunately, in real-world scenarios, something *is* known about the expected input. In many Internet applications, for example, historical logs serve as a very reliable indicator of future behavior, and can be sampled directly. More sophisticated applications may "project" these logs into the future, essentially sampling from these logs with appropriate re-weighting. For this paper, we will assume that we have access to such a black box that allows us to sample users arriving in the future. Note that unlike scenario sampling used in stochastic optimization literature and the Sample-Average-Approximation method (SAA, [7, 1]), we are sampling *individual* users, and not complete scenarios of future input.

An orthogonal motivation for sampling is the potential size of the problem. Even if we know all of the parameters to $\mathcal{P}$, finding an optimal $x \in \mathcal{P}$ may be unrealistic in a reasonable amount of time. In this section we use the machinery developed in the previous section to quantify the effects of sampling on the optimality of the final solution.

Our key insight stems from Theorem 2. It allows us to work in the dual space of $\alpha \in \mathbb{R}^J$, rather than the actual solution space $\mathbb{R}^E$. One of the most important consequences of this is that it gives us the ability to correctly interpolate the allocation on the unseen supply.

Given "tweaked" supply s', Theorem 2 spells out the par-

ticular conditions on the "tweaked" demand d', namely that

$$\sum_{i \in \Gamma(j)} s'_i \tilde{x}_{ij}(\alpha^*) \geq d'_j \ ,$$

with equality whenever $\sum_{i \in \Gamma(j)} s_i \tilde{x}_{ij}(\alpha^*) = d_j$. In general, this could mean that d' and d are quite different. In this section, we show that if we sample properly then this is not the case: d' and d are close with high probability.

Our results depend on the sensitivity of F to small changes in the input. Intuitively, the higher the effect of small changes in the input on the value of F, the more samples we will need to draw to ensure the desired result. Formally, we define the *stretch of* F, $\mathcal{P}$ for a well-structured F. Let $\sigma_j = \sum_{i \in \Gamma(j)} s_i$. Define

$$\textsc{Str}(F, \mathcal{P}) = \frac{\alpha_{\max}}{\Delta}$$

$$\text{with } \Delta = \varepsilon \min_{i,j,x \in [0,1]^E} \{\frac{\partial}{\partial x_{ij}} f_{ij}(x)\} \min_j \{d_j/\sigma_j\}/2$$

where $\alpha_{\max}$ is an upper bound on the maximum value of α_j. We will provide a bound for $\alpha_{\max}$ that depends only on the graph G and objective function. Here, Δ is chosen so that varying α by at most Δ changes the value of $\tilde{x}(\alpha)$ by at most $\varepsilon d_j/\sigma_j$. We see when $F(s, x)$ is well-structured, the stretch is well-defined.

4.1 Sampling algorithm

We now describe our sampling method, a type of importance sampling that it similar to Karp-Luby sampling. Let $\mathcal{P} = \langle G, s, d \rangle$ be as above, and let F be a well-structured objective function. For any $\delta > 0$, create set $\hat{I}$ as follows:

- For each $j \in J$, choose $m_j = \frac{9|J|}{\varepsilon^2} \frac{\sigma_j}{d_j} \ln(3|J|\textsc{Str}(F, \mathcal{P})/\delta)$ supply nodes from I independently and with replacement, where we choose supply node i with probability $1/\sigma_j$.

Define $\hat{s}$ by $\hat{s}_i = (\sum_{j \in J} m_j/\sigma_j)^{-1}$ for each $i \in \hat{I}$, with $\hat{s}_i = 0$ for $i \notin \hat{I}$, and $\hat{d}$ by $\hat{d}_j = d_j + 4\varepsilon d_j$ for each $j \in J$. Let $\widehat{\mathcal{P}} = \langle G, \hat{s}, \hat{d} \rangle$ be the problem on sampled input. Notice that $\widehat{\mathcal{P}}$ (and its components) are random variables.

