Optimal allocation design without transfers
PRELIMINARY DRAFT

Itai Ashlagi Peng Shi

March, 2014

Abstract

This paper studies the allocation of goods to agents without monetary transfers and incomplete information. Agents’ have private, multi-dimensional utilities over goods, drawn from commonly known priors, possibly asymmetric across agents’ types. Both cardinal and ordinal mechanisms are studied. For tractability, a large market approximation is considered with a continuum of agents for each type. We characterize incentive compatible mechanisms that are Pareto-optimal with respect to each type. For general objective functions, this allows reducing the mechanism design problem to a well-defined optimization problem with a few continuous variables.

We apply this framework to real data from Boston 2012-2013 school choice reform. The goal is to design an optimal ordinal mechanism to assign students to public schools to maximize a linear combination of utilitarian and max-min welfare, subject to capacity and transportation constraints. We show how to optimally solve a large market model with over 868 types of students and 77 schools, translate the solution into the finite market setup and yield a feasible solution, which significantly outperforms the baseline plan chosen by the city in terms of efficiency, equity, and predictability.

1 Introduction

In various settings goods or services should be allocated to agents without the use of monetary transfers. Seats in public schools should be allocated to children, spaces in college dorms or courses are allocated to students, internships positions are allocated to doctors, etc. Social planners’ concerns are often involved including possibly social welfare, equity and system costs.

If agents’ preferences are known, the allocation becomes merely an optimization problem. This paper studies how to allocate services when agents’ preferences are private. We study a large market model including many agents of each type and many copies of each kind of service. The model allows multiple types of agents and multiple kinds of services, with agents having private, multi-dimensional preferences over services. The distribution of preferences for each type is common knowledge (so a type in this paper does not capture the private information of an agent but rather the public information). We require mechanisms to be incentive compatible that are Pareto optimal among agents with the same type, referred here by valid mechanisms. So while agents of the same type are treated symmetrically, agents with different types may be differentiated. The goal is to find a valid mechanism that maximizes the social planner’s objective, which can be fairly general.

1 For example the type may capture a neighborhood and a prior of families preferences from that neighborhood.
Since designing multi-dimensional mechanisms is traditionally difficult especially with general objectives, we model agents within each type as a continuum. A simple characterization is given for all valid mechanisms; under mild assumptions over utility priors, any valid mechanism can be described as a collection of Competitive Equilibriums with Equal Income (CEEI) for each type. In other words, agents of each type are given “virtual prices” for probabilities to each service, and the allocation can be interpreted by allowing agents to “purchase” with one unit of “virtual money” a bundle of services that maximize her expected utility. Prices for services may vary across types, but agents with the same type observe the same prices. This characterization reduces the search for the optimal mechanism to a well defined non-linear optimization problem with the only variables being the virtual prices.

In many contexts, only relative preferences between services are elicited rather then exact utility functions. Children submit a ranking over schools in Boston and NYC school choice programs, doctors submit rankings over residencies and vice versa to the National Residency Matching Program. Such mechanisms are called ordinal as opposed to cardinal mechanisms which have no information requirements. We also study the design of ordinal mechanisms subject to incentive compatibility and a suitable definition of Pareto optimal within type.

Under mild regularity assumptions it is shown that any valid ordinal mechanism can be described as “lottery-plus-cutoff”: each agent receives a uniformly random lottery number between zero and one. For each service and each type, there is a “lottery cutoff,” and an agent is “admitted” to a service if her lottery number is below the cutoff. Each agent is allocated her most preferred service for which she is admitted. This again simplifies the search of the optimal mechanism to a well-defined optimization problem with the only variables being the lottery cutoffs.

These structural results give insights on the types of mechanisms observed in practice. In many business schools course allocation is done by a bidding process, in which students are given a number of points and the highest bidders are assigned a seat. Given equilibrium prices, this mechanism is akin to the type-specific pricing mechanism described above. In Boston, New York City, New Orleans and San Francisco, students submit a ranking of schools they prefer within their menu of options. For example, until 2013 Boston city was divided into three zones and each student can only rank schools within her zone. A centralized mechanism uses submitted ranking lists, pre-defined priorities and a random lottery number given to each student in order to determine the assignment, which is analogue to the lottery-plus-cutoff mechanism. Of course, in contrast to our model all of these markets are finite.

One of the main technical contributions in this paper is to efficiently find the optimal ordinal mechanism in an empirical relevant environment. By relying on the theoretical characterization, one can encode the optimal ordinal mechanism in an exponentially sized linear program. We show that the dual of this program can be efficiently solved given an oracle to an “optimal menu” sub-problem, which in turn can be solved efficiently when utilities are based on a logit discrete choice model.

To demonstrate the relevance of our large market model, we apply it to school choice in Boston. Seats in Boston Public Schools have historically been allocated using an ordinal mechanism; families submit a ranked list of preferred schools within an individualized menu of choices and a centralized algorithm allocates based on various priorities and lottery numbers. An important question is how to design the menus from which families will choose schools? In 2012-2013, Boston launched a reform to reduce the size of the menus (from the 3 zones), while still offering sufficient choices and equity of access to quality. One of the main costs in the education system came from from busing.
The outcome of the reform was based on a simulation analysis, which used historical choice data to fit a utility model, evaluating various proposed plans on a portfolio of metrics. The plan chosen by the city is referred here by the Baseline mechanism (the Baseline can also be viewed as a valid mechanism). The goal of the empirical exercise is to design a mechanism that will improve the Baseline mechanism in various dimensions: yield significantly higher social welfare, higher max-min welfare, and improves students’ chances of getting their top choices, while staying within the same transportation budget. All of the analysis use real data from Boston Public Schools (BPS).

Although the school choice problem is defined as a finite market problem, using our analytical results we find the optimal mechanism for the large market approximation of the finite market model, translate it back into the finite market, and get an “asymptotically optimal” solution. Once the optimal (large market) cutoffs are found, we apply the well-known Deferred Acceptance (DA) algorithm (Gale and Shapley [1962]) on the finite market, using the cutoffs to guide priorities of students. We evaluate the mechanism in the finite market setting, and show by simulation that it significantly improves upon the Baseline in all aspects. In terms of welfare, the improved mechanism effectively decreases students’ distance to schools by 0.5 miles, and for max-min utility the improvement is about 2.5 miles. This is significant since the Baseline only improves over the most Naive plan by 0.6 miles, so we effectively double the gain. Furthermore, the improved mechanism increases students’ chances of getting their first choice by .15.

Finally, more theoretical results are derived in simplified settings that suggest that the type of mechanism, whether ordinal or cardinal, may result in a significant different outcomes. We give an example with one service of limited capacity and an outside option with infinite capacity, in which the ratio between the optimal social welfare from cardinal mechanisms and the optimal social welfare from ordinal mechanisms may be arbitrarily large.

Our work connects three strands of previous research. First is the matching literature, which traditionally focuses on designing mechanisms that satisfy certain properties, such as Pareto efficiency, various fairness conditions, and strategyproofness. These models are able to handle multiple types of goods and services. Hylland and Zeckhauser [1979] study cardinal mechanisms that achieves Pareto efficiency and propose Competitive Equilibrium with Equal Incomes (CEEI). Bogomolnaia and Moulin [2001] studies ordinal mechanisms that satisfy an ordinal notion of Pareto optimality called ordinal efficiency, and proposes a mechanism called Probabilistic Serial, which Che and Kojima [2011] shows is approximately is asymptotically equivalent in the large market to the more commonly known Random Serial Dictatorship (RSD), in which agents are ordered uniformly randomly and take turns picking items. Liu and Pycia [2012] extend this result, and show that in the large market all ordinal mechanisms that are asymptotically efficient, symmetric, and asymptotically strategyproof coincide in the limit with Probabilistic Serial. Abdulkadiroğlu and Sönmez [2003], Abdulkadiroğlu et al. [2009] and Abdulkadiroğlu et al. [2006] apply matching theory to school choice, and their work have been influential in the adoption of strategyproof ordinal mechanisms over non-strategyproof alternatives in cities such as New York, Boston, Chicago, New Orleans, and San Francisco. However, the matching literature hardly assume priors on agents utilities (especially not asymmetric priors), and do not seek to optimize global objective. Our work can be

---

2 This solution is asymptotically optimal in the sense that if the market is scaled up with independent copies of itself, then the finite market model converges to the large market model and the finite market solution also converge to the large market optimum.

3 For surveys, see Roth and Sotomayor [1990], Abdulkadiroğlu and Sönmez [2010].

4 More recent work on school choice include Pathak and Sethuraman [2011], Erdil and Ergin [2008], Abdulkadiroğlu et al. [2008], and Echenique and Yenmez [2012].
viewed as bridging the matching literature and the mechanism design/auction literature (see e.g. Myerson [1981] who models every agent with a common prior, and Budish [2012] who compares matching and standard mechanism design emphasizing the absence of different priors and objectives in the matching literature). By considering prior information, allocating for example seats in a city may have a significant impact on the school choice design once concerns other than efficiency are considered.

Another strand is optimal mechanism design without money in the finite market framework. Miralles [2012] tackles the multi-dimensional case by considering the Bayesian optimal cardinal mechanism with two goods and many symmetric bidders whose utility priors are symmetric across the two goods. He shows that under certain regularity conditions, the optimal ex-interim allocation rules can be described as resulting from Competitive Equilibrium with Equal Incomes (CEEI), and this can be implemented ex-post using a combination of lotteries, insurances, and virtual auctions, in which agents use probabilities for the less desirable good as a virtual currency to bid for the more desirable good. However, even with two goods, the analysis is difficult, and still requires reducing the valuation space to a single-dimension, by taking the ratio of each agent’s utilities for the two goods, so it is not “truly” multi-dimensional and does not generalize to more than two goods. Hartline and Roughgarden [2008], Condorelli [2012] and Chakravarty and Kaplan [2013] study models in which agents cannot pay money but may “burn money” or exert costly effort to signal their valuation. Similar to our work, the social planner has priors on agents’ valuations, but their analyses only allow single-dimensional valuations. One insight from their work is that if the tail of the utility prior is not too thick, or more precisely, if the prior satisfies the commonly assumed Monotone Hazard Rate condition, then requiring agents to exert costly effort is unnecessary and a lottery maximizes social welfare. However, their work cannot be easily extended to multi-dimensional preferences, which is the more realistic assumption in settings such as school choice, where there are multiple types of services.

Finally, a third strand of related research is large market models with a continuum of agents. In many such models, the analysis greatly simplifies over the finite market analog, and stronger, cleaner results may be possible, while still yielding empirically relevant insights. Such models are common in Industrial Organization. Tirole [1988]. There is previous work that have the flavor of our characterization for valid cardinal mechanisms, although they do not imply our result. Aumann [1964] shows that with a continuum of agents, the game theoretic core from trading coincides with the set of competitive equilibria, although not necessarily with equal incomes. Aumann [1964] shows conditions in which any Pareto efficient allocation is supported by equilibrium prices, although not necessarily from equal incomes. His analysis crucially depends on the unboundedness of the space of allocations, which in our cases is the bounded unit simplex \( \Delta \). Zhou [1992] and Thomson and Zhou [1993] show that under certain notions of Pareto efficiency and envy-freeness, the only possible mechanisms are again CEEI. However, their analyses depend on the space of allocation being open, which in our case is not true, and their envy-freeness condition is stronger than our incentive compatibility condition. Azevedo and Leshno [2012] study matching markets with continuum of agents, but contrary to our result they do not consider a global optimization perspective. While continuum models often provide cleaner results, finding the actual mechanism may still be hard.

