Optimal Allocation without Money: an Engineering Approach*

Itai Ashlagi
MIT Sloan School of Management, Cambridge, MA 02139, iashlagi@mit.edu

Peng Shi
MIT Operations Research Center, Cambridge, MA 02139, pengshi@mit.edu

We study the allocation of heterogeneous services to agents without monetary transfers under incomplete information. Agents have private, multi-dimensional utilities over services, drawn from commonly known priors. The social planner's goal is to maximize a possibly complex public objective. For tractability, we take an "engineering" approach, in which we solve a large market approximation, and convert the solution into a feasible finite market mechanism that still yields good results.

We apply this framework to real data from Boston to design a mechanism that assigns 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 a finite market mechanism, which significantly outperforms the baseline plan chosen by the city in terms of efficiency, equity, and predictability.

Key words: market design; priors; assignment; school choice; optimal design; large market approximation

1. Introduction

In many settings, goods or services are allocated to agents without the use of monetary transfers. Examples include the allocation of seats in public schools, spaces in college dorms or courses, and positions in medical residency programs. Social planners' concerns are often multifaceted, including possibly social welfare, equity and system costs.

One example of such a problem was faced by Boston in the 2012-2013 school assignment reform. Seats in Boston Public Schools (BPS) have historically been allocated as follows: the city was divided into three "zones" and each family submitted a ranked list of preferred schools within an individualized menu of choices, which depended on which of three zones the family lived in and which schools were within a one mile radius of the family's home; a centralized algorithm allocated based on priorities and lottery numbers. These large menus resulted in unsustainably high busing costs, representing 10% of total school board budget (Russell and Ebbert (2011)). In 2012, a city committee was charged with the task of reducing the choice menus, while maintaining sufficient variety of choice and equity between various neighborhoods. The outcome of the reform was based on a simulation analysis, which used historical choice data to fit a utility model, and evaluated

* We thank Itay Fainmesser, Steve Graves and Ozalp Ozer for helpful discussions. Ashlagi acknowledges the research support of the National Science Foundation grant SES-1254768.

a shortlist of proposed plans on a portfolio of metrics. However, the plans in the shortlist were proposed based in an ad-hoc manner. This paper studies whether we can use the same inputs (the utility model and the city's objectives) to optimally design the allocation in a systematic way.

When agents' preferences are publicly known, the allocation reduces to an optimization problem. This paper studies how to allocate when preferences are privately known. There are multiple kinds of services, and agents have private, multi-dimensional utilities over services, drawn from common knowledge priors, which may depend on agents' observable information. Since designing multi-dimensional mechanisms is traditionally difficult especially with general objectives, we take an "engineering" approach: first solve a simpler, large market model. Then convert it to a feasible finite market mechanism and evaluate it by simulating with real data.

In the large market model, there are finitely many "types" of agents, and a continuum of agents of each type. "Types" in this paper represent the public information of agents. For example in school choice, a type may represent students from a certain neighborhood of a certain race or socio-economic status. We require mechanisms to be incentive compatible and Pareto optimal among agents with the same type, and we refer to such mechanisms as *valid* mechanisms. While agents of the same type are treated symmetrically, agents of 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.

We first characterize all valid mechanisms. Under mild assumptions over utility priors, we show that any valid mechanism can be described as a collection of Competitive Equilibria with Equal Income (CEEI), which we also refer to as "type-specific-pricing." More precisely, agents of each type are given "virtual prices" for probabilities to each service, and the allocation can be interpreted as giving agents one unit of "virtual money" and allowing them to "purchase" their preferred bundle of probabilities to services (a related mechanism was introduced by Hylland and Zeckhauser (1979)). 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 decision variables being the virtual prices.

In many contexts, only relative preferences are elicited, but not preference intensities. For example, in school choice systems in Boston and New York City, children submit rankings over schools, but not how much they prefer a school over another. In the National Residency Matching Program, doctors submit rankings over residency programs and vice versa. 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 optimality within type.

Under mild regularity assumptions, we show 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.[1] 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 preference rankings over schools, and a centralized mechanism uses submitted preferences, pre-defined priorities and a random lottery number given to each student to determine the assignment.[2] Given ex-post lottery number cutoffs at each school for each priority class of students, this is analogous to the lottery-plus-cutoff mechanism. Notice, however that in contrast to our model all of these markets are finite.

A major technical contribution of this paper is to efficiently find the optimal large market ordinal mechanism in an empirical relevant environment. By relying on the theoretical characterization, one can encode the optimal ordinal mechanism in the large market model by an exponential sized linear program. We show that this can be efficiently solved by considering the dual, which can be decomposed into a collection of "optimal menu" sub-problems, which in turn can be solved efficiently when utilities are based on a multinomial-logit discrete choice model.

To demonstrate the relevance of our large market model, we use our methodology to optimally design school choice in Boston in a systematic way. We take the mechanism chosen by the city, which we refer to as the *Baseline* mechanism, and compute its expected average busing distance per student. Then we seek to optimize a linear combination of utilitarian welfare and max-min welfare using the same amount of busing. All of the analyses use real data from Boston Public Schools (BPS).

Although the school choice problem is defined as a finite market problem, we define a *large market approximation* and use our theoretical and computational results to first find the optimal mechanism in this large market model, which is encoded by a set of lottery cutoffs for each type of students. Using these optimal (large market) cutoffs, we design the corresponding menus and priorities and use the well-known Deferred Acceptance (DA) algorithm (Gale and Shapley (1962))

[1] See, e.g., Sönmez and Ünver (2010) and Budish and Cantillon (2012).

[2] For more information, see Abdulkadiroğlu and Sönmez (2003), Abdulkadiroğlu et al. (2006) and Abdulkadiroğlu et al. (2009).

to define a feasible finite market mechanism that can be seen as "asymptotically optimal."[3] We evaluate this 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 increases average utility by an amount equivalent to decreasing students' average distance to schools by 0.5 miles, and it improves the minimum by about 2.5 miles. This is significant since the Baseline only improves over the most naive plan in average utility by 0.6 miles and in minimum utility by 1.7 miles, so we effectively double the gains. Furthermore, the improved mechanism increases students' chances of getting their first choice by an additive gain of 15%.

Our results yield several insights. The characterizations imply that when the market is large, social planners may restrict attention to a few types of mechanisms observed in practice: "virtual auctions" in the cardinal setting, or "Deferred Acceptance" with menus, priorities and lottery numbers in the ordinal setting. Despite the lack of monetary transfers, a mechanism can still optimize an allocation using differentiated priorities. Such optimization may yield large benefits: in the school choice case we were able to simultaneously improve social welfare, max-min welfare, and predictability, while staying within the same transportation budget. Examining the optimal mechanism found in our empirical exercise shows that it exhibit a quality/quantity trade-off: the optimized plan offers less popular schools to larger areas to attract idiosyncratic preferences, a pattern also seen in the Baseline mechanism chosen by the city.