THEOREM 3. *Given problem instance $\mathcal{P} = \langle G, s, d \rangle$ that is 8ε-feasible, let $\widehat{\mathcal{P}}$ be obtained by sampling, as described above. Let $F(s, x)$ be a well-structured objective function. Further, let $\tilde{x}$ be the continuous function guaranteed by Theorem 2 (which is the same for both $\mathcal{P}$ and $\widehat{\mathcal{P}}$).*

Then with probability $1 - \delta$ (over the choice of $\widehat{\mathcal{P}}$), there is an $\alpha \in \mathbb{R}^J$ so that $\tilde{x}(\alpha)$ is the optimal solution for $\widehat{\mathcal{P}}$ and $\tilde{x}(\alpha)$ is an 8ε-good solution to $\mathcal{P}$, under F.

At first, it might appear that this proof follows easily. In fact, given a *fixed* allocation x, we can show that with high probability, over the choice of $\hat{s}$, that $\sum_{i \in \Gamma(j)} \hat{s}_i x_{ij} \approx \sum_{i \in \Gamma(j)} s_i x_{ij}$. In effect, this shows that an optimal solution in the unsampled space (using s) is also an optimal solution in the sampled space (using $\hat{s}$), with only minor tweaking of the demands. However, the solution we obtain in the sampled space is itself implicitly a function of $\hat{s}$, and so the simpler argument does not apply.

We might hope that the optimal solution in the sampled space and the optimal solution in the unsampled space are close (or perhaps that their respective α values are close). However, this turns out to be false in general. Indeed, consider optimizing a function subject to any set of linear constraints. Perturbing these constraints, even by a small amount, can result in large changes in an optimal solution (although the value of the objective function may change by only a very small amount).

Essentially, we would like to show that for whatever solution arises in the sampled space, say x', we have that $\sum_{i \in \Gamma(j)} s_i x'_{ij} \geq d_j$. To do this, we cover the dual space of $\alpha \in \mathbb{R}^J$ with points spaced just Δ apart, with Δ set as above; we argue that varying α by at most Δ in each coordinate results in only a small variation in $\tilde{x}(\alpha)$. Denote this set of points as A.

We further bound the maximum value α_j may take for any j, a value we denote $\alpha_{\max}$. Since α_j is bounded above, this also yields an upper bound on $|A|$, namely $(\alpha_{\max}/\Delta)^{|J|}$, i.e. $(\textsc{Str}(F, \mathcal{P}))^{|J|}$. Thus, we may apply a union bound (on all $(\textsc{Str}(F, \mathcal{P}))^{|J|}$ points in A) to guarantee that $\sum_{i \in \Gamma(j)} \hat{s}_i \tilde{x}_{ij}(\alpha) \approx \sum_{i \in \Gamma(j)} s_i \tilde{x}_{ij}(\alpha)$ for all such points $\alpha \in A$. Since α^*, the actual solution found in the sampled space, must be within Δ (in each coordinate) of some $\alpha' \in A$, it will follow by our choice of Δ and the fact that $\tilde{x}(\alpha^*) \approx \tilde{x}(\alpha')$, that $\sum_{i \in \Gamma(j)} \hat{s}_i \tilde{x}_{ij}(\alpha) \geq d_j$.

To complete the proof, we appeal to Theorem 1, allowing us to show that the solution obtained in the sampled space is 8ε-good in the original space.

Thus, our key technical tools in this proof are (1) working in the dual space over $\mathbb{R}^J$, rather than the original allocation space over $\mathbb{R}^E$, (2) analyzing the sensitivity of the function $\tilde{x}(\alpha)$, and (3) showing that for a fixed allocation, sampling does not perturb the demand constraints by much.

We begin by showing that small changes in α do not affect $\tilde{x}(\alpha)$ too much.