5 For a survey, see Schummer and Vohra [2007].
6 Extending their analyses to multi-dimensional preferences requires a breakthrough in characterizing incentive compatibility in multi-dimensional domains, for which the currently known characterization of cyclic-monotonicity is difficult to work with. (See Rochet [1987])
This paper contributes to this literature by actually being able to find the optimal mechanism in an empirically relevant context.

2 Model

A social planner needs to allocate services to a continuum of agents. There is a finite set $T$ of agent types\(^7\) and a mass $n_t$ of agents for each type $t \in T$. There is a finite set $S$ of services. Every agent must be allocated exactly one service. However, allocations might be probabilistic, so the set of possible allocations for each agent is the distribution of services,

$$\Delta = \{ \mathbf{p} \in \mathbb{R}^{|S|} : \mathbf{p} \geq 0, \sum_{s} p_s = 1 \}.$$  

For agents of type $t$, their utilities for various services are distributed according to a continuous\(^8\) measure $F_t$ over utility space $U = \mathbb{R}^{|S|}$. Each $\mathbf{u} \in U$ is a possible utility vector where each component denotes utility for a service. Note that our valuation space is multi-dimensional. For any measurable subset $A \subseteq U$, $F_t(U)$ denotes the mass of agents having utilities in $A$. Since the total mass for type $t$ is $n_t$, the total measure $F_t(U) = n_t$. The distributions $F_t$’s are common knowledge, while the exact utilities of each agent is private knowledge. The social planner must design a mechanism to truthfully elicit this information.

A mechanism $x$, is a collection of allocation rules $x_t$ for each type, where each $x_t$ is a mapping from reported utilities to a possible allocation, $x_t : U \to \Delta$. An allocation rule is incentive compatible if it is in the agent’s best interest to report the truth:

$$\mathbf{u} \cdot x_t(u) \in \arg \max_{u' \in U} \mathbf{u} \cdot x_t(u')$$

An allocation rule is Pareto efficient if the agents within this type cannot trade among themselves to improve. Precisely speaking, $x_t$ is Pareto efficient if there does not exist another function $x'_t : U \to \Delta$ such that $x'_t$ has the same average allocation,

$$\int_U x'_t(u) dF_t = \int_U x_t(u) dF_t$$

and $x'_t$ is weakly preferred by all agents

$$\mathbf{u} \cdot x'_t(u) \geq \mathbf{u} \cdot x_t(u)$$

and strictly preferred for a positive measure of agents $A \subseteq U$, such that $F_t(A) > 0$.

We call an allocation rule valid if it is both incentive compatible (IC) and Pareto efficient (PE). Requiring IC is without loss of generality by the revelation principle, as long as we assume that agents would eventually play a Bayes Nash equilibrium. We require PE as a “stability” criterion: our setup implicitly assumes that the social planner must treat each type symmetrically, without the ability to discriminate based on the identity of the agent within the type, so it may be unreasonable

---

\(^7\)In contrast to mechanism design convention, our notion of “type” does not denote the agent’s private information, but rather her public information.

\(^8\)This assumption can be relaxed, but we choose to adopt it to simplify analysis, since when the distribution is continuous we can associate agents with utilities and not worry about one utility having a positive mass of agents.
to enforce that agents of the same type cannot trade among themselves post-allocation, and so distort the original allocation. Hence, we desire that the mechanism “foresees” any such trades and incorporates the non-existence of Pareto improving trades within types as a constraint.

The set of allocation rules for all types makes up the mechanism. The social planner’s goal is to find a mechanism $x$ that maximizes his own objective function $W(x)$, subject to all the allocation rules being valid. For now, we allow the objective function $W(x)$ to arbitrarily depend on all the allocation rules $x_t$, hence allowing it to incorporate agents’ welfare, capacity constraints, differential costs in providing various services to various types of agents, and other complex considerations. As examples of how of how such measures can be represented in terms of the mechanism $x$, observe that the expected utility of an average type $t$ agent is

$$v_t = \frac{1}{n_t} \int_U u \cdot x_t(u) dF_t.$$  

So we can incorporate social welfare by including $\sum_t n_t v_t$ in the objective function. As an another example, the total amount of service $s$ allocated is

$$m_s = \sum_t \int_U x_{ts}(u) dF_t.$$  

Using this we can model a hard capacity limit quality of service $s$ on service $s$ by setting $W(x)$ to be negative infinity when $m_s > quality_s$. Alternatively, we can model a smooth penalty for exceeding capacity by including a term $C(\max\{0, m_s - quality_s\})$ in the objective where $C$ is a convex cost function.

### 2.1 Characterization of Valid Allocation Rules

We show that under mild regularity conditions on $F_t$, any incentive compatible and Pareto efficient allocation rule $x_t$ corresponds to a Competitive Equilibrium from Equal Incomes (CEEI) of an artificial “currency.” This means that there exist “prices” $a_s \in (0, \infty]$ (possibly infinite) in terms of units of probability of a service per unit of artificial currency, such that the allocation is what agents would do if they had 1 unit of artificial currency and purchased a probability bundle that maximized their expected utility:

$$x_t(u) \in \arg \max_{p \in \Delta} \{u \cdot p : a \cdot p \leq 1\}.$$  

Figure II illustrates a CEEI with 3 services.

The price vector $a$ is the same for all agents of this type, but may be different for agents of different types. This result implies that the search for the optimal mechanism can be restricted to searching over the set of price vectors for each type to optimize the induced objective. For each type, the space of price vectors is only $|S|$-dimensional, as opposed to the space of allocation rules, which is the space of all functions $g : U \rightarrow \Delta$.

Our result requires one mild regularity condition on the distributions $F_t$, which says that a-priori, an agent’s range of relative preferences, which is her normalized preference intensities, could with positive probability be anything in the unit sphere. This is used in our analysis to guarantee “trade”: for any “trading direction,” there is a positive measure of agents who would improve benefit by moving in that direction.
Figure 1: Illustration of a Competitive Equilibrium from Equal Incomes (CEEI) with \(|S| = 3\). The triangle represents the space of possible allocations \(\Delta\). The shaded region is \(\{p \in \Delta : a \cdot p \leq 1\}\). This represents the “allowable allocations” for this type, which is the convex hull of \(\{x_t(u)\}\); she could obtain allocations in the interior of this region by randomizing over several reports. For utility report \(u\), the agent receives an allocation \(p\) that maximizes expected utility \(u \cdot p\) subject to \(p\) being in the “allowable region.” This corresponds to having price vector \(a\) and budget 1, and the agent can “purchase” any allowable allocation subject to the budget constraint.

Definition 2.1. (Full relative support) Define the set of relative preferences \(U_{rel} = \{u \in U, \|u\| = 1\}\), where \(\|\cdot\|\) is the Euclidean norm. This is a sphere in \(|S|\)-dimensional space and can be endowed with the topology of a \((|S| - 1)\)-sphere. This induces a topology on the set of cones \(C \subseteq U\) by defining \(C\) as open iff \(C \cap U_{rel}\) is open in \(U_{rel}\). Distribution \(F_t\) has full relative support if for every non-empty open cone \(C \subseteq U\), \(F_t(C) > 0\).

Theorem 2.2. For a given type, suppose its utility distribution \(F\) over \(U\) is continuous and has full relative support, then any incentive compatible and Pareto efficient allocation rule can be supported as a Competitive Equilibrium from Equal Incomes (CEEI) with some price vector \(a \in (0, \infty)^{|S|}\).

The full proof of this contains fairly technical steps and is deferred to the Appendix. However, we explain the intuition behind the proof here.

In the standard mechanism design setup with monetary payments and quasi-linear utilities, incentive compatibility with multi-dimensional utilities is difficult to work with as incentive compatibility simply becomes requiring that the set \(X_t = \{x_t(u)\}\) lies on the boundary of its convex hull, and moreover that \(x_t(u)\) maximizes the linear function \(u \cdot x\) over this convex hull. This yields a correspondence between an incentive compatible \(x_t\) and a convex subset of \(\Delta\). Any IC \(x_t\) maps to a unique convex subset, and any convex subset correspond to IC \(x_t\)’s that gives the utilities \(u \cdot x_t(u) = \max_{p \in \Delta} \{p \cdot u\}\). Label the convex set that corresponds to \(x_t\) as \(X_t\).

Now, any convex set can be specified by a family of supporting hyperplanes. If there is one and only one supporting hyperplane that intersects \(x_t\) in the interior of \(\Delta\), then we are done since this hyperplane can be represented by a price vector. If there are two such hyperplanes that yield

9 A cone is a set \(C\) in which \(x \in C\) implies \(\lambda x \in C\) \(\forall \lambda \in (0, \infty)\).

10 The condition is called “cyclic monotonicity.” [Rochet 1987]
different points of tangency with $X_t$, then we show that there is a “trading direction” $d$ by which some positive measure of agents may move allocations in direction $d$ and others in direction $-d$ and all these agents strictly improve in utility, while maintaining the same average allocation for this type, thus contradicting Pareto efficiency. The existence of such positive measures of agents to make the trade possible is guaranteed by the full relative support assumption, which can be interpreted as a “liquidity” criterion.

3 Ordinal Mechanism

Many allocation mechanisms in practice do not elicit preference intensities but only relative rankings over preferences. Such mechanisms are called ordinal mechanisms. (In contrast, mechanisms that may elicit preference intensities are called cardinal mechanisms.) We develop a formulation and characterization of optimal ordinal mechanisms in a large market environment, analogous to our theory for cardinal mechanisms in section 2.

As before, there is a finite set $S$ of services. The space of allocations is the space of distributions $\Delta$ over the services. There is a finite set $T$ of agent types, and for each type $t \in T$ there is a measure $F_t$ describing the mass of agents with various utilities. $F_t$’s are common knowledge, while each agent’s utility is private knowledge. Since we assumed that $F_t$ is continuous, preference rankings are strict with probability one.

Let $\Pi$ be the set of permutations of $S$. Every $\pi \in \Pi$ represents a strict preference ranking over $S$. Let $U(\pi) \subseteq U$ be the set of utilities consistent with ranking $\pi$, in the sense that

$$u_{\pi(1)} > u_{\pi(2)} > \cdots > u_{\pi(|S|)}.$$ 

Let $F_t(\pi) = F_t(U(\pi))$ be the measure of agents of type $t$ that adhere to the strict preference ranking $\pi$.

An ordinal allocation mechanism is a mapping between preference rankings and distributions over services, $x_t : \Pi \rightarrow \Delta$. $x_t$ is incentive compatible if truth-telling maximizes utility: $\forall u \in U(\pi)$,

$$x_t(\pi) \in \arg \max_{\pi' \in \Pi} u \cdot x_t(\pi').$$

$x_t$ is ordinal efficient if agents cannot trade probabilities and all improve in the sense of first-order stochastic dominance. This is the ordinal analog to Pareto efficiency. Precisely speaking, $x_t$ is ordinal efficient if there does not exist another function $x_t' : \Pi \rightarrow \Delta$ with the same average allocation

$$\int_{\Pi} x_t' dF_t = \int_{\Pi} x_t dF_t,$$

but $x_t'$ always first-order stochastically dominates $x_t$, which means that $\forall \pi \in \Pi$, $\forall 1 \leq k \leq |S|$, 

$$\sum_{j=1}^{k} x_{t\pi(j)}(\pi) \geq \sum_{j=1}^{k} x_{t\pi'(j)}(\pi),$$

and the inequality is strict for some $k$ and some $\pi$ of positive measure, $F_t(\pi) > 0$.