1.1. Related Literature

Our work connects three strands of previous research. The first is the matching literature, which traditionally focuses on designing mechanisms that satisfy certain properties, such as Pareto efficiency, various fairness conditions, and strategyproofness (see, e.g. Roth and Sotomayor (1990), Abdulkadiroglu and Sönmez (2010)). 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), which also arises in our characterization of valid cardinal mechanisms. Bogomolnaia and Moulin (2001) study ordinal mechanisms that satisfy an ordinal notion of Pareto optimality called ordinal efficiency, and propose a mechanism called Probabilistic Serial, which Che and Kojima (2011) show is asymptotically equivalent in the large market to the more widely 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 with the Probabilistic Serial in the limit. This is

[3] 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 converges to the large market optimum.

analogous to our characterization of valid ordinal mechanisms in the large market, but we consider a more general environment with heterogeneous agent types. Abdulkadiroğlu and Sönmez (2003), Abdulkadiroğlu et al. (2009) and Abdulkadiroğlu et al. (2006) apply matching theory to school choice, and their work has 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.[4] However, the matching literature hardly assumes priors on agents utilities (especially not asymmetric priors), and do not seek to optimize a global objective.[5] Our work can be viewed as bridging the matching literature and the mechanism design/auction literature.[6] In particular, considering prior information may have a significant impact on the design of the mechanism once concerns other than efficiency are considered.

Another strand is optimal mechanism design without money in the finite market framework.[7] Miralles (2012) tackles the multi-dimensional case by considering the Bayesian optimal cardinal mechanism with two services and many symmetric bidders whose utility priors are symmetric across the two services. 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 services, 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 services, so it is not "truly" multi-dimensional and does not generalize to more than two services. Our work shows stronger results by leveraging a large market approximation. Hoppe et al. (2009), 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 allows 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 priors satisfy 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

[4] More recent work on school choice includes Pathak and Sethuraman (2011), Erdil and Ergin (2008), Abdulkadiroglu et al. (2010), and Echenique and Yenmez (2012).

[5] Some exceptions include Ehlers and Massó (2007), Niederle and Yariv (2009).

[6] See e.g. Myerson (1981) who models agents' preferences with Bayesian priors, and Budish (2012) who compares matching and standard mechanism design. He emphasizes the absence of heterogeneous priors and the absence of global objectives in the matching literature.

[7] For a survey, see Schummer and Vohra (2007).

extended to multi-dimensional preferences[8], which is the more realistic assumption in settings such as school choice, where there are multiple types of services. Gershkov et al. (2013) find the optimal incentive compatible mechanism without transfers (or costly signals) in a social choice setting with cardinal, single crossing utility functions. In their setting preferences are also single-dimensional.

A third strand of related research is concerned with 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 the Industrial Organization literature. (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 conditions in which with with a continuum of agents, 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. 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 allocations being open, which in our case is closed. Azevedo and Leshno (2012) study matching markets with a continuum of agents, but contrary to our result they do not consider a global optimization. While continuum models often provide cleaner results, computing the actual mechanism may still be hard. This paper contributes to this literature by actually computing 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 and a mass n_t of agents for each type $t \in T$. In contrast to mechanism design convention, our notion of "type" does not denote the agent's private information, but rather her public information. For example, the type of an agent may be the neighborhood of the agent in school choice, the program or year of study in course allocation, etc. There is a finite set S of services. Every agent must be allocated exactly one service. (Outside options can be accommodated by including a "null service" that represents the outside option.) However, allocations might be probabilistic, so the set of possible allocations for each agent is the probability simplex,

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

[8] 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).)

For agents of type t, their utilities for various services are distributed according to a continuous[9] 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 $A \subseteq U$, $F_t(A)$ 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 cardinal *mechanism* $\mathbf{x}$ is a collection of *allocation rules* $\mathbf{x}_t$ for each type, where each $\mathbf{x}_t$ is a mapping from reported utilities to a possible allocation, $\mathbf{x}_t : U \to \Delta$, and is measurable with respect to F_t. An allocation rule is *incentive compatible* if it is in the agent's best interest to report the truth:

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

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

$$\int_U \mathbf{x}'_t(\mathbf{u}) dF_t = \int_U \mathbf{x}_t(\mathbf{u}) dF_t$$

and $\mathbf{x}'_t$ is weakly preferred by all agents

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

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) within type. Requiring IC is without loss of generality by the revelation principle, as long as we assume that agents' plays are in equilibrium. We require PE as a "stability" criterion: our setup implicitly assumes that the social planner must treat agents within a given type symmetrically, without the ability to discriminate based on the exact identity of the agent; so it may be unreasonable to enforce that agents of the same type cannot trade among themselves post-allocation. Hence, we desire that the mechanism "foresees" any such trades and incorporates the non-existence of Pareto improving trades within each type 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 $\mathbf{x}$ that maximizes his own objective function $W(\mathbf{x})$, subject to all the allocation rules being valid. For now, we allow the objective function $W(\mathbf{x})$ to arbitrarily depend on all the allocation rules $\mathbf{x}_t$, hence allowing it to incorporate agents' welfare, capacity constraints, differential

[9] This assumption can be relaxed, but we choose to adopt it to simplify analysis.

costs in providing various services to various types of agents, and other complex considerations. As an examples of how such performance metrics can be represented in terms of the mechanism $\mathbf{x}$, observe that the expected utility of an agent of type t is

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

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

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

Using this we can model a hard capacity limit m_s on service s by setting $W(\mathbf{x})$ to be negative infinity when $q_s > m_s$. Alternatively, we can model a smooth penalty for exceeding capacity by subtracting a penalty term $C(\max\{0, q_s - m_s\})$ from 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 $\mathbf{x}_t$ 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 for an unit of artificial currency, such that the allocation is what agents would buy if they had 1 unit of artificial currency and were offered probabilities to various services at these prices:

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

Figure 1 illustrates a CEEI with 3 services.

The price vector $\mathbf{a}$ 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 $\mathbf{x}_t : U \to \Delta$.

Since everyone must be assigned somewhere, only relative preferences matter, and a utility report $\mathbf{u}$ gives the same information if all coordinates were changed by the same additive constant. To take away this extra degree of freedom, let $D = \{\mathbf{u} \in \mathbb{R}^n : \mathbf{u} \cdot \mathbf{1} = 0\}$. This is the subspace normal to the all one's vector, and it represents the directions in which utility reports are informative. Given a preference report $\mathbf{u}$, we call the projection of $\mathbf{u}$ onto D the *relative preference*. Given any set $A \subseteq D$, define $U(A)$ to be the subset of U whose projections onto D is in A.

Our result requires one mild regularity condition on the distributions F_t, which says that a-priori, an agent's relative preference could with positive probability take any direction in D. This is used

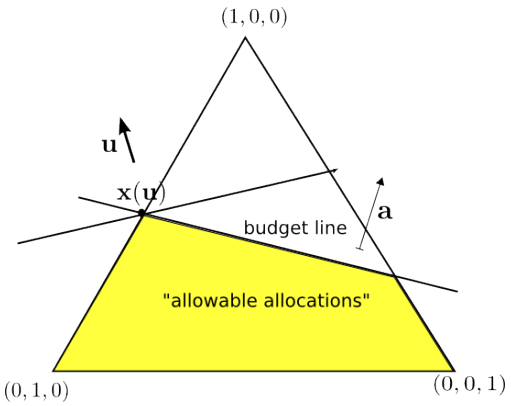


Figure 1 **Illustration of Competitive Equilibrium from Equal Incomes (CEEI) with $|S| = 3$. The triangle represents the space of possible allocations Δ. The shaded region is $\{\mathbf{p} \in \Delta : \mathbf{a} \cdot \mathbf{p} \leq 1\}$. This represents the "allowable allocations" for this type, which is the convex hull of $\{\mathbf{x}_t(\mathbf{u})\}$; she could obtain allocations in the interior of this region by randomizing over several reports. For utility report $\mathbf{u}$, the agent receives an allocation $\mathbf{p}$ that maximizes expected utility $\mathbf{u} \cdot \mathbf{p}$ subject to $\mathbf{p}$ being in the "allowable region." This corresponds to having price vector $\mathbf{a}$ and budget 1, and the agent can "purchase" any allowable allocation subject to the budget constraint.**

in our analysis to guarantee "trade": for any "trading direction," there is a positive measure of agents who would improve by moving in that direction.