LEMMA 2. *Fix $\varepsilon > 0$. Let $F(s, x)$ be a well-structured function, let $\mathcal{P} = \langle G, s, d \rangle$ be a problem instance as above, and define the function $\tilde{x}$ as in Theorem 1. (Note that this $\tilde{x}$ is identical to the one obtained using $\widehat{\mathcal{P}} = \langle G, \hat{s}, \hat{d} \rangle$ since they use the same graph G.) Set Δ as above. Let $\alpha, \alpha' \in \mathbb{R}^J$, and suppose that for all $j \in J$, $\alpha_j \leq \alpha'_j \leq \alpha_j + \Delta$. Then we have for all $(i, j) \in E$,*

$$|\tilde{x}_{ij}(\alpha) - \tilde{x}_{ij}(\alpha')| \leq \varepsilon d_j/\sigma_j$$

PROOF IDEA. We begin by bounding β_i (which is implicitly a function of α) using Lemma 1. The rest of the proof then follows using calculus and the fact that $\tilde{x}_{ij}$ can be written in terms of α_j, β_i, and the inverse of f_{ij} (again, by Lemma 1). □

Define Δ as in the previous lemma, and define $A \subseteq \mathbb{R}^J$ as follows:

$$A = \{(\Delta n_1, ..., \Delta n_{|J|}) \ : \\ \text{For all } j \in J, \ 0 \leq n_j \leq \textsc{Str}(F, \mathcal{P}), \text{ with } n_j \in \mathbb{Z}.\}$$

Notice that $|A| \leq \textsc{Str}(F, \mathcal{P})^{|J|}$. However, we still need to argue that $\alpha_{\max}$ actually exists (and that it is dependent on G, F, but independent of s, d.) The following lemma proves exactly that.

LEMMA 3. *Fix $\mathcal{P} = \langle G, s, d \rangle$, and let $F(s, x)$ be a well-structured function. Finally, let $\tilde{x}$ be the continuous function*

guaranteed in Theorem 1. Then for any s, d, there is an α^* such that $\tilde{x}(\alpha^*)$ is the optimal solution of $\mathcal{P}$, and for all j, we have that $0 \leq \alpha_j^* \leq 2|J|f_{\max}$, where $f_{\max}$ is defined as $\max_{i,j,z\in[0,1]}\{f_{ij}(z)\}$.

Furthermore, given any α satisfying the KKT conditions of the problem, there is a polynomial-time algorithm that calculates such an α^*.

PROOF IDEA. The proof follows from repeated use of the fact that

$$f_{ij}(x^*) - \alpha_j + \beta_i \geq 0 \text{ with equality unless } x_{ij}^* = 0.$$

We are most interested in edges for which $x_{ij}^* \neq 0$, since the above inequality is tight there. Consider starting at a supply node i for which $\beta_i = 0$ and finding a path through G walking along only edges for which $x_{ij}^* \neq 0$. Then we may immediately bound the dual value for every vertex we reach, using the above equality. In fact, we see that every step we take on this path, we have increased the upper bound by at most $f_{\max}$. (Notice that the equality implies $\alpha_j \leq \beta_i + f_{\max}$ as well as $\beta_i \leq \alpha_j + f_{\max}$ whenever $x_{ij}^* \neq 0$.) Since any simple path in this bipartite graph has length at most $2|J|$ (recall $|J| < |I|$), this gives us the upper bound of the lemma.

But what if there is no i for which $\beta_i = 0$, or at least none in a given component (when only considering the edges of G for which $x_{ij}^* \neq 0$)? In this case, we argue that we may shift both the values of β and α uniformly by the same amount so that we still maintain the same primal solution. The algorithm for finding the proper α is a direct extension of the full proof. □

For the budgeted bidders problem, we may bound $\alpha_{\max}$ by $f_{\max}$ divided by the smallest query cost (but always divided by at least 1). In the most general setting, $\alpha_{\max}$ is bounded by $|E|\lambda f_{\max}$, where $|E|$ is the total number of edges in G, and λ is the ratio of the largest minor (in absolute value) of the full constraint matrix, divided by the smallest (in absolute value) non-zero minor. Note that none of these quantities involve the supply or the demand.

There is one subtlety in Lemma 3. Although the optimal primal solution is unique, the optimal dual solution in general is not. However, given any optimal dual solution, there is a simple algorithm to find an optimal dual solution that falls within the bounds of $\alpha_{\max}$.