As before, we call an ordinal allocation rule valid if it is incentive compatible and ordinal efficient. The collection of ordinal allocation rules for all types makes up an ordinal mechanism $x$. The objective is to optimize an arbitrary function of the mechanism, $W(x)$, subject to each $x_t$ being valid.
3.1 Characterization of Valid Ordinal Allocation Rules

We show that in the large market model, any valid ordinal allocation rule can be represented as “lottery-plus-cutoff”: agents are given lottery numbers distributed Uniform[0,1], and there is a type-specific lottery cutoff for each service; an agent is “admitted” to a service iff her lottery number does not exceed the cutoff. An agent chooses her most preferred service among those that she is admitted to. This is illustrated in Figure 2.

Definition 3.1. An ordinal allocation rule \( x : \Pi \rightarrow \Delta \) is lottery-plus-cutoff if there exists “cutoffs” \( a_s \in [0,1] \) such that

\[
x_{\pi(k)}(\pi) = \max_{j=1}^{k} a_{\pi(j)} - \max_{j=1}^{k-1} a_{\pi(j)}.
\]

Figure 2: Illustration of “lottery-plus-cutoff”. The vertical axis represents lottery numbers, which are uniformly distributed from 0 to 1. The dotted lines are lottery cutoffs for this type. The columns represent various preference reports. For preference report \( \pi_1 \), the allocation probability for service 3 is the difference \( a_3 - a_1 \), which represents lottery numbers for which the first choice service 1 is not accessible but the second choice service 3 is accessible.

Analogous to full relative support, the regularity condition on the utility distribution is that every choice ranking \( \pi \in \Pi \) is possible.

Definition 3.2. (Full ordinal support) \( F_t \) satisfies full ordinal support if \( F_t(\pi) > 0 \) for every preference ranking \( \pi \in \Pi \).

Theorem 3.3. For a given type, suppose its utility prior \( F \) has no ties and full ordinal support, and let \( x(\pi) \) be any incentive compatible and ordinal efficient allocation rule, then \( x(\pi) \) is lottery-plus-cutoffs for some cutoffs \( a_s \in [0,1]^{|S|} \).

The proof is given in Appendix [C]. The intuition behind it is similar to Theorem [2,2] In the cardinal case, incentive compatible allocation rules are associated with arbitrary convex subsets of \( \Delta \). In the ordinal setup, instead of a convex set we have a polymatroid. More precisely, for
any incentive compatible ordinal allocation rule $x_t$, the set $\{x_t\}$ is the vertex set of the base polymatroid of a monotone submodular function $f$, and any monotone submodular $f$ induces an incentive compatible allocation rule. Using this characterization, we show that subject to incentive compatibility, the full ordinal support condition implies that unless the allocation rule is lottery-plus-cutoff, there are agents who can trade with each other and yield allocations that first-order stochastically dominates their current allocation, which implies that the only ordinal efficient rules are lottery-plus-cutoff.

To give additional intuition on what lottery-plus-cutoff looks like, consider a given type and cutoffs $a$. Relabel the services so that $a_1 \leq a_2 \leq \cdots \leq a_{|S|}$. This sorts the services in increasing order of cutoffs, which can be intuitively interpreted as increasing order of “accessibility.” Let $M_k = \{k, \cdots, |S|\}$ denote the list of services from the $k$th in this order to the last. Note that $S = M_1 \supseteq M_2 \supseteq \cdots \supseteq M_{|S|}$. Lottery-plus-cutoff can be interpreted as follows: if an agent of this type gets the best lottery numbers $[0, a_1)$, then she is given the largest menu of choices $M_1 = S$, and can get any service she wants. Similarly, if her lottery number is in $(a_{k-1}, a_k]$, she chooses her most preferred service in menu $M_k$. It is straightforward to show that this implements the same assignment probabilities as in Definition 3.2. Hence, we also give lottery-plus-cutoff the name randomized menus with nested menus.

### 3.2 Comparing Cardinal and Ordinal Mechanisms

The proofs of Theorem 2.2 and 3.3 yield intuition on the nature of valid cardinal and ordinal mechanisms: in a valid cardinal mechanism, agents of the same type can trade probabilities of various services at different ratios, hence expressing their preferences not only for which service but also how much do they value each. In a valid ordinal mechanism, agents of the same type can also trade probabilities, but they must trade services one-for-one, hence they can only express preference rankings. In a certain sense, the value of a cardinal mechanism lies in its ability to differentiate agents with extreme preference intensities, so intuitively it is less valuable if agents preferences for various services are more equal, and is more valuable if preferences are more extreme.

We give an example in in Appendix C.3 that shows that with unrestricted utilities, the ratio between the optimal social welfare from a valid cardinal mechanism and the optimal social welfare of a valid ordinal mechanism may be arbitrarily large.

### 4 Empirical Application: Public School Assignment

We demonstrate how the framework developed in this paper can be applied to a real world problem and yield empirically relevant results. The problem we examine is based on the 2012-2013 Boston school assignment reform. One of the authors of this paper helped the city committee in charge of the reform to simulate a list of potential plans and evaluate based on a given portfolio of metrics by simulation. In this section, we ask the reverse question: given the objective function and constraints, what the optimal plan might have been like. Although we use real data, the problem presented here has been simplified for conceptual clarity. We recognize that in order to produce implementable

\[ a_1 \leq a_2 \leq \cdots \leq a_{|S|}. \]

\[ S = M_1 \supseteq M_2 \supseteq \cdots \supseteq M_{|S|}. \]
recommendations, the precise objective function and constraints must be scrutinized and debated over by all stakeholders and constituents, which has not yet taken place. Hence, the purpose of this section is not to give concrete policy recommendations for Boston, but to serve as a proof of concept and to showcase how our large market framework can be applied in a real world setting.

We first give a finite market formulation of the problem that the city committee faced during the reform. This is an asymmetric, multi-dimensional mechanism design problem with a complex objective and additional constraints, for which an exact optimal solution remains elusive. As a baseline, we describe the actual plan adopted by Boston Public Schools (BPS), which can be seen as an intuitive and relatively simple heuristic solution to the original problem. To apply the techniques in this paper, we first consider a large market approximation of the original problem, for which we can use the characterization results in this paper to efficiently solve for the optimal solution. We then define a finite market analog of this large market solution, which is a feasible mechanism in the original, finite market model, and which is asymptotically optimal in the sense that it becomes the large market optimum as the finite market model is scaled up. We compare this “asymptotically optimal” solution with the baseline, quantify the improvements, and discuss insights.

4.1 Finite Market Formulation

About 4000 students apply to Boston Public Schools (BPS) each in for grade Kindergarten 2 (K2), which is the main entry grade to elementary schools. A social planner is charged with designing an assignment system for K2 that is efficient, equitable, and that respects certain institutional, capacity, and budget constraints. The social planner partitions Boston into 14 neighborhoods. Based on historical data in 2010-2013 (4 years of data), the social planner estimates that the number of K2 applicants from each neighborhood is the product of two normally distributed random variables. The first term is common across neighborhoods, and has mean 4294 and standard deviation 115. This intuitively captures the overall number of applicants. The second term is specific for each neighborhood, and captures neighborhood specific variation in application rates. The mean and standard deviation for each neighborhood is estimated using historical data, and is shown in Table 4 in Appendix B.

Each of the 14 neighborhoods is broken down further into geocodes. The geocodes partition the city into 868 small contiguous blocks. The social planner uses geocodes to model student types, so there are 868 types. As an approximation, we use each geocode’s centroid as the reference location for all students in that geocode. Given the number of students from each neighborhood, we assume that each of these students is distributed among the geocodes of that neighborhood according to the historic average in years 2010-2013.

Each student is to be assigned to one of the 77 schools that has grade K2. The distribution of students across geocodes and the capacities of schools are plotted in Figure 4. The capacities are the actual numbers used in 2013.

12 Language Learners (ELL) and disabled students, accommodating continuing students, and allowing special status for students who have older siblings already attending a school. In this paper, we ignore grandfathering, specialized programs, continuing students, and whether or not students have older siblings at a school. However, all of these complications could in principle be accommodated in our approach, and we leave more realistic modeling to future work.

12 In 2010-2013, the average number of applicants to K2 in BPS for round 1 is about 4294 and the sample standard deviation is about 115.
Table 1: Parameters of the random utility model estimated using 2013 choice data, using grade K1-2 non-continuing, regular education students. (We use K1 data as well since families face similar choices and having more data allows greater precision.) The values can be interpreted in units of miles, as in how many miles a student is willing to travel for one unit of this variable.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>quality</td>
<td>0–6.29</td>
<td>Quality of schools. For a school of $\Delta q$ additional quality, holding fixed other components, a student would be willing to travel $\Delta q$ miles further. These values are graphically displayed in Figure 5. We normalize the smallest value to be 0.</td>
</tr>
<tr>
<td>$\omega$</td>
<td>0.86</td>
<td>Additional utility for going to a school within walk-zone.</td>
</tr>
<tr>
<td>$b$</td>
<td>1.88</td>
<td>Standard deviation of idiosyncratic taste shock.</td>
</tr>
</tbody>
</table>

Using historical choice data, the social planner estimates the following utility model for student $i$ in geocode $t$ and school $s$:

$$u_{is} = quality_s - Distance_{ts} + \omega \text{Walk}_{ts} + \beta \epsilon_{is}$$

where $quality_s$ represents “school quality,” which encapsulates all common propensities that families have to choose a school, such as facilities, test scores, teachers’ quality, and school environment. $Distance_{ts}$ is the distance from geocode $t$ to school $s$ in miles, estimated using Google Maps walk distance. The coefficient of $-1$ before the distance allows us to measure utility in the unit of miles, so that one additional unit of utility can be interpreted as “moving schools one mile closer.” $Walk_{ts}$ is an indicator for whether geocode $t$ is within the walk-zone of school $s$, which is an approximately 1-mile radius around the school in which students do not require bus transportation to the school. $\omega$ represents additional utility for walk-zone schools, since these schools are in the immediate neighborhood and students do not have to deal with the sometimes unpleasant experience of busing. $\epsilon_{is}$ is an idiosyncratic taste shock assumed to be i.i.d. standard Gumbel distributed, and $\beta$ represents the strength of the idiosyncratic taste shock. Variables starting with capital letters, $Distance_{ts}$ and $Walk_{ts}$, are directly from data, while variables in lower case, $quality_s$, $\omega$, and $\beta$ are estimated via maximum likelihood using historical choices. Using maximum likelihood, and real choice data from 2013, we estimate the model and plot the school qualities in Figure 5 and tabulate the other coefficients in Table 1. Although more sophisticated demand models are possible, since the focus of this paper is mechanism design rather than demand modeling, we use the above as the “true” utility distribution.