DEFINITION 1. (Full relative support) Let D be the set of relative preferences. Define the set of normalized relative preferences, $\tilde{D} = \{d \in D, \|d\| = 1\}$, where $\|\cdot\|$ is the Euclidean norm. This is a sphere in $(|S|-1)$-dimensional space and can be endowed with the topology of a $(|S|-2)$-sphere. This induces a topology on the set of cones[10] $C \subseteq D$ by defining C as open if and only if $C \cap \tilde{D}$ is open in $\tilde{D}$. Distribution F_t has *full relative support* if for every non-empty open cone $C \subseteq D$, $F_t(U(C)) > 0$.

THEOREM 1. *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 Competitive Equilibrium from Equal Incomes (CEEI) with some price vector $\mathbf{a} \in (0, \infty]^{|S|}$.*

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.[11] This is a major impediment to the search for positive theoretical results in multi-dimensional mechanism design.

[10] A cone is a set C in which $x \in C$ implies $\lambda x \in C \ \forall \lambda \in (0, \infty)$.

[11] The condition is called "cyclic monotonicity." (Rochet (1987))

However, in our setup without monetary payments, incentive compatibility simply becomes requiring that the set $X_t = \{\mathbf{x}_t(\mathbf{u})\}$ lies on the boundary of its convex hull, and moreover that $\mathbf{x}_t(\mathbf{u})$ maximizes the linear function $\mathbf{u} \cdot \mathbf{x}$ over this convex hull. This yields a correspondence between an incentive compatible $\mathbf{x}_t$ and a convex set. Any incentive compatible allocation rule maps to a unique convex set, and any convex set corresponds to an incentive compatible rule, which is given by the optimal solution of maximizing the linear functional $\mathbf{u} \cdot \mathbf{x}$ over the set. Label the convex set that corresponds to the allocation rule 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 the feasibility simplex, then we are done since this hyperplane can be represented by a price vector. If there are two such hyperplanes that yield different points of tangency with X_t, then we show that there is a "trading direction" $\mathbf{d}$ by which some positive measure of agents may move allocations in direction $\mathbf{d}$ and others in direction $-\mathbf{d}$ 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 carry out the trade 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 (as opposed to cardinal mechanisms which elicit preference intensities). 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 unit simplex Δ, the space of probability vectors over the $|S|$ 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 utilities are private knowledge. Since we assumed that F_t is continuous, preference rankings are strict with probability one.

Let Π 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 π, in the sense that the utilities are ranked according to the permutation:

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

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

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

$$\mathbf{x}_t(\pi) \in \arg\max_{\pi' \in \Pi} \mathbf{u} \cdot \mathbf{x}_t(\pi').$$

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

$$\int_\Pi \mathbf{x}'_t dF_t = \int_\Pi \mathbf{x}_t dF_t,$$

but $\mathbf{x}'_t$ always first-order stochastically dominates $\mathbf{x}_t$, which means that $\forall \pi \in \Pi$, $\forall 1 \le k \le |S|$,

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

and the inequality is strict for some k and some π 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 $\mathbf{x}$. The objective is to optimize an arbitrary function of the mechanism, $W(\mathbf{x})$, subject to each $\mathbf{x}_t$ 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 as `Uniform`$[0,1]$, and there is a type-specific lottery cutoff for each service; an agent is "admitted" to a service if and only if 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 2. An ordinal allocation rule $\mathbf{x} : \Pi \to \Delta$ is *lottery-plus-cutoff* if there exists "cutoffs" $a_s \in [0,1]$ such that

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

Analogous to full relative support, we define a regularity condition on the utility distribution, which says that every choice ranking $\pi \in \Pi$ is possible.

Definition 3. (Full ordinal support) F_t satisfies full ordinal support if $F_t(\pi) > 0$ for every preference ranking $\pi \in \Pi$.

Theorem 2. *For a given type, suppose its utility prior F induces strict preference rankings with probability one and has full ordinal support, and let $\mathbf{x}(\pi)$ be any incentive compatible and ordinal efficient allocation rule, then $\mathbf{x}(\pi)$ is lottery-plus-cutoffs for some cutoffs $\mathbf{a} \in [0,1]^{|S|}$.*

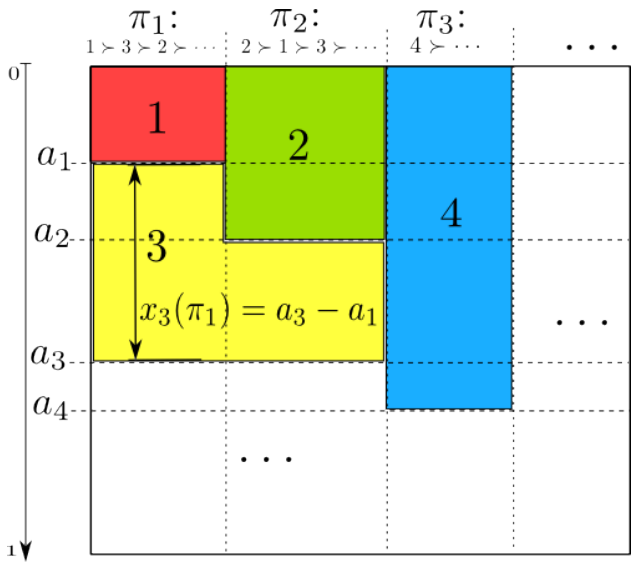


Figure 2 **Illustration of "lottery-plus-cutoff". The vertical axis represents lottery numbers, which are uniformly distributed from 0 to 1. The dotted lines are lottery cutoffs for this type. The columns represent various preference reports. For preference report π_1, the allocation probability for service 3 is the difference $a_3 - a_1$, which represents lottery numbers for which the agent is not admitted to her first choice of service 1 but is admitted to her second choice of service 3.**

The proof is given in Appendix C.[12] Recall that ihe intuition behind it is similar to Theorem 1. In the cardinal case, incentive compatible allocation rules are associated with arbitrary convex subsets of Δ. Here in the ordinal case, instead of a convex set we have a polymatroid. More precisely, for any incentive compatible ordinal allocation rule $\mathbf{x}_t$, the set $\{\mathbf{x}_t\}$ 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 dominate their current allocations, 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 $\mathbf{a}$. 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 kth in this

[12] A similar theorem appears in Ashlagi and Shi (2013), but the setting is slightly different here and we give a different proof.