Finally, we wish to show that for all $j \in J$ and for all $\alpha \in A$, that $\sum_{i\in\Gamma(j)} s_i\tilde{x}_{ij}(\alpha) \approx \sum_{i\in\Gamma(j)} \hat{s}_i\tilde{x}_{ij}(\alpha)$, unless the sum is bounded far below or far above d_j. Unfortunately, we will not be able to apply standard Chernoff-Hoeffding bounds. Instead, we need the somewhat stronger Bernstein Inequalities, which allows us to show that the probability of being far from the mean drops exponentially with the inverse of the variance. This allows us to show that for any fixed $\alpha \in A$, the probability of a large error is exponentially small. Thus, we may apply a union bound over all $\alpha \in A$, yielding a good result for all $\alpha \in A$ with high probability.

LEMMA 4. *Let $\mathcal{P}$, $\widehat{\mathcal{P}}$, F, and A be defined as above. With probability at least $1-\delta$, the following holds: For all $\alpha \in A$,*

- *If $\sum_{i\in\Gamma(j)} s_i x_{ij}(\alpha) < d_j/2$, then $\sum_{i\in\Gamma(j)} \hat{s}_i x_{ij}(\alpha) < 3/4d_j$.*
- *If $\sum_{i\in\Gamma(j)} s_i x_{ij}(\alpha) > 2d_j$, then $\sum_{i\in\Gamma(j)} \hat{s}_i x_{ij}(\alpha) > 3/2d_j$.*
- *Otherwise, $|\sum_{i\in\Gamma(j)} \hat{s}_i x_{ij}(\alpha) - \sum_{i\in\Gamma(j)} s_i x_{ij}(\alpha)| < \varepsilon d_j$.*

Furthermore, $\sum_{i\in\Gamma(j)} \hat{s}_i \leq 2\sigma_j$ for all j. (Recall that $\sigma_j = \sum_{i\in\Gamma(j)} s_i$.)

PROOF IDEA. It is straightforward to see that the expected value of $\sum_{i\in\Gamma(j)} \hat{s}_i x_{ij}(\alpha)$ is precisely $\sum_{i\in\Gamma(j)} s_i x_{ij}(\alpha)$. With some algebra and the correct inequalities, we may also bound the variance of the quantity. The last claim of the lemma follows easily using this analysis, but the three bullets take a little more work.

The most straightforward case is the third bullet. Here, an application of the Bernstein Inequalities yields the proper results, since the variance and error are in balance with each other. For the first bullet, the variance in relation to the expected value is actually too high, and it may not be the case that $\sum_{i\in\Gamma(j)} \hat{s}_i x_{ij}(\alpha) \approx \sum_{i\in\Gamma(j)} s_i x_{ij}(\alpha)$. However, the probability that the left-hand side is very large is still quite small, which is all that is needed. For the second bullet, the variance is good in relation to the expected value, but too high with respect to d_j, the quantity we really care about. So, although $\sum_{i\in\Gamma(j)} \hat{s}_i x_{ij}(\alpha)$ and $\sum_{i\in\Gamma(j)} s_i x_{ij}(\alpha)$ are not within εd_j with high probability, the first quantity will still be much greater than d_j; again, this is all we need.

Finally, we apply a union bound to all $\text{STR}(F, \mathcal{P})^{|J|}$ points in A to obtain the claim. Notice that in order for our bounds to be tight enough, this requires an extra factor of $\lg(\text{STR}(F, \mathcal{P})^{|J|})$ number of samples. □

We are now ready to complete the proof of the main theorem in this section.

PROOF OF THEOREM 3. With probability $1-\delta$, $\widehat{\mathcal{P}}$ satisfies the conditions of Lemma 4. So we need only consider such $\widehat{\mathcal{P}}$. Let α^* be the optimal α guaranteed by Lemma 3. Let $\alpha \in A$ be such that $\alpha_j \leq \alpha^* \leq \alpha_j + \Delta$; by Lemma 3 and the definition of A, such an α exists. By Lemma 2, $|\tilde{x}_{ij}(\alpha) - \tilde{x}_{ij}(\alpha^*)| < \varepsilon d_j/\sigma_j$ for all $(i,j) \in E$. In particular, this means that