The social planner’s objective, $W$, is to maximize a linear combination of average welfare and max-min welfare,

$$W = \alpha \sum_t w_t v_t + (1 - \alpha) \min_t v_t,$$

---

13 In practice, there is a slight difference between bus ineligibility and being inside the walk-zone, as the walk-zone includes the whole geocode even if only a part of it is within 1-mile, while bus ineligibility only includes the part of the geocode strictly within the mile. However, this difference is small as geocodes are small, so for conceptual clarity we ignore it in this exercise.

14 The Gumbel distribution is chosen because it makes the model easy to estimate via maximum likelihood, as the likelihood function has a closed form expression.
where $v_t$ is the expected utility of a student from geocode $t$, $w_t$ is the proportion of students in geocode $t$ (taking the expected number of students from each geocode and normalizing so the weights sum to 1), and $\alpha$ is a parameter specifying the desired trade-off between efficiency and equity. $\alpha = 1$ represents social welfare (maximizing the average expected utility); $\alpha = 0$ represents max-min welfare (maximizing the expected utility for the worst-off geocode); $\alpha = .5$ represents an equal weighting of the two.

There are capacity limits $m_s$ for each school $s$, which are the number of seats available for K2 students. However, a certain set of schools $S_c \subseteq S$, which we call capacity schools, there is additional capacity available for students who live in the school's catchment region. We assume that the catchment region of a capacity school $s \in S_c$ is exactly the geocodes for which the $s$ is the closest capacity school. For capacity school $s$, the limit $m_s$ only applies to students outside of its catchment region, and we assume that it can accept unlimited number of students from its catchment region. This guarantees that even if no capacity is available elsewhere, each student can at least be assigned to the closest capacity school.\footnote{This is only approximately true in reality. Although BPS has committed to expand the capacity schools as needed by adding hiring new teachers and adding modular classrooms if needed, in reality there are hard space constraints and BPS has a more complicated system for guaranteeing that each student is assigned, which is based on distance but also many other factors. In BPS literature, capacity schools were later renamed “option schools.”}

There are 19 such capacity schools distributed across the city.

In addition to capacity constraints, the social planner faces a busing constraint. Let

$$B_{ts} = \begin{cases} \text{Distance}_{ts} & \text{if geocode } t \text{ is not in school } s \text{'s walk-zone.} \\ 0 & \text{otherwise.} \end{cases}$$

This represents the travel distance on a bus for a student from geocode $t$ to school $s$. Suppose that the social planner budgets $C$ miles of busing per student in expectation, then the busing constraint is

$$\sum_t w_t B_{ts} p_{ts} \leq C$$

where $p_{ts}$ is the probability that a random student in geocode $t$ is assigned to school $s$.\footnote{Note that this is a “soft” budget constraint in that the budget only has to be satisfied in expectation. We choose this because typically in school board operations the initial budget is only a projection but may be revised later if needed.}

In reality, busing cost is much more complicated, having to do with the routing, the number of buses used, the kinds of buses used, and legal requirements for more expensive door-to-door busing for certain Special Education students. We leave finer modeling of transportation costs in Boston to future work.

The institutional constraint is that the social planner must design a mechanism that is incentive compatible, “ex-post efficient” within type, and only requires eliciting preference rankings, rather than preference intensities. The reason we limit to preference rankings is that the system has used rankings from 1988 to 2012, so families are used to gathering and submitting this information. The reason we require incentive compatibility is that in 2006, a non-incentive compatible mechanism was rejected by the Boston school committee in favor of an incentive compatible one, and since then BPS has committed to having a mechanism that allows families to submit truthful preferences without worrying that it might negatively affect their chances.\footnote{For more details of that reform, see \textit{Abdulkadiroğlu et al. 2006}.} Ex-post efficiency within type means that
after the assignment of students to schools, students in the same type (geocode) should not be able to trade with one another and each improve in utility. This is to mitigate public discontentment with the mechanism, as students in the same geocodes are likely to compare assignments with one another.

In the following section, we describe a feasible solution to the above problem, which is the one actually adopted by the city after the reform. This is a baseline for comparison. The goal is to find a mechanism that uses the same transportation budget as the baseline, but improves significantly in efficiency and equity, as measured by utilitarian welfare and max-min welfare.

4.2 Baseline: Actual Implementation

The actual plan adopted by BPS for 2014 is called the Home Based Plan. It is based on a proposal in Shi [2013], although there are significant deviations. An input to this plan is a classification of BPS schools into 4 tiers using standardized test scores, with Tier 1 being the best and Tier 4 being the worst. Each student’s choice menu consists of any school within 1 mile, plus a certain number of closest schools of various types, as well as some idiosyncratic additions. For details of the Home Based Plan, see Appendix A. In this paper, we call this the “Baseline.”

The priority structure is as follows. Students who live in East Boston students get priority for East Boston schools, while non-East Boston students get priority for non-East Boston schools. We encode this using auxiliary variable $h_{ts}$, which is an indicator for whether both geocode $t$ and school $s$ are on the same side of the bridge.

Each student $i$ is given a random lottery number distributed Uniform[0, 1]. Her score to school $s$ is defined as $\sigma_{is} = r_i - h_{ts}$. This encodes the student’s priority to school $s$, with lower scores having higher priority.

Having defined the choice menu and priorities, the plan computes the assignment by the (student-proposing) Deferred Acceptance (DA) algorithm. which is as follows,

1. An arbitrary student $i$ applies to her top choice $s$.

2. School $s$ tentatively accepts the student.

3. If this acceptance causes the capacity of school $s$ to be exceeded, then the school finds the tentatively accepted student with the highest (worst) score and bumps her out. This school is then removed from that student’s choice ranking and the student applies to her next choice.

4. Iterate steps 1-3 until all unassigned students have empty choice rankings.

It is well known that this algorithm does not depend on the order of students’ application in step 1, and that the result is strategyproof, which means that it is a dominant strategy for all students to report their truthful preference rankings.

We simulate this plan 10000 times according to the environment described in Section 4.1, and tabulate the plan’s transportation burden, efficiency, equity, and predictability in Table 3, under the column “Baseline.”

---

18 The actual plan also contains priorities for continuing students, students with siblings to a school, and students wait-listed from previous rounds. Since we do not model these complexities, the priority structure is simpler.

19 The reason for this is that there is a cumbersome bridge between East Boston and the rest of Boston.

20 Note that if every student includes in her ranking her closest capacity school, then in the end no student will be unassigned.

21 roth-sotomayor
4.3 Solving the Large Market Approximation

We define the large market approximation to the model in Section 4.1 as follows. Replace each student with a continuum of independent students of mass 1. Instead of a stochastic mass of students of each type, approximate the scenario with a deterministic mass $n_t$ of students, setting $n_t$ to be its expected value.

This yields exactly the setting in Section 3, since the capacity constraints and the busing constraint can be incorporated into the objective function by setting regions for which any constraint is violated to negative infinity. Moreover, ex-post efficiency is equivalent to ordinal efficiency in the large market setting.\footnote{For works that study this equivalence, see Che and Kojima \cite{Che2011} and Liu and Pycia \cite{Liu2012}.}

By the characterization result in Section 3.1, it suffices to consider randomized menu mechanisms with nested menus, and as a relaxation it suffices to consider randomized menu mechanisms. For a menu of services $M \subseteq S$, abuse notation slightly and let $v_t(M)$ denote the expected utility of the best service in this menu for an agent of type $t$,

$$v_t(M) = \frac{1}{n_t} \int u \max_{s \in M} u_s dF_t(u).$$

Let $p_t(s, M)$ denote the probability that an agent of type $t$ would choose service $s$ to be her most preferred in menu $M \subseteq S$,

$$p_t(s, M) = \mathbb{P}\{s \in \arg \max_{s' \in M} u_{s'} : u \propto F_t\}.$$

Let $z_t(M)$ denote the probability an agent of type $t$ is shown menu $M \subseteq S$. Let $T_s$ denote the set of geocodes for which the capacity limit for school $s$ applies. For capacity schools, this is all geocodes that are not in its catchment region; for other schools, this is all geocodes. The optimal randomized menu mechanism is encoded by the following LP:

\begin{equation}
\text{(LargeMarketLP) max } \quad W = \alpha \sum_{t, M} w_t v_t(M) z_t(M) + (1 - \alpha) y \\
\text{s.t. } \quad y - \sum_{M} v_t(M) z_t(M) \leq 0 \quad \forall t \in T \\
\quad \sum_{M} z_t(M) = 1 \quad \forall t \in T \\
\quad \sum_{t \in T_s, M} n_t p_t(s, M) z_t(M) \leq m_s \quad \forall s \in S \\
\quad \sum_{s, t, M} n_t p_t(s, M) B_{ts} z_t(M) \leq C \\
\quad z_t(M) \geq 0 \quad \forall t \in T, M \subseteq S
\end{equation}

Since there are $2^{|S|} - 1$ possible menus $M \subseteq S$, the number of variables of this LP is exponential in $|S|$. However, it turns out that if the utility distribution has special structure, then an optimal solution to this LP can be found in time polynomial in $|T|$ and $|S|$.

**Definition 4.1.** Utility prior $F_t$ is multinomial-logit if the utilities can be written as

$$u_{is} = a_{ts} + b_t \epsilon_{is},$$

where $b_t \geq 0$ and $a_{ts}$ are given parameters, and $\epsilon_{is}$ are i.i.d. standard Gumbel distributed.
Note that the utility distribution in our model is *multinomial-logit*. This demand model implies that \( v_t(M) = b_t \log \left( \sum_{s \in M} \exp \left( \frac{a_{ts}}{b_t} \right) \right) \), and \( p_t(s, M) = \frac{\exp \left( \frac{a_{ts}}{b_t} \right)}{\sum_{s' \in M} \exp \left( \frac{a_{ts'}}{b_t} \right)} \).

**Theorem 4.2.** Suppose that utility distributions \( F_t \)'s are all multinomial-logit, and \( \alpha > 0 \), and weights \( w_t > 0 \) for all \( t \), then an optimal solution to the exponential sized LargeMarketLP can be found in time polynomial in \(|T|\) and \(|S|\).

The solution involves taking the dual of the LP, which has a small number of variables but exponentially number of constraints. This dual can be decomposed into a master problem that is polynomial sized and convex, and \(|T|\) independent sub-problems that have small number of variables but exponentially many constraints. The multinomial-logit assumption allows each of these sub-problems to be solvable in \(|S| \log |S|\) time, which allows the dual to be solved in polynomial time. Having solved the dual, we can efficiently find a polynomial subset of constraints that are tight, and discard all the other variables in the the original LP. This yields an optimal feasible solution to the original problem. The full proof is in Appendix C.

From the optimal solution to LargeMarketLP, we can infer the optimal cutoffs for each school and each geocode. For each school, students in geocodes that have cutoff 0 are those that should not be able to rank \( s \); students in geocodes with cutoff 1 should always be able to get into the school if they choose it; students in geocodes with intermediate cutoffs can be admitted to \( s \) if and only if they get a good enough lottery number.

For the transportation budget, we use 0.6 miles, which is just under the 0.63 used in the Baseline (see Table 3. For the objective, we consider \( \alpha = 1 \) (utilitarian welfare), \( \alpha = 0 \) (max-min welfare) and \( \alpha = 0.5 \) (equal weighting). We evaluate the optimal plan in the large market model and tabulate the results in Table 2. As seen, setting \( \alpha = 0.5 \) yields near optimal efficiency (average expected utility) and equity (expected utility for worst off type), so for the remainder of this paper we use \( \alpha = 0.5 \).