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 lottery-plus-cutoff. Hence, we also give lottery-plus-cutoff the name *randomized menus with nested menus*.

3.2. Comparing Cardinal and Ordinal Mechanisms

The proofs of Theorems 1 and 2 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* 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. Intuitively, the value of a cardinal mechanism over an ordinal mechanism lies in its ability to differentiate agents with extreme preference intensities. So if agents' preferences for various services are of similar relative intensities, then we would expect ordinal to perform well compared to the cardinal; if preferences exhibit extremely heterogeneous relative intensities, then we would expect the cardinal to outperform.

We give an example in Appendix C.3 that shows with unrestricted utilities, the ratio between the optimal social welfare achievable from a valid cardinal mechanism and the optimal social welfare achievable from 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, which was based on simulating a list of potential plans and evaluating based on a given portfolio of metrics. 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.[13] We recognize that in order to produce implementable recommendations, the precise objective function and constraints must be scrutinized and debated over by all stakeholders and constituents, which has not yet taken place.

[13] In the actual problem faced by the city committee, one needs to consider several additional complications: grandfathering of the previous plan for students in the transitional period, handling specialized programs for English 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 refined modeling to future work.

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 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 each year apply to Boston Public Schools (BPS) for the grade of Kindergarten 2 (K2), which is the main entry grade to elementary schools. The social planner (the city in this case) 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.[14] 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 these students are distributed among the geocodes of that neighborhood according a multinomial distribution with probabilities matching the historic average in years 2010-2013.

[14] 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.

Each student is to be assigned to one of 77 schools. The distribution of students across geocodes and the capacities of schools are plotted in Figure 4a. The capacities are the actual numbers from 2013.

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

$$u_{is} = \text{quality}_s - \text{Distance}_{ts} + \omega \text{Walk}_{ts} + \beta \epsilon_{is}$$

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 the equivalent of "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.[15] ω 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. ϵ_{is} is an idiosyncratic taste shock assumed to be i.i.d. standard Gumbel distributed, and β represents the strength of the idiosyncratic taste shock.[16] Variables starting with capital letters, Distance_{ts} and Walk_{ts}, are directly from data, while variables in lower case, quality_s, ω, and β are estimated via maximum likelihood using historical choices. We plot the school qualities in Figure 4b and tabulate the other coefficients in Table 1. Although more sophisticated demand models are possible, in this mechanism optimization exercise, 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 minimum welfare,

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

where v_t is the expected utility of a student from geocode t, w_t is the proportion of all students who live in geocode t (taking the expected number of students from each geocode and normalizing so the weights sum to 1), and α is a parameter specifying the desired trade-off between efficiency

[15] 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.

[16] 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.

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 in fitting the utility model as well as K2 data since families face similar choices in the two grades and more data allows greater precision.) The values can be interpreted in units of miles: how many miles a student is willing to travel for one unit of this variable.

Parameter	Value	Interpretation
quality_s	0–6.29	Quality of schools. For a school of Δq additional quality, holding fixed other components, a student would be willing to travel Δq miles further. These values are graphically displayed in Figure 4b. We normalize the smallest value to be 0.
ω	0.86	Additional utility for going to a school within walk-zone.
b	1.88	Standard deviation of idiosyncratic taste shock.

and equity. $\alpha = 1$ represents maximizing the average expected utility; $\alpha = 0$ represents maximizing the expected utility for the worst-off geocode; $\alpha = .5$ represents an equal weighting of the two.

For each school s, there is a capacity limit m_s, which is the number of seats available at the school for K2 students. Moreover, for 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 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 an unlimited number of students inside 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.[17] 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 (busing is needed only for students outside a 1-mile walk-zone). Suppose that the social planner budgets C miles of busing per student in expectation, then the busing constraint is

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

where p_{ts} is the probability that a random student in geocode t is assigned to school s.[18] In reality, busing cost is much more complicated, having to do with the routing, the number of buses used,

[17] This is only approximately true in reality. Although BPS has committed to expanding the capacity schools as needed by adding hiring new teachers and adding modular classrooms, in reality there are hard space constraints and BPS uses a more complicated, ad-hoc system for guaranteeing that each student is assigned, which is based on distance and many other factors. In BPS literature, capacity schools were later renamed "option schools."

[18] 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.

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 Pareto 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.[19] *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 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.[20] One of the 14 neighborhoods is East Boston. In the assignment plan, East Boston students get priority for East Boston schools, while non-East Boston students get priority for non-East Boston schools.[21] We encode this using auxiliary variable h_{ts}, which is an indicator for whether geocode t and school s are both in East Boston or both outside of East Boston.

[19] For more details of that reform, see Abdulkadiroğlu et al. (2006).

[20] 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, our priority structure is simpler.

[21] The reason is that East Boston and the rest of Boston are separated by water, and only connected by a few bridges and tunnels, so it may be inconvenient to traverse across.

Each student i is given a random lottery number distributed Uniform$[0,1]$. Her *score* to school s is defined as $\sigma_{is} = r_i - h_{ts}$, where r_i is a Uniform[0,1] random variable, independently drawn across i's. 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 within her menu.
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 within her menu.
4. Iterate steps 1-3 until all unassigned students have empty choice rankings.[22]

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.[23]

We simulate this plan 10000 times according to the assumptions described in Section 4.1, and tabulate the plan's transportation burden, average expected utility, expected utility of the worst-off type, and predictability measures in Table 3, under the column "Baseline."

4.3. Solving the Large Market Approximation

We define the *large market approximation* to the model in Section 4.1 as follows. Replace each agent with a continuum of infinitesimal agents 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 the 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 be negative infinity. Moreover, ex-post Pareto efficiency is equivalent to ordinal efficiency in the large market setting.[24]

By the characterization result in Section 3.1, it suffices to consider mechanisms that are randomized menu with nested menus, and as a relaxation it suffices to consider randomized menu mechanisms. (Such mechanisms offer menu $M \subseteq S$ to type t with probability $z_t(M)$. Given her menu, an agent picks her most preferred school in her menu.) For a menu of services $M \subseteq S$, abuse

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

[23] See Roth and Sotomayor (1990) and Abdulkadiroğlu and Sönmez (2003)

[24] For works that study this equivalence, see Che and Kojima (2011) and Liu and Pycia (2012).

notation slightly and let $v_t(M)$ denote the expected utility of the best service in this menu for an agent of type t,

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

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 $M \subseteq S$,

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

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:

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

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 a special structure, then an optimal solution to this LP can be found in time polynomial in $|T|$ and $|S|$.

Definition 4. Utility prior F_t is *multinomial-logit* if the utilities can be written as

$$u_{is} = \bar{u}_{ts} + b_t \epsilon_{is},$$

where $b_t > 0$ and $\bar{u}_{ts}$ are given parameters, and ϵ_{is}'s are i.i.d. standard Gumbel distributed.

Note that the utility distribution in our model is *multinomial-logit*. This implies that $v_t(M) = b_t \log(\sum_{s \in M} exp(\frac{\bar{u}_{ts}}{b_t}))$, and $p_t(s, M) = \frac{exp(\frac{\bar{u}_{ts}}{b_t})}{\sum_{s' \in M} exp(\frac{\bar{u}_{ts'}}{b_t})}$.