$$|\sum_{i\in\Gamma(j)} s_i\tilde{x}_{ij}(\alpha) - \sum_{i\in\Gamma(j)} s_i\tilde{x}_{ij}(\alpha^*)| \leq \sum_{i\in\Gamma(j)} s_i(\varepsilon\frac{d_j}{\sigma_j}) = \varepsilon d_j, \text{ and}$$

$$|\sum_{i\in\Gamma(j)} \hat{s}_i\tilde{x}_{ij}(\alpha) - \sum_{i\in\Gamma(j)} \hat{s}_i\tilde{x}_{ij}(\alpha^*)| \leq \sum_{i\in\Gamma(j)} \hat{s}_i(\varepsilon\frac{d_j}{\sigma_j}) = 2\varepsilon d_j.$$

Since $\widehat{\mathcal{P}}$ satisfies the conditions of Lemma 4, we have three cases to consider.

1. If $\sum_{i\in\Gamma(j)} s_i\tilde{x}_{ij}(\alpha) < d_j/2$, then $\sum_{i\in\Gamma(j)} \hat{s}_i\tilde{x}_{ij}(\alpha) \leq 3/4d_j$. Hence, $\sum_{i\in\Gamma(j)} \hat{s}_i\tilde{x}_{ij}(\alpha^*) \leq 3/4d_j + 2\varepsilon d_j < d_j$, a contradiction.
2. If $\sum_{i\in\Gamma(j)} s_i\tilde{x}_{ij}(\alpha) > 2d_j$, then

$$\sum_{i\in\Gamma(j)} s_i\tilde{x}_{ij}(\alpha^*) > 2d_j - \varepsilon d_j \geq d_j \ .$$

With an eye towards the second part of the proof, also notice that

$$\sum_{i\in\Gamma(j)} \hat{s}_i x_{ij}(\alpha^*) \geq \sum_{i\in\Gamma(j)} \hat{s}_i x_{ij}(\alpha) - 2\varepsilon d_j > 3/2d_j - 2\varepsilon d_j > d_j(1+4\varepsilon).$$

That is, the j-th demand constraint is not tight for the sampled problem.

3. In the last case, we have that

$$|\sum_{i\in\Gamma(j)} \hat{s}_i\tilde{x}_{ij}(\alpha) - \sum_{i\in\Gamma(j)} s_i\tilde{x}_{ij}(\alpha)| \le \varepsilon d_j$$

Hence, we see

$$\begin{aligned}
&|\sum_{i\in\Gamma(j)} \hat{s}_i\tilde{x}_{ij}(\alpha^*) - \sum_{i\in\Gamma(j)} s_i\tilde{x}_{ij}(\alpha^*)| \\
&\le |\sum_{i\in\Gamma(j)} \hat{s}_i\tilde{x}_{ij}(\alpha^*) - \sum_{i\in\Gamma(j)} \hat{s}_i\tilde{x}_{ij}(\alpha)| \\
&\quad + |\sum_{i\in\Gamma(j)} \hat{s}_i\tilde{x}_{ij}(\alpha) - \sum_{i\in\Gamma(j)} s_i\tilde{x}_{ij}(\alpha)| \\
&\quad + |\sum_{i\in\Gamma(j)} s_i\tilde{x}_{ij}(\alpha) - \sum_{i\in\Gamma(j)} s_i\tilde{x}_{ij}(\alpha^*)| \\
&\le 2\varepsilon d_j + \varepsilon d_j + \varepsilon d_j = 4\varepsilon d_j
\end{aligned}$$

But $\sum_{i\in\Gamma(j)} \hat{s}_i\tilde{x}_{ij}(\alpha^*) \ge d_j(1+4\varepsilon)$ since α^* was chosen to satisfy the demand constraints with $d'_j = d_j(1+4\varepsilon)$. Hence,

$$\sum_{i\in\Gamma(j)} s_i\tilde{x}_{ij}(\alpha^*) \ge d_j \ .$$

Again, looking towards the second part of the proof, we also note that if the j-th demand constraint is tight in the sampled problem (i.e. $\sum_{i\in\Gamma(j)} \hat{s}_i\tilde{x}_{ij}(\alpha^*) = d_j(1+4\varepsilon)$), then $\sum_{i\in\Gamma(j)} s_i\tilde{x}_{ij}(\alpha^*) \le d_j(1+8\varepsilon)$.