<table>
<thead>
<tr>
<th></th>
<th>( \alpha = 1 )</th>
<th>( \alpha = 0.5 )</th>
<th>( \alpha = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average expected utility</td>
<td>7.78</td>
<td><strong>7.66</strong></td>
<td>7.39</td>
</tr>
<tr>
<td>Expected utility for worst off type</td>
<td>2.52</td>
<td><strong>7.39</strong></td>
<td>7.39</td>
</tr>
</tbody>
</table>

### 4.4 Converting the Optimal Large Market Mechanism to a Feasible Finite Market Mechanism

To convert the optimal large market mechanism to a feasible finite market mechanism, we simply use the Deferred Acceptance (DA) algorithm and use the optimal large market cutoffs to guide the priorities.

Let \( a_{ts} \) be the cutoffs from any large market mechanism which incorporates the capacity constraints. For a student in type \( t \), define her menu to be schools for which her cutoff \( a_{ts} \) is positive. For schools in her menu, define her *score* to school \( s \) as \( \sigma_{is} = r_i - a_{ts} \). We use these menus and scores in the DA algorithm as defined in Section 4.2. We call this the DA analog to the optimal large market mechanism.

In the limit in which the students and school capacities are duplicated with many independent copies, then running after the DA analog, a student \( i \) is assigned to school \( s \) if and only if her score is
negative. In this case, each student’s assignment probabilities becomes identical to the probabilities she would have gotten in the large market approximation. Thus, using this method we can use a large market optimum to define a finite market mechanism that nevertheless is “asymptotically optimal” as the market is scaled up independently. (The independent scaling also removes the stochasticity in number of students of each type.)

4.5 Numerical Results

Let $\text{ApproximateOpt1}$ denote the DA analog to the optimal large market mechanism with $\alpha = 0.5$ and busing budget of 0.60 miles. We simulate 10000 times and compare its performance to Baseline in Table 3. It turns out that this plan evaluated in the finite market model uses 0.71 miles of busing, exceeds the 0.63 miles of Baseline. So we let $\text{ApproximateOpt2}$ to denote the DA analog to the optimal large market mechanism with $\alpha = 0.5$ and busing budget of 0.50. Evaluated in the finite market model, ApproximateOpt2 stays within the busing budget of Baseline, and significantly improves over it in terms of average utility, utility of worst-off type, and % of getting top 1 or top 3 choices. To give a frame of reference for the magnitude of the improvement, we also evaluate what we call the “Minimum” plan, which has no priorities, and only includes in the menu for each type the capacity school and schools with zero transportation cost (schools in the walk-zone). As seen, while the Baseline uses 0.29 miles of additional busing per student over the Minimum, it improves the average utility by 0.64 miles and the utility of the worst type by 1.67 miles. In offering these additional options it sacrifices slightly on predictability. However, ApproximateOpt2 uses only 0.28 miles of additional busing per student over the Minimum, while improving average utility by 1.18 miles (almost twice that of Baseline), and improving expected utility of worst-off type by 4.16 miles. It also significantly improves students’ chances of getting into top 1 and top 3 choices.

Table 3: Evaluating a variety of plans in the finite market model using 10000 independent simulations.

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Baseline</th>
<th>ApproximateOpt1</th>
<th>ApproximateOpt2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miles of busing/student</td>
<td>0.35</td>
<td>0.64</td>
<td>0.71</td>
<td>0.63</td>
</tr>
<tr>
<td>Average expected utility</td>
<td>6.31</td>
<td>6.95</td>
<td>7.62</td>
<td>7.49</td>
</tr>
<tr>
<td>Expected utility for worst off type</td>
<td>2.86</td>
<td>4.53</td>
<td>7.05</td>
<td>7.02</td>
</tr>
<tr>
<td>% getting top 1 choice</td>
<td>0.66</td>
<td>0.64</td>
<td>0.80</td>
<td>0.79</td>
</tr>
<tr>
<td>% getting top 3 choice</td>
<td>0.88</td>
<td>0.85</td>
<td>0.94</td>
<td>0.93</td>
</tr>
</tbody>
</table>

To gain intuition on what ApproximateOpt2 is doing, we compute for each school its “availability area” in each plan. This is the total area of the geocodes for which this school shows up in the menu. We then divide schools into quartiles by their quality score, with Q1 being the best and Q4 being the worst. We compare the average availability areas for different quality quartiles in Baseline and ApproximateOpt2 in Figure 3. As seen, ApproximateOpt2 offers lower quality schools to larger areas. The intuition is that the higher quality schools already have high demand from nearby areas, so it is more efficient in terms of transportation to restrict access to them to the nearby areas. However, to compensate the students who do not live near high quality schools, the plan offers them further away lower quality schools, in hope that the student will have high idiosyncratic tastes for them. Interestingly, the Baseline mimics the same behavior for Q1, Q2, Q3 schools, but not for Q4 schools, which makes sense in retrospect because some of the Q4 schools may be in risk of being closed.
Figure 3: Comparison of availability areas for schools of various quality in Baseline and ApproximateOpt2. Q1 is the best quartile in quality, Q4 is the worst.

5 Discussion

This paper studies the allocation of services to agents with private information and without monetary transfers. Priors over agents’ utilities are known to the social planner who is interested in maximizing an objective function. The approach in this paper sacrifices exact analysis of a finite market situation by a continuum approximation, in order to gain analytical tractability to handle large-scale applications with complex objectives and many types of agents and many kinds of services. In some sense, the thrust of this paper is to take mechanism design further into the “engineering realm,” focusing on tractable and useful approximations rather than complex, exact analysis.

We provide characterizations of incentive compatible mechanisms that are Pareto-optimal within each type in large markets and show how to find the optimal ordinal mechanism. An open question regards the efficient computation of the optimal cardinal mechanism.\textsuperscript{23} As known from the auction literature, prior information can impact significantly the mechanism as observed in our empirical exercise.

One difficulty in actually implementing optimal mechanism design without money is in estimating the utility distributions. With transfers, valuation estimation becomes simpler as the social planner may infer willingness to pay from past transactional data. However, without money, it is harder to infer preferences. The demand modeling used in our empirical application assumes a particular functional form for the utilities, but the underlying utilities are not observable, so one may question its validity. A more fundamental question is whether human behavior can indeed be captured with such models.\textsuperscript{24} Nevertheless, if a utility model can be estimated and trusted, the methods used in this paper can be used to compute the optimal mechanism.

\textsuperscript{23} After discretizing the prices, our current techniques can reduce this to exponential sized linear program, but we have not found any interesting distributions for which we can efficiently solve the optimal-menu subproblem.

\textsuperscript{24} One project that seeks to validate the use of utility models in Boston school choice, compared to an alternative model based on Marketing or salience, is [Pathak and Shi 2014], in which the authors uses various methods to predict how families will choose schools after the 2012-2013 reform, pre-commit to the predictions before the new choice data is collected, and compare the prediction accuracies after.
References


A  Details of the Home Based Plan

In the Home Based Plan implemented in 2014, a student’s choice menu is the union of the following sets.

- any school within 1 mile straight line distance;
- the closest 2 Tier 1 schools;
- the closest 4 Tier 1 or 2 schools;
- the closest 6 Tier 1, 2 or 3 schools;
- the closest school with Advanced Work Class (AWC);
- the closest Early Learning Center (ELC), which are extended-day kindergartens;
• the 3 closest capacity schools.
• the 3 city-wide schools, which are available to everyone in the city.

Furthermore, for students living in parts of Roxbury, Dorchester, and Mission Hill, their menu includes the Jackson/Mann school in Allston/Brighton.

B Additional Tables and Figures

Table 4: Means and standard deviations of the proportion of K2 applicants from each neighborhoods. This is estimated using 4 years of historic data.

<table>
<thead>
<tr>
<th>Neighborhood</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allston-Brighton</td>
<td>0.0477</td>
<td>0.0018</td>
</tr>
<tr>
<td>Charlestown</td>
<td>0.0324</td>
<td>0.0024</td>
</tr>
<tr>
<td>Downtown</td>
<td>0.0318</td>
<td>0.0039</td>
</tr>
<tr>
<td>East Boston</td>
<td>0.1335</td>
<td>0.0076</td>
</tr>
<tr>
<td>Hyde Park</td>
<td>0.0588</td>
<td>0.0022</td>
</tr>
<tr>
<td>Jamaica Plain</td>
<td>0.0570</td>
<td>0.0023</td>
</tr>
<tr>
<td>Mattapan</td>
<td>0.0759</td>
<td>0.0025</td>
</tr>
<tr>
<td>North Dorchester</td>
<td>0.0522</td>
<td>0.0047</td>
</tr>
<tr>
<td>Roslindale</td>
<td>0.0771</td>
<td>0.0048</td>
</tr>
<tr>
<td>Roxbury</td>
<td>0.1493</td>
<td>0.0096</td>
</tr>
<tr>
<td>South Boston</td>
<td>0.0351</td>
<td>0.0014</td>
</tr>
<tr>
<td>South Dorchester</td>
<td>0.1379</td>
<td>0.0065</td>
</tr>
<tr>
<td>South End</td>
<td>0.0475</td>
<td>0.0022</td>
</tr>
<tr>
<td>West Roxbury</td>
<td>0.0638</td>
<td>0.0040</td>
</tr>
</tbody>
</table>

C Omitted Proofs

C.1 Characterization for Cardinal Mechanisms

of Theorem 2.2 The proof uses a series of lemmas. For exposition purposes, we first show the main proof, and we prove the lemmas later.

Lemma C.1. A cardinal allocation rule is incentive compatible if and only if there exists closed convex set $X \subseteq \Delta$ such that $x(u) \in \arg \max_{y \in X} \{u \cdot y\}$, $\forall u \in U$. We call $X$ the closed convex set that corresponds to incentive compatible allocation rule $x$.

Lemma C.1 says that any incentive compatible allocation rule can be represented by a closed convex set $X$ in which $y = x(u)$ is a maximizer to the linear objective $u \cdot y$ subject to $y \in X$. This is illustrated in Figure 6.

We proceed to prove the theorem by induction on $|S|$. For $|S| = 1$, there is nothing to prove as $\Delta$ is one point. Suppose we have proven this theorem for all smaller $|S|$. Let $X$ be the convex set that corresponds to $x$. Suppose $X$ does not intersect the interior of $\Delta$, $\text{int}(\Delta) = \{y \in \mathbb{R}^{|S|} : y >$
Figure 4: Distribution of students and capacity of schools. Each blue circle represents a geocode, with its area proportional to the expected number of students from that geocode. Each yellow circle represents a school, with its area proportional to the number of K2 seats available. The distribution of students is based on 4 years of real data, in 2010-2013. The capacities are based on data from 2013.

Let $H(u, \alpha)$ denote the $|S| - 1$ dimensional hyperplane $\{y \in \mathbb{R}^{|S|} : u \cdot y = \alpha\}$. Let $H^-(u, \alpha)$ denote the half-space $\{y : u \cdot y \leq \alpha\}$, and $H^+(u, \alpha)$ denote $\{y : u \cdot y \geq \alpha\}$. Let $\text{span}(\Delta)$ denote the linear extension of $\Delta$, $\text{span}(\Delta) = \{y \in \mathbb{R}^{|S|} : \sum_s y_s = 1\}$. The following lemma allows us to express tangents of $X$ in $\text{span}(\Delta)$ in terms of a price vector $a \in (0, \infty)^{|S|}$. 