Theorem 3. *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 computation involves taking the dual of the LP, which has a small number of variables but an exponential 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 a small number of variables but exponentially many constraints. The multinomial-logit assumption allows each of

these sub-problems to be solved in $|S|\log|S|$ time, so the whole dual can 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 utilitarian welfare and max-min welfare, so for the remainder of this paper we use $\alpha=0.5$.

Table 2 Performance in the large market model of the optimal plans under various choices of α.

	$\alpha=1$	$\alpha=0.5$	$\alpha=0$
Utilitarian welfare	7.78	**7.66**	7.39
Max-min welfare	2.52	**7.39**	7.39

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 $\sigma_{is}=r_i-a_{ts}$, where r_i is Uniform[0,1] random variable, independent across the i's. 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, after running the DA analog described above, a student i is assigned to school s if and only if her score is negative. In this case, each student's assignment probabilities become identical to the probabilities she would have gotten in the large market approximation. Thus, this finite market mechanism is "asymptotically optimal" as the market is scaled up independently. (The independent scaling also removes the stochasticity in number of students of each type.)

Observe that the converted mechanism remains incentive compatible and ex-post Pareto-efficient within type. It is incentive compatible because Deferred Acceptance is strategyproof for the students when the priorities are exogenous, not depending on students' preference submissions. (Note that the cutoffs are estimated using previous years' choice data and a-priori demand modeling and thus students cannot manipulate the cutoffs.) Ex-post Pareto efficiency within type follows from the priorities a_{ts} being the same for students of each type, and from our using for each student the same random number r_i at all schools.[25]

4.5. Numerical Results

Let *ApproximateOpt1* denote the DA analog to the optimal large market mechanism with $\alpha = 0.5$ and a 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, which exceeds the 0.63 miles of Baseline. So we let *ApproximateOpt2* denote the DA analog to the optimal large market mechanism with $\alpha = 0.5$ and a 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 "Most Naive" 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 Most Naive, 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 less miles of busing per student than the Baseline, while improving average utility over the Most Naive by 1.18 miles (almost twice the improvement from Baseline), and improving expected utility of worst-off type by 4.16 miles (more than twice the improvement from Baseline). It also significantly improves students' chances of getting into top 1 and top 3 choices.

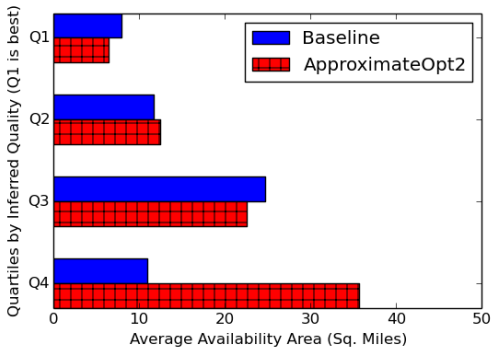
To gain intuition on what ApproximateOpt2 is doing, we compute for each school its "availability area" in each plan. This is the total area of the geocodes for which this school shows up in the

[25] Ex-post Pareto efficiency means there cannot be improvement cycles within the same type. For a fixed type, consider all students of this type and all the schools they are assigned to (counting multiplicity of seats) from Deferred Acceptance. Consider the student of the best (smallest) random number r_i out of these, then this student must be assigned to her most preferred of these schools, otherwise there would be instability (in the two sided matching sense) as all schools prefer this student to others of the same type by the score structure. Hence, this student would take part in any improvement cycle. Removing this seat from our list of schools seats, the student with the second best random number must be assigned to her most preferred of the remaining seats. She cannot get the first student's seat, and so she cannot participate in any improvement cycle. Continuing in this way, we rule out all improvement cycles and get that the final assignment is ex-post Pareto efficient within type.

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

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

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.



menu. We then divide schools into quartiles by their quality score (the term that shows up in the utility model), 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 at risk of being closed.

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 a public 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 compute the optimal ordinal mechanism in an empirically relevant special case. An open question is the efficient computation of the optimal ordinal mechanism with other distributions[26], and the efficient computation of the optimal cardinal mechanism.[27]

In our empirical exercise, we use past demand patterns to inform the city on how to allocate the limited amount of busing between various neighborhoods to maintain efficient and equitable access to schools. (The busing budget is justified as the cost eventually comes back to tax payers, so a limit on busing is needed to control the negative externalities of a family using city resources to travel to far away schools.) If one were to implement this in practice, the optimization of menus and priorities should not be done more often than once every 5-10 years, in order to maintain predictability for families.

One question is whether to conduct an ordinal or cardinal mechanism. We find in a simplified settings that the type of mechanism, whether ordinal or cardinal, may results in a significant different social welfare. Understanding the tradeoff for various distributions of preferences may lead to design implications (see Abdulkadiroglu et al. (2011) and Che and Tercieux (2013) who study the impact of various ordinal mechanisms on efficiency). Even broader is the question of when to conduct centralized mechanisms or allow agents to exert effort, for example by maintaining a queue as often is a given mechanism in the operations literature.

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, especially preference intensities. 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 fundamental question is whether human behavior can indeed be captured with such models.[28] Nevertheless, *if* a utility model can be estimated and trusted, the methods used in this paper can be used to compute the optimal mechanism.

---

[26] Our current techniques reduce this to solving an "optimal-menu subproblem."

[27] Our current techniques reduce this to a polynomial sized non-linear program, and we have not found any interesting distributions for which the structure significantly simplifies.

[28] One project that examines the validity 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 evaluate the prediction accuracy.

References

Abdulkadiroglu, A., Y. Che, Y. Yasuda. 2010. Expanding 'choice' in school choice. Working Papers 10-23, Duke University, Department of Economics.

Abdulkadiroglu, A., T. Sönmez. 2010. Matching markets: Theory and practice.

Abdulkadiroglu, Atila, Yeon-Koo Che, Yosuke Yasuda. 2011. Resolving conflicting preferences in school choice: The boston mechanism reconsidered. *The American Economic Review* **101**(1) 399–410.

Abdulkadiroğlu, A., P. A. Pathak, A. E. Roth. 2009. Strategy-proofness versus efficiency in matching with indifferences: Redesigning the nyc high school match. *American Economic Review* **99**(5) 1954–1978.

Abdulkadiroğlu, A., P. A. Pathak, A. E. Roth, T. Sönmez. 2006. Changing the boston school choice mechanism. Boston College Working Papers in Economics 639, Boston College Department of Economics.

Abdulkadiroğlu, A., T. Sönmez. 2003. School choice: A mechanism design approach. *American Economic Review* **93** 729–747.

Ashlagi, I., P. Shi. 2013. Improving community cohesion in school choice via correlated-lottery implementation. Working paper.

Aumann, R. J. 1964. Markets with a continuum of traders. *Econometrica* **32**(1) 39–50.

Azevedo, E., J. Leshno. 2012. A supply and demand framework for two-sided matching markets. Working paper.

Bogomolnaia, A., H. Moulin. 2001. A new solution to the random assignment problem. *Journal of Economic Theory* **100**(2) 295–328.

Budish, E. 2012. Matching versus mechanism design. *ACM SIGecom Exchanges* **11**(2) 4–15.

Budish, E., E. Cantillon. 2012. The multi-unit assignment problem: Theory and evidence from course allocation at harvard. *American Economic Review* **102**(5) 2237–71.