In every viable case, $\tilde{x}(\alpha^*)$ satisfies the demand constraints on the true underlying problem instance $\mathcal{P}$, as we wanted.

To finish the proof, we appeal to Theorem 1. Consider the following problem instance $\mathcal{P}' = \langle G, s, d'\rangle$, where G and s are in the underlying instance, and d' is defined by

$$d'_j = \sum_{i\in\Gamma(j)} s_i\tilde{x}_{ij}(\alpha^*) \qquad \text{if } \sum_{i\in\Gamma(j)} \hat{s}_i\tilde{x}_{ij}(\alpha^*) = d_j(1+4\varepsilon)$$

$$d'_j = d_j \qquad \text{otherwise}$$

(Notice the first case corresponds to the demand constraint being tight in the sampled problem.) By Theorem 1, we have that $\tilde{x}(\alpha^*)$ is in fact the optimal solution for $\mathcal{P}'$. Further, by the above case analysis, we see that $d_j \le d'_j \le d_j(1+8\varepsilon)$. Hence, $\tilde{x}(\alpha^*)$ is 8ε-good. (Specifically, $\tilde{x}(\alpha^*)$ is a feasible solution for $\mathcal{P}'' = \langle G, s, (1+8\varepsilon)d\rangle$. Since it is an optimum for a space containing $\mathcal{P}''$, it must be at least as good as the optimum for $\mathcal{P}''$.) □

We remark that the result of Theorem 3 assumes we sample from the true supply s. In general, the techniques we use here extend to handle sampling from both a biased supply — in which the true supply may vary by as much as, say a $(1+\varepsilon)$ factor from the forecasted supply — and from a noisy supply — in which the true supply is a random variable where only the mean is known. However, we do not prove any formal statements in this extended abstract.

5. OVERALL ALGORITHM

In this section we put together the results of previous theorems to obtain a simple algorithm for the linear assignment problem. Given a problem $\mathcal{P}$, Theorem 3 shows that only a limited sample of the supply nodes $I' \subseteq I$ is necessary to compute a near optimal allocation plan $\tilde{x}(\alpha)$. In this section we show how to use this plan in an online fashion:

Given problem instance $\mathcal{P} = \langle G, s, d\rangle$, and well-structured objective function $F(s, x)$:

1. In the pre-processing phase, first create the sampled problem, $\widehat{\mathcal{P}}$, as described in Section 4. Then find $\alpha^* \in \mathbb{R}^J$ such that $\tilde{x}(\alpha^*) = \arg\min_{x\in\widehat{\mathcal{P}}}\{F(s,x)\}$ and $\alpha^*_j \le \alpha_{\max}$ for all $j \in J$.
2. In the online phase, as each $i \in I$ arrives, together with s_i, $\Gamma(i)$:
   - If $\sum_{j\in\Gamma(i)} \hat{g}_{ij}(\alpha_j) \le 1$, then set $\beta_i = 0$. Otherwise, find β_i such that $\sum_{j\in\Gamma(i)} \hat{g}_{ij}(\alpha_j - \beta_i) = 1$.
   - Allocate $\tilde{x}_{ij}(\alpha) = \hat{g}_{ij}(\alpha_j - \beta_i)$ for each $j \in \Gamma(i)$.

When the objective function is quadratic in the x_{ij} values, we can solve for β_i exactly. Unfortunately, we will not be able to produce the exact β_i in general. Instead, we approximate β_i within an additive Δ using the bisection method, where Δ is the same as in the proof of Theorem 3. Notice that using such a Δ guarantees that our reconstructed solution is at most another $\varepsilon \min_k\{t_k\}$ away in each demand constraint. Further, the proof bounding the α values also bounds the β values, showing that $\beta_i \le \alpha_{\max}$ for each $i \in I$. Thus, the bisection search method takes as most $O(\lg(\alpha_{\max}/\Delta)) = O(\lg \textsc{Str}(F, \mathcal{P}))$ iterations.

We may now state the main theorem of this section.