$0, \sum_s y_s = 1$, then some component of $x$ must be restricted to zero, so we can set the price for that service to infinity, ignore that service, and arrive at a scenario with smaller number of services, for which the theorem is true by induction. Thus, it suffices to consider the case $X \cap \text{int}(\Delta) \neq \emptyset$. 

Let $H(u, \alpha)$ denote the $|S| - 1$ dimensional hyperplane $\{y \in \mathbb{R}^{|S|} : u \cdot y = \alpha\}$. Let $H^-(u, \alpha)$ denote the half-space $\{y : u \cdot y \leq \alpha\}$, and $H^+(u, \alpha)$ denote $\{y : u \cdot y \geq \alpha\}$. Let $\text{span}(\Delta)$ denote the linear extension of $\Delta$, $\text{span}(\Delta) = \{y \in \mathbb{R}^{|S|} : \sum_s y_s = 1\}$. The following lemma allows us to express tangents of $X$ in $\text{span}(\Delta)$ in terms of a price vector $a \in (0, \infty)^{|S|}$. 

22
Figure 5: Estimates of quality, (inferred quality) from 2013 choice data. The size of the circle is proportional to the estimated quality, with higher quality schools having larger circles. The other parameter estimates are shown in Table I.

Lemma C.2. Any tangent hyperplane of $X$ in $\text{span}(\Delta)$ can be written as $H(a, 1) \cap \text{lin}(\Delta)$, for some $a \in (0, \infty)^{|S|}$, with $a$ pointing outward from $X$ and not co-linear with $1$. ($a \cdot y \leq 1$, $\forall y \in X$, and $a \neq \lambda 1$ for any $\lambda \in \mathbb{R}$.)

For any set $A \subseteq U$, define the average allocation of agents in $A$ to be

$$\bar{x}(A) = \int_A x(u) dF(u).$$

For any vector $a \in \mathbb{R}^{|S|}$ not co-linear with $1$, let $\text{Proj}_U(a)$ denote its projection onto $U$, and let $\tilde{a}$ denote its unit projection $\frac{\text{Proj}_U(a)}{\|\text{Proj}_U(a)\|}$. Since type-specific-pricing without infinite prices is the
Figure 6: An incentive compatible allocation rule with $|S| = 3$. $X$ is an arbitrary closed convex subset of feasibility simplex $\Delta$. $y = x(u)$ is maximizer of linear objective $u \cdot y$ with $y \in X$.

same as having $X = H^-(a, 1) \cap \Delta$, it suffices to show that the convex set $X$ has only one tangent in $\text{int}(\Delta)$. Intuitively, if it has two different tangents $H(a, 1)$ and $H(a', 1)$, with non-zero and unequal unit projections $\tilde{a} \neq \tilde{a}'$, then we can find a unit vector $d \in U$ s.t. $d \cdot a > 0 > d \cdot a'$. Since $a$ and $a'$ are tangent normals, we can perturb $x(u)$ in direction $d$ for $u$ near $\tilde{a}$, and perturb $x(u)$ in direction $-d$ for $u$ near $\tilde{a}'$, thus Pareto improving $x(\cdot)$ but keeping average $\bar{x}(U)$ fixed. This is illustrated in Figure [7]. However, defining a feasible move with positive measure in all cases is non-trivial, as prior $F$ and closed convex set $X$ are general. To do this, we prove the following lemma.

Figure 7: Exchange argument to Pareto improve the allocation rule by expanding $X$ along opposite directions, when there is more than one tangent of $X$ intersecting $\text{int}(\Delta)$.

**Lemma C.3.** Suppose that $H(a, 1)$ is an outward pointing supporting hyperplane of $X$ that intersects $X$ in the relative interior of the feasibility simplex, $\text{int}(\Delta)$. Then for any unit vector $d \in U$ such that $d \cdot a > 0$, there exists $\delta_0 > 0$ such that for all $\delta \in (0, \delta_0)$, there exists allocation rule $x'$ that strictly dominates $x$, with $\bar{x}'(U) = \bar{x}(U) + \delta d$.

Using this, we can rigorously carry out the above argument: suppose that $H(a, 1)$ and $H(a', 1)$ are two outward-pointing supporting hyperplanes of $X$ that intersect $X$ in $\text{int}(\Delta)$, with different non-zero unit projections onto $U$, $\tilde{a} \neq \tilde{a}'$. Take any unit vector $d \in \text{int}(H+(a, 0) \cap H^-(a', 0)) \cap U$.
(Such \( \mathbf{d} \) exists since \( \mathbf{a} \neq \mathbf{a}' \).) Then \( \mathbf{d} \cdot \mathbf{a} > 0 > \mathbf{d} \cdot \mathbf{a}' \). Using Lemma \text{C.3} there exists allocation rule \( \mathbf{x}' \) and \( \mathbf{x}'' \) which both strictly dominate \( \mathbf{x} \), one of which has average allocation \( \mathbf{x}(U) + \delta \mathbf{d} \), and the other \( \mathbf{x}(U) - \delta \mathbf{d} \). Taking \( \mathbf{x}''' = \frac{1}{2}(\mathbf{x}' + \mathbf{x}'') \), we have that \( \mathbf{x}''' \) also strictly dominates \( \mathbf{x} \), but \( \mathbf{x}'''(U) = \mathbf{x}(U) \), contradicting the cardinal efficiency of \( \mathbf{x}(\cdot) \). Therefore, \( \mathbf{X} \) has only one supporting hyperplane in \( \Delta \) that intersects it in the interior \( \text{int}(\Delta) \). \( \square \)

of Lemma \text{C.4} Suppose cardinal allocation rule \( \mathbf{x}(\mathbf{u}) \) is incentive compatible. Let \( \mathbf{X} \) be the convex closure of its range. Then \( \forall \mathbf{u} \in \mathbf{U}, \mathbf{x}(\mathbf{u}) \in \arg \max_{\mathbf{y} \in \mathbf{X}} \{ \mathbf{u} \cdot \mathbf{y} \} \). Conversely, if for some closed convex set \( \mathbf{X} \), for any \( \mathbf{u} \in \mathbf{U}, \mathbf{x}(\mathbf{u}) \in \arg \max_{\mathbf{y} \in \mathbf{X}} \{ \mathbf{u} \cdot \mathbf{y} \} \), then \( \forall \mathbf{u}' \in \mathbf{U}, \mathbf{x}(\mathbf{u}') \in \mathbf{X} \), so \( \mathbf{u} \cdot \mathbf{x}(\mathbf{u}) \geq \mathbf{u} \cdot \mathbf{x}(\mathbf{u}') \). So \( \mathbf{x} \) is incentive compatible. \( \square \)

of Lemma \text{C.2} Any tangent hyperplane \( \mathbf{Y} \) of \( \mathbf{X} \) in \( \text{span}(\Delta) \) is a \( |\mathbf{S}| - 2 \) dimensional subspace of the \( |\mathbf{S}| - 1 \) dimensional linear space \( \text{span}(\Delta) \). Take an arbitrary point \( \mathbf{z} \in \Delta \) on the same side of \( \mathbf{Y} \) as \( \mathbf{X} \). Consider the \( |\mathbf{S}| - 1 \) dimensional hyperplane \( \mathbf{H} \) passing through \( \mathbf{Y} \) and \( (1+\epsilon)\mathbf{z} \). For some sufficiently small \( \epsilon > 0 \), by continuity, \( \mathbf{H} \) has all positive intercepts, so \( \mathbf{H} = \{ \mathbf{y} : \mathbf{a} \cdot \mathbf{y} \leq 1 \} \) for some \( \mathbf{a} > 0 \), and by construction, \( \mathbf{a} \) is not co-linear with \( \mathbf{1} \). Now, \( \mathbf{a} \cdot \mathbf{z} = \frac{1}{1+\epsilon} < 1 \), so \( \mathbf{a} \) points outward from \( \mathbf{X} \). \( \square \)

The following rather technical lemma is used in the proof of Lemma \text{C.3}

\textbf{Lemma C.4.} Given any bounded closed convex set \( \mathbf{X} \subseteq \mathbb{R}^n \), non-empty open cone \( \mathbf{C} \subseteq \mathbb{R}^n \) and measurable function \( \mathbf{x} : \mathbf{C} \rightarrow \mathbb{R}^n \) such that \( \mathbf{x}(\mathbf{u}) \in \arg \max_{\mathbf{y} \in \mathbf{X}} \{ \mathbf{u} \cdot \mathbf{y} \} \). Let \( \mathbf{F} \) be an atomless measure with full relative support on \( \mathbf{C} \), with \( \mathbf{F}(\mathbf{C}) > 0 \). Define

\[ \tilde{\mathbf{X}} = \{ \tilde{\mathbf{x}}(\mathbf{A}) = \int_{\mathbf{A}} \mathbf{x}(\mathbf{u}) d\mathbf{F}(\mathbf{u}) / \mathbf{F}(\mathbf{A}) : \mathbf{F}(\mathbf{A}) > 0, \mathbf{A} \subseteq \mathbf{C} \} \]

Then \( \forall \mathbf{u} \in \mathbf{C} \), \( \mathbf{u} \cdot \mathbf{x}(\mathbf{u}) \in \text{closure}(\tilde{\mathbf{X}}) \) (closure of the convex closure of \( \tilde{\mathbf{X}} \)).

of Lemma \text{C.4} We first show that \( \tilde{\mathbf{X}} \) is convex following the proof of Lemma 3.3 in [Zhou 1992]. For any \( \mathbf{A} \subseteq \mathbf{C} \), define the \( n + 1 \) dimensional measure

\[ m(\mathbf{A}) = \left( \int_{\mathbf{A}} \mathbf{x}(\mathbf{u}) d\mathbf{F}(\mathbf{u}), \mathbf{F}(\mathbf{A}) \right) \]

By Lyapunov’s convexity theorem, since \( \mathbf{F} \) is atomless, the range of this measure, denoted \( \mathbf{M} \), is convex. Therefore, the cone generated by \( \mathbf{M} \), \( \text{cone}(\mathbf{M}) \), is convex, and so its intersection with the hyperplane \( (\cdot, 1) \) is convex. This intersection is non-empty since \( \mathbf{F}(\mathbf{C}) > 0 \). This intersection, restricted to first \( n \) components is exactly \( \tilde{\mathbf{X}} \), so \( \tilde{\mathbf{X}} \) is convex. Moreover, since every point of \( \tilde{\mathbf{X}} \) is a convex combination of points in \( \mathbf{X} \) and since \( \mathbf{X} \) is closed and convex, \( \text{closure}(\tilde{\mathbf{X}}) \subseteq \mathbf{X} \).

Since every point of \( \tilde{\mathbf{X}} \) is a convex combination of points in \( \mathbf{X} \) and since \( \mathbf{X} \) is closed and convex, \( \text{closure}(\tilde{\mathbf{X}}) \subseteq \mathbf{X} \).