Chakravarty, S., T. R. Kaplan. 2013. Optimal allocation without transfer payments. *Games and Economic Behavior* **77** 1–20.

Che, Y., F. Kojima. 2011. Asymptotic equivalance of probabilistic serial and random priority mechanisms. *Econometrica* **78**(5) 1625–1672.

Che, YK, O. Tercieux. 2013. Efficiency and stability in large matching markets. Working paper.

Condorelli, D. 2012. What money cant buy: Efficient mechanism design with costly signals. *Games and Economic Behavior* **75**(2) 613–624.

Echenique, F., B. Yenmez. 2012. How to control controlled school choice. Working paper.

Ehlers, L., J. Massó. 2007. Incomplete information and singleton cores in matching markets. *Journal of Economic Theory* **136**(1) 587–600.

Erdil, A., H. Ergin. 2008. What's the matter with tie-breaking? improving efficiency in school choice. *American Economic Review* **95**.

Gale, D., L. S. Shapley. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* **69** 9–15.

Gershkov, A., B. Moldovanu, X. Shi. 2013. Optimal voting rules. Working paper.

Hartline, J. D., T. Roughgarden. 2008. Optimal mechanism design and money burning. *Proceedings of the 40th annual ACM symposium on Theory of computing*. STOC '08, ACM, New York, NY, USA, 75–84.

Hoppe, H. C., B. Moldovanu, A. Sela. 2009. The theory of assortative matching based on costly signals. *The Review of Economic Studies* **76**(1) 253–281.

Hylland, A., R. Zeckhauser. 1979. The efficient allocation of individuals to positions. *Journal of Political Economy* **87**(2) 293–314.

Liu, Q., M. Pycia. 2012. Ordinal efficiency, fairness, and incentives in large markets. Working paper.

Miralles, A. 2012. Cardinal bayesian allocation mechanisms without transfers. *Journal of Economic Theory* **147** 179–206.

Myerson, R. B. 1981. Optimal auction design. *Mathematics of operations research* **6**(1) 58–73.

Niederle, M., L. Yariv. 2009. Decentralized matching with aligned preferences. National Bureau of Economic Research.

Pathak, P. A., J. Sethuraman. 2011. Lotteries in student assignment: An equivalence result. *Theoretical Economics* **6**(1).

Pathak, P. A., P. Shi. 2014. Demand modeling, forecasting, and counterfactuals, part i. Working Paper 19859, National Bureau of Economic Research.

Rochet, J. C. 1987. A necessary and sufficient condition for rationalizability in a quasilinear context. *Journal of Mathematical Economics* **16**(2) 191–200.

Roth, A. E., M. Sotomayor. 1990. *Two-sided Matching: a Study in Game-theoretic Modeling and Analysis*. Econometric Society monographs, Cambridge University Press.

Russell, J., S. Ebbert. 2011. The high price of school assignment. *Boston Globe* (12 Jun. 2011).

Schummer, J., R. Vohra. 2007. Mechanism design without money. Noam Nisan, Tim Roughgarden, Eva Tardos, Vijay V. Vazirani, eds., *Algorithmic Game Theory*. Cambridge University Press, Cambridge, UK, 243–266.

Shi, P. 2013. Closest types: A simple non-zone-based framework for school choice. Http://www.mit.edu/ pengshi/papers/closest-types.pdf.

Sönmez, T., M. U. Ünver. 2010. Course bidding at business schools. *International Economic Review* **51**(1) 99–123.

Thomson, W., L. Zhou. 1993. Consistent allocation rules in atomless economies. *Econometrica* **61**(3) 575–587.

Tirole, J. 1988. *The theory of industrial organization*. MIT press, Cambridge, MA.

Zhou, L. 1992. Strictly fair allocations in large exchange economies. *Journal of Economic Theory* **57**(1) 160–175.

Appendix 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);[29]
- the 3 closest capacity schools;[30]
- 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.

Appendix B: Additional Tables and Figures

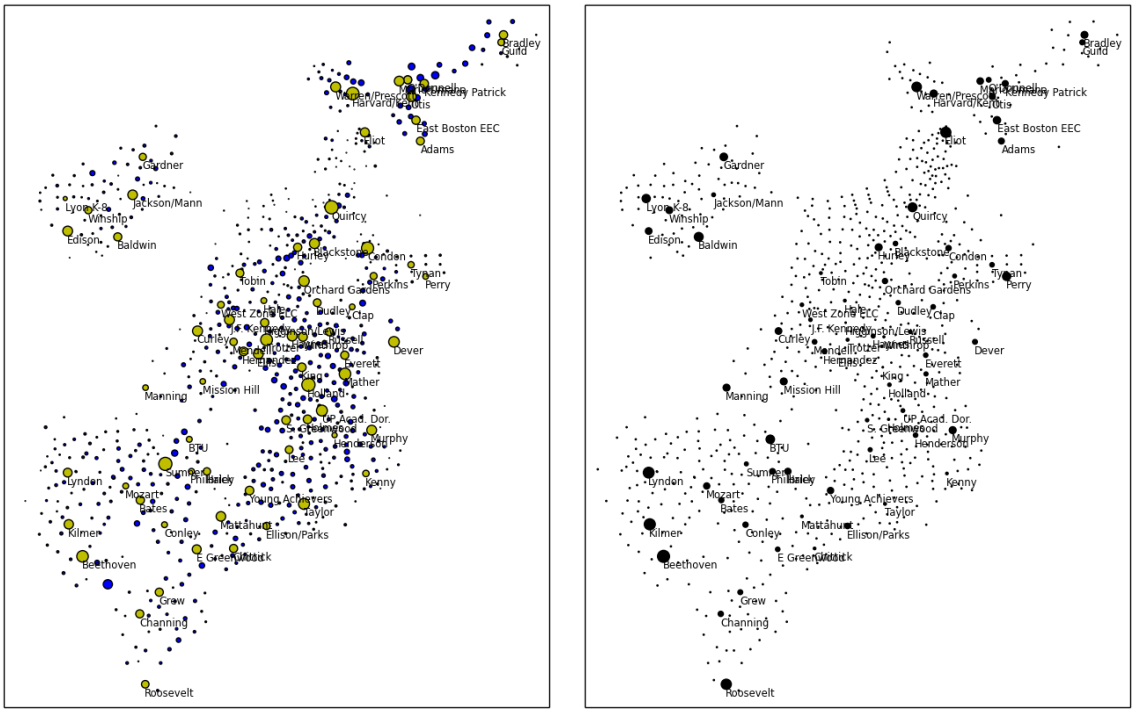
Table 4 shows the forecasted proportion of students applying from each neighborhood. Figure 4a and 4b give a big picture view of the distribution of supply and demand for schools and of inferred school quality in Boston.

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

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

[29] ELCs are extended-day kindergartens.

[30] Recall that capacity schools are those which BPS has committed to expanding capacity as needed to accommodate all students. In the 2014 implementation of the Home Based Plan, for elementary schools, capacity schools are exactly the Tier 4 schools.