THEOREM 4. *There is an online algorithm whose starting input consists of a set of demand nodes, a well-structured objective function, F, and a compact plan, denoted α^*, and whose online input consists of a sequence of requests, each request being a supply node i, along with its supply s_i, and its neighborhood $\Gamma(i)$ (a subset of the demand nodes), and whose online output is an allocation from i to each of its adjacent demand nodes j (denoted y_{ij}); for any $\varepsilon > 0$, this algorithm has the following properties:*

- *For $\mathcal{P} = \langle G, s, d\rangle$, create $\widehat{\mathcal{P}}$ by sampling as described in Section 4, and find α^* as described in step (1) above. Then the solution y output by the algorithm is 9ε-good for $\mathcal{P}$.*
- *The compact plan described above requires just $|J|$ values.*
- *The processing time per request is $O(|\Gamma(i)| \lg \textsc{Str}(F, \mathcal{P}))$.*

6. CONCLUSION

We introduce the online assignment with forecast problem, in which an Internet publisher equipped with a forecast of future user visits must compute and implement an assignment of users to contracts in an online fashion. We show that the dual space of the resulting optimization problem is useful not just for analysis, but also as a tool for creating allocation plans used to guide the online algorithm. By further moving to a subspace of the dual space, we not only create an incredibly compact allocation plan, we also produce a robust online algorithm that generalizes to unseen input. In practice, this result has an important consequence: the ability to solve a problem (nearly) optimally and actually serve it in a distributed, essentially stateless fashion, with a lightweight plan and no communication between servers.

More generally, this framework allows us to prove that an allocation plan produced over even a sampled problem

instance still yields near-optimal results. There are two equally valid reasons why we may chose to work with such a restricted version: (1) only some information about the overall problem is known ahead of time, or (2) the non-linear optimization problem cannot be solved on larger inputs. These insights allow us to tackle the new class of online problems we propose: online problems with forecast, a theoretical model motivated by the need to produce online algorithms that function under more realistic assumptions about online input.

Acknowledgements

Thanks to Vijay Bharadwaj for many helpful discussions, and for his insights on the generalization to Lemma 3.

7. REFERENCES

[1] Moses Charikar, Chandra Chekuri, and Martin Pál. Sampling bounds for stochastic optimization. In Chandra Chekuri, Klaus Jansen, José D. P. Rolim, and Luca Trevisan, editors, *APPROX-RANDOM*, volume 3624 of *Lecture Notes in Computer Science*, pages 257–269. Springer, 2005.

[2] Nikhil R. Devenur and Thomas P. Hayes. The adwords problem: online keyword matching with budgeted bidders under random permutations. In John Chuang, Lance Fortnow, and Pearl Pu, editors, *ACM Conference on Electronic Commerce*, pages 71–78. ACM, 2009.

[3] Jon Feldman, Aranyak Mehta, Vahab S. Mirrokni, and S. Muthukrishnan. Online stochastic matching: Beating 1-1/e. 2009.

[4] Arpita Ghosh, Preston McAfee, Kishore Papineni, and Sergei Vassilvitskii. Bidding for representative allocations for display advertising. In Stefano Leonardi, editor, *WINE*, volume 5929 of *Lecture Notes in Computer Science*, pages 208–219. Springer, 2009.

[5] R. M. Karp, U. V. Vazirani, and V. V. Vazirani. An optimal algorithm for on-line bipartite matching. In *STOC '90: Proceedings of the twenty-second annual ACM symposium on Theory of computing*, pages 352–358, New York, NY, USA, 1990. ACM.

[6] Aranyak Mehta, Amin Saberi, Umesh Vazirani, and Vijay Vazirani. Adwords and generalized online matching. *J. ACM*, 54(5):22, 2007.

[7] David B. Shmoys and Chaitanya Swamy. An approximation scheme for stochastic linear programming and its application to stochastic integer programs. *J. ACM*, 53(6):978–1012, 2006.

[8] Erik Vee, Sergei Vassilvitskii, and Jayavel Shanmugasundaram. Optimal online assignment with forecasts. Yahoo Technical Report 2009-005, 2009.