Suppose on the contrary that \( \exists \mathbf{u}_0 \in \mathbf{C} \) and \( \mathbf{y}_0 \in \arg \max_{\mathbf{y} \in \mathbf{X}} \{ \mathbf{u}_0 \cdot \mathbf{y} \} \) s.t. \( \mathbf{y}_0 \not\in \text{closure}(\tilde{\mathbf{X}}) \). Since \( \text{closure}(\tilde{\mathbf{X}}) \) is closed, convex, and bounded, there exists strictly separating hyperplane \( \mathbf{H}(\mathbf{u}_1, \alpha_1) \) s.t. for some \( \delta_1 > 0 \), \( \mathbf{u}_1 \cdot \mathbf{y}_0 \geq \alpha_1 + \frac{\delta_1}{2} > \alpha_1 - \frac{\delta_1}{2} \geq \mathbf{u}_1 \cdot \mathbf{y}, \forall \mathbf{y} \in \text{closure}(\tilde{\mathbf{X}}) \). Now \( \mathbf{u}_0 \cdot \mathbf{y}_0 \geq \mathbf{u}_1 \cdot \mathbf{y} \) \( \forall \mathbf{y} \in \text{closure}(\tilde{\mathbf{X}}) \subseteq \mathbf{X} \) by construction of \( \mathbf{y}_0 \). So since \( \mathbf{X} \) is open, and \( \mathbf{u}_0 \in \mathbf{C} \), for sufficiently small \( \epsilon > 0 \) we have \( \mathbf{u}_2 = \mathbf{u}_0 + \epsilon \mathbf{u} \in \mathbf{C} \) and \( \mathbf{u}_2 \) is a strictly separating hyperplane. This is because if
Let \( y_0 = y \in X \cap H(a, 1) \cap \text{int}(\Delta) \). Let \( u_0 = \text{Prof}_U(a) \) (the projection of \( a \) onto \( U \)). Let \( y \) be the distance from \( y_0 \) to the boundary of \( \Delta \), then \( y = 0 \) since \( y_0 \) is in the interior of the simplex. Take \( \epsilon = \frac{\gamma}{2} \). Define the cones

\[
C_1 = \left\{ u : u \cdot x(u) < u \cdot y_0 + \frac{\epsilon}{2} \| u_0 \| \right\}
\]

\[
C_2 = \left\{ u : u \cdot d > \left\| u \right\| \frac{u_0 \cdot d}{\| u_0 \|} \right\}
\]

\( C_1 \) is a cone because by the convex set characterization in Lemma C.1 when \( u \) is scaled by a positive scalar \( \lambda > 0 \), \( u \cdot x(u) \) is scaled up also by \( \lambda \). \( C_2 \) is clearly a cone. \( C_1 \) is open because \( g(u) = u \cdot x(u) \) is a continuous function of \( u \) as this is the objective of the linear maximizer over convex set \( X \). \( C_2 \) is also open by continuity of \( u \cdot d - b\| u \| \) for constants \( d, b \). Therefore, the set

\[
C_3 = C_1 \cap C_2
\]

is an open cone. Moreover, it is non-empty since \( u_0 \in C_3 \).

Now, \( X \) is bounded, closed and convex. \( C_3 \) is a non-empty open cone.

Therefore by Lemma C.4 \( y_0 \in \arg \max_{y \in X} \{ u_0 \cdot y \} \subseteq \text{closure}(X) \), so there exists \( \bar{A} \subseteq C_3 \), \( F(A) > 0 \), s.t. \( \| \bar{x}(A) - y_0 \| \leq \frac{\epsilon u_0 \cdot d}{\| u_0 \|} \). Now, consider the alternative allocation rule

\[
y(u) = \begin{cases} 
\bar{x}(A) + 6\epsilon d & \text{if } u \in A \\
 x(u) & \text{otherwise.}
\end{cases}
\]

\( y(u) \in \Delta \forall u \in U \) since \( 7\epsilon < \gamma \).

By construction, \( y(U) = \bar{x}(U) + 6\epsilon F(A) d \). Moreover, \( y \) strictly Pareto improves \( x \) because \( \forall u \in A, \)

\[
u \cdot \bar{x}(A) + 6\epsilon u \cdot d - u \cdot x(u)
\]

\[
> u \cdot y_0 - \epsilon \left( \frac{u_0 \cdot d}{\| u_0 \|} \| u \| + 3\epsilon \| u \| \frac{u_0 \cdot d}{\| u_0 \|} \right) - u \cdot x(u)
\]

\[
> \epsilon \left( \| u \| \frac{u_0 \cdot d}{\| u_0 \|} \right) - u \cdot x(u)
\]

\[
> 0
\]

Now, let \( \delta_0 = 6\epsilon F(A) \), for any \( \delta \in (0, \delta_0) \), if we set

\[
x'(u) = \frac{\delta}{\delta_0} y(u) + (1 - \frac{\delta}{\delta_0}) x(u).
\]

Then \( x' \) still strictly Pareto improves \( x \) since objectives are linear, but \( x'(U) = \bar{x}(U) + \delta d \), which is what we needed.
C.2 Characterization for Ordinal Mechanisms

of Theorem 3.3. The proof is similar to that of Theorem 2.2 in that we first find an equivalent description of incentive compatibility and then use an exchange argument to derive the lottery-plus-cutoffs structure. The difference is that instead of a closed convex set as in the proof of Theorem 2.2, we have the base polytope of a polymatroid. The exchange argument is also simpler because the space of permutations \( \Pi \) is discrete and every member has positive probability due to full relative support.

The proof of the following lemmas are in Appendix C.

Lemma C.5. An ordinal allocation rule \( x(\pi) \) is incentive compatible if and only if there exists monotone submodular set function \( f : 2^{|S|} \rightarrow [0, 1] \) s.t. \( \forall 1 \leq k \leq |S|, \)
\[
x_{\pi(k)}(\pi) = f(\{\pi(1), \ldots, \pi(k)\}) - f(\{\pi(1), \ldots, \pi(k-1)\}).
\]
We call \( f \) the monotone submodular set function that corresponds to \( x \).

If \( X \) is the range of \( x \), then the above lemma says that \( x \) is incentive compatible if and only if \( X \) is the vertex set of the base polytope of polymatroid defined by \( f \):
\[
\sum_{s \in M} x_s \leq f(M) \quad \forall M \subseteq S \\
\sum_{s \in S} x_s = 1 \\
x \geq 0
\]

The following lemma embodies the exchange argument.

Lemma C.6. Let \( f \) be the monotone submodular set function that corresponds to incentive compatible allocation rule \( x \). If \( x \) is ordinal efficient, then for any \( M_1, M_2 \subseteq S \),
\[
f(M_1 \cup M_2) = \max\{f(M_1), f(M_2)\}.
\]

Let \( a_s = f(\{s\}) \). An easy induction using Lemma C.6 yields \( \forall M \subseteq S, f(M) = \max_{s \in M} a_s \), which together with Lemma C.5 implies that \( x \) is lottery-plus-cutoffs.

of Lemma C.5. If \( x(\pi) \) is an incentive compatible ordinal allocation rule, then for any \( M \subseteq S \), define
\[
f(M) = \sum_{j=1}^{|M|} x_{\pi(j)}(\pi), \quad \text{where} \{\pi(1), \pi(2), \ldots, \pi(|M|)\} = M.
\]

This is well-defined because incentive compatibility requires each agent’s chances of getting a service in \( M \), conditional on ranking these first in some order (ranking all of \( M \) before all of \( S \setminus M \), to be fixed, regardless of the relative rank between services in \( M \) and services in \( S \setminus M \). If this were not the case, then for some large \( b > 0 \) and small \( \epsilon \geq 0 \), consider an agent with utilities \( u_s = b + \epsilon_s \) for \( s \in M \) and \( u_s = -c + \epsilon_s \), where \( \{\epsilon_s < \epsilon\} \) are small distinct numbers that sum to zero, and \( c = b|M|/(|S| - |M|) \) is set to preserve normalization \( \sum_s u_s = 0 \). If the agent’s chance of getting a service in \( M \) can be altered by changing relative order in \( M \) and relative order in \( S \setminus M \), while she
ranks $M$ before $S \setminus M$, then the agent would for some $\{\epsilon_s\}$’s gain $b$ times a positive number and lose at most $|S|\epsilon$, so for sufficiently large $\frac{b}{\epsilon}$ she has incentive to mis-report.

We now show that $f$ is submodular. Suppose on the contrary that $f$ is not submodular, then there exists $M_1 \subseteq M_2$, and $s \notin M_2$, such that

$$f(M_1 \cup \{s\}) - f(M_1) < f(M_2 \cup \{s\}) - f(M_2).$$

However, for sufficiently large $b$, let $u_s = b$ for $s \in M_1 \cup \{s\}$, and $u_s = -c$ for $s \in M_2 \setminus M_1$, where $c = b(|M_1| + 1)/(|M_2|)$ is set to preserve the normalization $\sum_s u_s = 0$. She would have incentive to rank elements of $M_2 \setminus M_1$ before $s$ and improve expected utility. This contradicts incentive compatibility.

Now, the construction of $f$ implies $f$ is monotone, and that $\forall \pi \in \Pi$ and $1 \leq k \leq |S|$, $x_{\pi(k)}(\pi) = f(\{\pi(1), \ldots, \pi(k)\}) - f(\{\pi(1), \ldots, \pi(k-1)\})$.

Conversely, if $f$ is a monotone submodular set function. We show that if we define $\mathbf{x}$ so that $x_{\pi(k)}(\pi) = f(\{\pi(1), \ldots, \pi(k)\}) - f(\{\pi(1), \ldots, \pi(k-1)\})$, then $\mathbf{x}$ is incentive compatible. Note that the range of $\mathbf{x}$ defined this way is simply the vertex set of the base polytope of the polymatroid defined by $f$:

$$\sum_{s \in M} x_s \leq f(M) \quad \forall M \subseteq S$$
$$\sum_{s \in S} x_s = 1$$
$$x_s \geq 0 \quad \forall s \in S.$$

Now, the agent’s utility $\mathbf{u} \cdot \mathbf{x}$ is linear in $\mathbf{x}$, so using the fact that the greedy algorithm optimizes a linear objective over a polymatroid (and also the base polytope), we get that for any $\mathbf{u}$, if we re-label $S$ so that

$$u_1 \geq u_2 \geq \cdots \geq u_{|S|}.$$  

Then an optimal point of the base polytope simply sets $x_1$ to $f(\{1\})$, and $x_2$ to $f(\{1, 2\}) - f(\{1\})$ and so on, which is exactly how we defined $\mathbf{x}$. Thus, $\mathbf{x}(\pi) \in \arg \max_{\pi \in \Pi} \mathbf{u} \cdot \mathbf{x}(\pi')$, and $\mathbf{x}$ is incentive compatible. 

\[\square\]

of Lemma [C.6] By monotonicity of $f$, $f(M_1 \cup M_2) \geq \max\{f(M_1), f(M_2)\}$. What we need to show is that $f(M_1 \cup M_2) \leq \max\{f(M_1), f(M_2)\}$. By monotonicity, it suffices to show this for the case in which $M_1 \cap M_2 = \emptyset$.