(a) Supply and Demand (b) School Quality

Figure 4 **The left shows the distribution of students and capacities 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. The capacities are based on data from 2013. The right shows estimates of quality$_s$ (inferred quality) from the 2013 data. The size of the circle is proportional to the estimated quality$_s$, with higher quality schools having larger circles.**

Appendix C: Omitted Proofs

C.1. Characterization for Cardinal Mechanisms

Proof of Theorem 1. The proof uses a series of lemmas. For clarity of exposition, we first show the main proof, and prove the lemmas later.

Recall that $D = \{\mathbf{u} \in U : \mathbf{u} \cdot \mathbf{1} = 0\}$ is the space of relative utilities. This is the space of informative utilities since everyone must be assigned somewhere.

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

Lemma 1 says that any incentive compatible allocation rule can be represented by a closed convex set X in which $\mathbf{y} = \mathbf{x}(\mathbf{u})$ is a maximizer to the linear objective $\mathbf{u} \cdot \mathbf{y}$ subject to $\mathbf{y} \in X$. This is illustrated in Figure 5.

We proceed to prove the theorem by induction on $|S|$. For $|S| = 1$, there is nothing to prove as Δ is one point. Suppose we have proven this theorem for all smaller $|S|$. Let X be the convex set that corresponds to

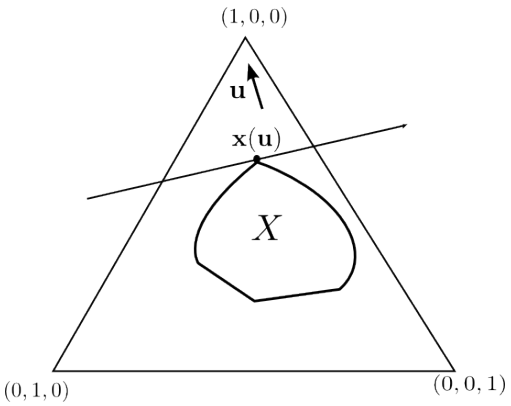


Figure 5 **An incentive compatible allocation rule with $|S|=3$. X is an arbitrary closed convex subset of the feasibility simplex Δ. $\mathbf{y}=\mathbf{x}(\mathbf{u})$ is maximizer of the linear objective $\mathbf{u}\cdot\mathbf{y}$ with $\mathbf{y}\in X$.**

allocation rule $\mathbf{x}$. Suppose X does not intersect the relative interior of Δ, $\texttt{int}(\Delta)=\{\mathbf{y}\in\mathbb{R}^{|S|}:\mathbf{y}>0,\sum_s y_s=1\}$, then some component of $\mathbf{x}$ 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 a smaller number of services, for which the theorem is true by induction. Thus, it suffices to consider the case $X\cap\texttt{int}(\Delta)\neq\emptyset$.

Let $H(\mathbf{u},\alpha)$ denote the $|S|-1$ dimensional hyperplane $\{\mathbf{y}\in\mathbb{R}^{|S|}:\mathbf{u}\cdot\mathbf{y}=\alpha\}$. Let $H^-(\mathbf{u},\alpha)$ denote the half-space $\{\mathbf{y}:\mathbf{u}\cdot\mathbf{y}\leq\alpha\}$, and $H^+(\mathbf{u},\alpha)$ denote $\{\mathbf{y}:\mathbf{u}\cdot\mathbf{y}\geq\alpha\}$. Let $\texttt{aff}(\Delta)$ denote the affine hull Δ, $\texttt{aff}(\Delta)=\{\mathbf{y}\in\mathbb{R}^{|S|}:\sum_s y_s=1\}$. Note that X is a convex subset of $\texttt{aff}(\Delta)$. The following lemma allows us to express tangents of X in $\texttt{aff}(\Delta)$ in terms of a price vector $\mathbf{a}\in(0,\infty)^{|S|}$.

LEMMA 2. *If $X\subseteq\Delta$, any tangent hyperplane of X in $\texttt{aff}(\Delta)$ can be written as $H(\mathbf{a},1)\cap\texttt{aff}(\Delta)$, for some $\mathbf{a}\in(0,\infty)^{|S|}$, with $\mathbf{a}$ pointing outward from X and not co-linear with $\mathbf{1}$. ($\mathbf{a}\cdot\mathbf{y}\leq 1$, $\forall\mathbf{y}\in X$, and $\mathbf{a}\neq\lambda\mathbf{1}$ for any $\lambda\in\mathbb{R}$.)*

For any set $A\subseteq D$, let $U(A)=\{\mathbf{u}\in U:\texttt{Proj}_D(\mathbf{u})\in A\}$. This is the set of utilities for which the projection in D is in A. The average allocation of agents with relative preference in A is

$$\bar{\mathbf{x}}(U(A))=\int_{U(A)}\mathbf{x}(\mathbf{u})dF(\mathbf{u}).$$

Since type-specific-pricing without infinite prices is the same as having $X=H^-(\mathbf{a},1)\cap\Delta$, it suffices to show that the convex set X has only one tangent in $\texttt{int}(\Delta)$. Intuitively, if it has two different tangents $H(\mathbf{a},1)$ and $H(\mathbf{a}',1)$, with non-zero and unequal unit projections onto D, then we can find a unit vector $\mathbf{d}\in D$ s.t. $\mathbf{d}\cdot\mathbf{a}>0>\mathbf{d}\cdot\mathbf{a}'$. Since $\mathbf{a}$ and $\mathbf{a}'$ are tangent normals, we can perturb $\mathbf{x}(\mathbf{u})$ in direction $\mathbf{d}$ for $\mathbf{u}$ near $\mathbf{a}$, and perturb $\mathbf{x}(\mathbf{u})$ in direction $-\mathbf{d}$ for $\mathbf{u}$ near $\mathbf{a}'$, thus Pareto improving $\mathbf{x}(\cdot)$ but keeping average $\bar{\mathbf{x}}(U)$ fixed. This is illustrated in Figure 6. 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.

LEMMA 3. *Suppose that $H(\mathbf{a_0},1)$ is an outward pointing supporting hyperplane of X that intersects X in the relative interior of the feasibility simplex, $\texttt{int}(\Delta)$. Then for any unit vector $\mathbf{d}\in D$ such that $\mathbf{d}\cdot\mathbf{a_0}>0$, there exists $\delta_0>0$ such that for all $\delta\in(0,\delta_0)$, there exists allocation rule $\mathbf{x}'$ that strictly dominates $\mathbf{x}$, with $\bar{\mathbf{x}}'(U)=\bar{\mathbf{x}}(U)+\delta\mathbf{d}$.*

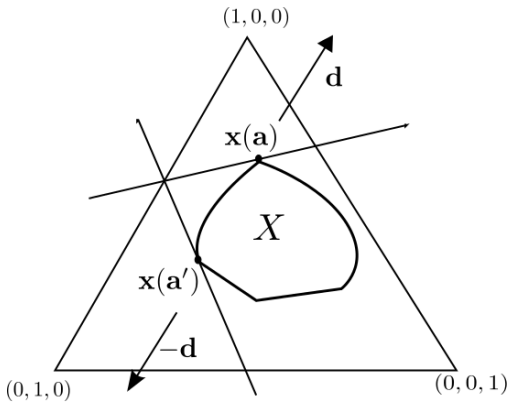


Figure 6 **Exchange argument to Pareto improve the allocation rule by expanding X along opposite directions, when there is more than one supporting hyperplane of X intersecting $\texttt{int}(\Delta)$.**

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

Proof of Lemma 1 Suppose cardinal allocation rule $\mathbf{x}(\mathbf{u})$ is incentive compatible. Let X be the convex closure of its range. Then since $\mathbf{u}\cdot\mathbf{x}(\mathbf{u}) \geq \mathbf{u}\cdot\mathbf{x}(\mathbf{u}') \quad \forall \mathbf{u}' \in U$, we have $\mathbf{u}\cdot\mathbf{x}(\mathbf{u}) \geq \mathbf{u}\cdot\mathbf{y} \quad \forall \mathbf{y} \in X$. So $\mathbf{x}(\mathbf{u}) \in \arg\max_{\mathbf{y}\in X}\{\mathbf{u}\cdot\mathbf{y}\}$.