Suppose that on the contrary that $f(M_1 \cup M_2) > \max\{f(M_1), f(M_2)\}$, $M_1 \cap M_2 = \emptyset$. Consider two preference rankings, $\pi_1$ and $\pi_2$: $\pi_1$ ranks services in $M_1$ first, followed by $M_2$, followed by other services in arbitrary order; $\pi_2$ ranks services in $M_2$ first, followed by $M_1$, followed by others. By Lemma [C.5] since $\mathbf{x}$ is incentive compatible, $\sum_{s \in M_1} x_{\pi_1}(\pi_1) = f(M_1 \cup M_2) - f(M_1) > 0$, and $\sum_{s \in M_1} x_{\pi_2}(\pi_2) = f(M_1 \cup M_2) - f(M_2) > 0$. Thus, agents with preference ranking $\pi_1$ can trade probabilities with agents with preference ranking $\pi_2$ and mutually improve in the first-order stochastic dominance sense. (Agents preferring $M_1$ get additional probabilities for services in $M_1$ in place of equal probabilities for $M_2$, while agents preferring $M_2$ get additional probabilities for $M_2$ in place of $M_1$.) By full relative support, there exist positive measures of both kinds of agents, so $\mathbf{x}$ is not ordinal efficient, contradiction. 

\[\square\]
C.3 Comparing Cardinal and Ordinal Mechanisms

We show an example in which the optimal social welfare from a cardinal mechanism is arbitrarily many times larger than the optimal social welfare from an ordinal mechanism. This examples uses the intuition that the value of a cardinal mechanism lies most in its ability to distinguish agents that have extremely large preferences for some services.

Let $M$ and $N$ be two positive real numbers larger than one, to be determined later. Suppose that there is only one service and an outside option (which can be supported in our model as another service). The service has capacity $\frac{1}{N^2}$, while the outside option has infinite capacity. Suppose there is only one type of mass 1, and $\frac{1}{N}$ of the agents have utilities $(M+1, 1)$ and the remaining agents $(2, 1)$.26 The objective is to maximize the social welfare. An optimal cardinal mechanism charges price vector $(N, 0)$, and agents have budget 1. The $\frac{1}{N}$ of agents who value the service a lot would purchase the bundle $(1, 1)$, while the other agents will opt for the outside option. Hence, the social welfare is $1 + \frac{M}{N^2}$, since everyone gets utility 1 from the outside option except the $\frac{1}{N^2}$ mass of agents that are assigned the service. On the other hand, an ordinal mechanism cannot differentiate between these agents, and everyone would choose the service over the outside option, and the best a lottery-plus-cutoff mechanism can do is to have cutoffs $(\frac{1}{N^2}, 1)$ for the two services, and social welfare is $1 + \frac{M}{N^3}$, since at best $\frac{1}{N^2}$ fraction of the service can be allocated to the agents who really value it. By making both $N$ and $\frac{M}{N^3}$ large, we can make the ratio of these arbitrarily large.

C.4 Computation Results of Theorem 4.2.

We take the dual:

\[
\text{(Dual) } \min C\gamma + m \cdot \lambda + \sum_{t \in T} \mu_t \\
\mu_t \geq (\alpha w_t + \nu_t)v_t(M) - n_t \sum_s p_t(s, M)(1(t \in T_s)\lambda_s + \gamma B_{ts}) \quad \forall t \in T, M \subseteq S \\
\sum_{t \in T} \nu_t \geq 1 - \alpha \\
\gamma, \lambda, \mu, \nu \geq 0
\]

Label the right hand side of the first inequality as $f_t(\gamma, \lambda, \nu, M)$. This can be interpreted as follows: suppose that one unit of expected utility for the agent of type $t$ contributes $\alpha w_t + \nu_t$ “credits” to the city, while assigning her to school $s$ costs the city $1(t \in T_s)\lambda_s + \gamma B_{ts}$ “credits,” then $f_t(\gamma, \lambda, \nu, M)$ is the expected number of “credits” an agent of type $t$ who is given menu $M$ contributes to the city, taking into account both her expected utility and the negative externalities of her occupying a slot of a service. Maximizing this over menus $M$ is thus a type of an “optimal-menu” problem.

**Definition C.7.** Given $\gamma, \lambda, \nu$, the optimal menu sub-problem is to find the solution set

$$\arg \max_{M \subseteq S} f_t(\gamma, \lambda, \nu, M).$$

Denote the optimal objective value $\mu_t(\gamma, \lambda, \nu) = \max_{M \subseteq S} f_t(\gamma, \lambda, \nu, M)$.

26Although this does not satisfy full relative support, we can trivially modify it to satisfy by having $\epsilon$ mass of agents with utilities uniform in $(\cos \theta, \sin \theta)$ with $\theta$ uniform in $[0, 2\pi)$.29
Lemma C.8. $\mu_t(\gamma, \lambda, \nu)$ is convex.

of Lemma C.8 Given $\gamma', \lambda', \nu'$ and $\gamma'', \lambda'', \nu''$. If $(\gamma, \lambda, \nu) = y(\gamma', \lambda', \nu') + (1 - y)(\gamma'', \lambda'', \nu'')$ for $y \in [0, 1]$, then if $M^* \in \arg \max_{M \subseteq S} f_t(\gamma, \lambda, \nu, M)$, then

$$
\mu_t(\gamma, \lambda, \nu) = f_t(\gamma, \lambda, \nu, M^*)
= y f_t(\gamma', \lambda', \nu', M^*) + (1 - y) f_t(\gamma'', \lambda'', \nu'', M^*)
\leq y \mu_t(\gamma', \lambda', \nu') + (1 - y) \mu_t(\gamma'', \lambda'', \nu'')
$$

Therefore, the dual can be written as a convex program with $|T| + |S| + k$ non-negative variables, objective $\min C\gamma + m \cdot \lambda + \sum_{t \in T} \mu_t(\gamma, \lambda, \nu)$ and a single linear constraint is $\sum_{t \in T} \nu_t \geq 1 - \alpha$. One difficulty is that the optimal menu sub-problem needs to optimize over all possible menus $M \subseteq S$, which are exponentially many. However, when preferences are multinomial-logit, we can efficiently solve the sub-problem.

**Proposition C.9.** Under logit utility priors, if $\omega_t + \nu_t > 0$, then the number of optimal solutions for the optimal menu sub-problem is at most $|S|$, and can all be found in time $|S| \log |S|$.

**Remark** If $b_t = 0$, then the utility is deterministic, and we simply return the single-item menus for the items $s$ with highest $(\omega_t + \nu_t) a_{ts} - n_t (t \in T) \lambda_s + \gamma B_{ts}$. Assume from now on that $b_t > 0$. Let $h_s = \frac{n_t (\lambda_s + \gamma B_{ts})}{(\omega_t + \nu_t) b_t}$, $z_s = \exp(h_s)$. The optimal menu sub-problem is equivalent to finding all solutions to

$$
\max_{M \subseteq S} \log(\sum_{s \in M} z_s) - \frac{\sum_{s \in M} h_s z_s}{\sum_{s \in M} z_s}.
$$

Consider the continuous relaxation of this, in which $y_s$ is a continuous variable constrained to be in $[0, z_s]$ and there are $|S|$ such variables:

$$
\max_{y_s \in [0, z_s]} \log(\sum_{s \in S} y_s) - \sum_{s \in S} h_s y_s.
$$

Now if $h_s < h_{s'}$ and $y_{s'} > 0$ but $y_s < z_s$, then we can decrease $\sum_s h_s y_s$ while keeping $\sum_s y_s$ the same, so this cannot occur at an optimum. Relabel services so that

$$
h_1 \leq h_2 \leq \cdots \leq h_{|S|}.
$$

We first consider the case in which the $\{h_s\}$ are all distinct. In this case, by the above, an optimal solution of the continuous relaxation must be of the form: for some $1 \leq m \leq |S|$,

$$
y_s = z_s \quad \forall s < m, \quad y_m \in [0, z_m], \quad y_s = 0 \quad \forall s > m.
$$
Now, let $d_1 = \sum_{s < m} z_s$, $d_2 = h_md_1 - \sum_{s < m} h_sz_s$. As a function of $y_m$, the objective and its first and second derivatives are

$$g(y_m) = \log(d_1 + y_m) + \frac{d_2}{d_1 + y_m} - h_m,$$

$$g'(y_m) = \frac{1}{d_1 + y_m} \left( 1 - \frac{d_2}{d_1 + y_m} \right),$$

$$g''(y_m) = \frac{1}{(d_1 + y_m)^2} \left( \frac{2d_2}{d_1 + y_m} - 1 \right).$$

Thus, if $y_m \in (0, z_m)$ is an interior optimum, $\frac{d_2}{d_1 + y_m} = 1$, and so the second derivative $g''(y_m) = \frac{1}{(d_1 + y_m)^2} > 0$, which implies that $y_m$ is a strict local minimum. Therefore, the objective is maximized when $y_m \in \{0, z_m\}$. This implies that all optimal solutions are restricted to be of the form $M_m = \{1, \cdots, m\}$ (the services are sorted in increasing order of $h_s$), and so we only need to search through $1 \leq m \leq |S|$. This can be done in $|S| \log |S|$ time as it is a linear search after sorting services in non-decreasing order of $h_s$.

Now if some of the $\{h_s\}$ are equal, then if we collapse them into one service in the continuous relaxation, and the above argument implies that an optimal menu $M$ either contains all of them or none of them. Thus, arbitrarily breaking ties when sorting $h_s$ in non-decreasing order and searching through the $M_m$’s still yields all optimal solutions.

The proof of Proposition C.9 reveals insight on what the optimal solution looks like with a logit utility model: based on “Lagrangian multiplier” shadow cost vector $\gamma$ for the budgets and shadow cost $\lambda_s$ for capacity of service $s$, the algorithm places a virtual “allocation cost” $1(t \in T_s)\lambda_s + \gamma B_{ts}$ on allocating an agent of type $t$ to service $s$. Services are put into the agent’s menu starting from the cheapest “allocation costs,” so that an agent is never able to access a service with higher allocation cost (the more over-demanded, “expensive” services) without being able to access a service with lower allocation cost (the less over-demanded, “cheaper” services). For type $t$, there is an “optimal” $m$ number of services to include, and this is chosen by balancing expected allocation costs with expected utility, with the weight on expected utility $\alpha w_t + \nu_t$ depending on how “important” this type is for the objective. The essence of the optimization is finding a set of choice menus that are desirable for the agent but that cause low strain to the system in terms of the capacity and budget limits.

Since the sub-problems are efficiently solvable, we can efficiently solve the dual. If $M_t$ is the solution set to the optimal menu sub-problem using optimal dual variables, then an optimal primal feasible solution can be recovered using complementary slackness by finding a feasible solution to
the polynomial sized LP:

\[
\sum_{M \in \mathcal{M}_t} v_t(M) z_t(M) \geq y \quad \forall t \in T \text{ with equality if } \nu_t > 0
\]

\[
\sum_{M \in \mathcal{M}_t} z_t(M) = 1 \quad \forall t \in T
\]

\[
\sum_{t \in T_s} \sum_{M \in \mathcal{M}_t} n_t p_t(s, M) z_t(M) \leq m_s \quad \forall s \in S \text{ with equality if } \lambda_s > 0
\]

\[
\sum_{s, t} \sum_{M \in \mathcal{M}_t} n_t p_t(s, M) B_{ts} z_t(M) \leq B \quad \forall 1 \leq j \leq k \text{ with equality if } \gamma_j > 0
\]

\[
z_t(M) \geq 0 \quad \forall t \in T, M \in \mathcal{M}_t
\]