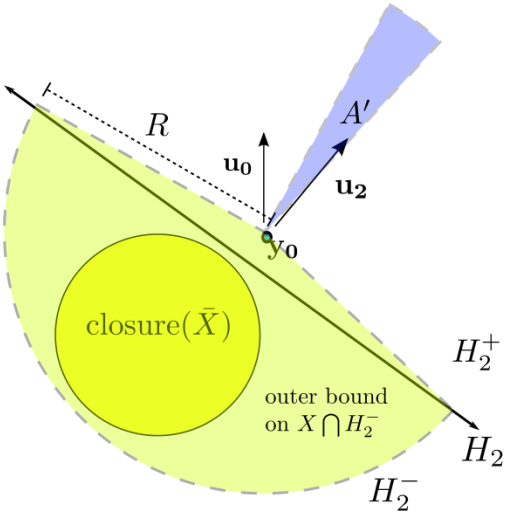
Conversely, if for some closed convex set X, for any $\mathbf{u} \in U$, $\mathbf{x}(\mathbf{u}) \in \arg\max_{\mathbf{y}\in X}\{\mathbf{u}\cdot\mathbf{y}\}$, then $\forall \mathbf{u}' \in U$, $\mathbf{x}(\mathbf{u}') \in X$, so $\mathbf{u}\cdot\mathbf{x}(\mathbf{u}) \geq \mathbf{u}\cdot\mathbf{x}(\mathbf{u}')$. So $\mathbf{x}$ is incentive compatible. □

Proof of Lemma 2 Any tangent hyperplane Y of X in $\texttt{aff}(\Delta)$ is a $|S|-2$ dimensional affine subset of the $|S|-1$ dimensional affine set $\texttt{aff}(\Delta)$. Take an arbitrary point $\mathbf{z} \in \Delta$ on the same side of Y as X. Consider the $|S|-1$ dimensional hyperplane H passing through Y and $(1+\epsilon)\mathbf{z}$. For some sufficiently small $\epsilon > 0$, by continuity, H has all positive intercepts, so $H = \{\mathbf{y} : \mathbf{a}\cdot\mathbf{y} \leq 1\}$ 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 X. □

In carrying out the exchange argument in Figure 6, we need to guarantee that a positive measure of agents benefit from the perturbation in the set X. If the points $x(\mathbf{a})$ and $x(\mathbf{a}')$ occur at a vertex, meaning that a positive measure of agents obtain each of these allocations, then there is nothing additional to show. The difficulty is if X is "smooth" at $\mathbf{a}$ and $\mathbf{a}'$, so we need to do an exchange for agents with utilities in a *neighborhood* of $\mathbf{a}$ and $\mathbf{a}'$, but in that case it is not clear that we can move in directions $\mathbf{d}$ and $-\mathbf{d}$ without going past the boundary of the feasibility simplex and it is not clear that we can obtain utility improvements for all these agents. For example, if the neighborhood is too large, then the move would not work. Lemma 3 is needed to guarantee that we can do this exchange with a positive measure of agents.

The proof of Lemma 3 uses the following rather technical lemma, which guarantees that for any $\delta > 0$, and any $\mathbf{u_0} \in D$, we can find a small open neighborhood of A of $\mathbf{u_0}$ such that the average allocation $\bar{\mathbf{x}}(A)$ is

Figure 7 **Illustration of the construction of cone A' in which any $\mathbf{u} \in A'$ has $\mathbf{x}(\mathbf{u})$ in the open half space H_2^+, so any convex combination of such $\mathbf{u}$ cannot be in closure$(\bar{X})$.**



δ within $\mathbf{x}(\mathbf{u_0})$. Note that for any $\mathbf{u} \in A \backslash \{\mathbf{u_0}\}$, $\mathbf{x}(\mathbf{u})$ itself may not be close to $\mathbf{x}(\mathbf{u_0})$, because $\mathbf{u_0}$ could be normal to the convex set X along a linear portion of X, in which case $\mathbf{x}(\mathbf{u})$ would veer off from $\mathbf{x}(\mathbf{u_0})$ until it reaches the end of the linear portion. But this lemma shows that by taking a convex combination of such $\mathbf{u} \in A$, we can have $\mathbf{x}(A)$ arbitrarily close to $\mathbf{x}(\mathbf{u_0})$.

LEMMA 4. *Given any bounded closed convex set $X \subseteq \mathbb{R}^n$, any non-empty open cone $C \subseteq \mathbb{R}^n$ and any measurable function $\mathbf{x} : C \to \mathbb{R}^n$ such that $\mathbf{x}(\mathbf{u}) \in \arg\max_{\mathbf{y} \in X} \mathbf{u} \cdot \mathbf{y}$. Let F be an atomless measure with $F(C) > 0$ and such that for any non-empty open cone $A \subseteq C$, $F(A) > 0$. Define*

$$\bar{X} = \{\bar{\mathbf{x}}(A) = \frac{\int_A \mathbf{x}(\mathbf{u}) dF(\mathbf{u})}{F(A)} : F(A) > 0, A \subseteq C\}.$$

Then $\forall \mathbf{u} \in C$, $\arg\max_{\mathbf{y} \in X} \{\mathbf{u} \cdot \mathbf{y}\} \subseteq$ closure$(\bar{X})$.

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

$$m(A) = (\int_A \mathbf{x}(\mathbf{u}) dF(\mathbf{u}), F(A)).$$

By Lyapunov's convexity theorem, since F is atomless, the range of this measure, denoted M, is convex. Therefore, the cone generated by M, $cone(M) = \{\lambda \mathbf{x} : \mathbf{x} \in M, \lambda > 0\}$, is convex, and so its intersection with the hyperplane $(\cdot, 1)$ is convex (last component restricted to 1). This intersection is non-empty since $F(C) > 0$. Moreover, this intersection, restricted to first n components is exactly $\bar{X}$, so $\bar{X}$ is convex.

The proof of the lemma proceeds by contradiction. Suppose on the contrary that there exists $\mathbf{u_0} \in C$ and $\mathbf{y_0} \in \arg\max_{\mathbf{y} \in X} \{\mathbf{u_0} \cdot \mathbf{y}\}$ but $\mathbf{y_0} \notin$ closure$(\bar{X})$. We will exhibit some open subset $A \subseteq C$ such that $\bar{\mathbf{x}}(A) \notin \bar{X}$, which contradicts the definition of $\bar{X}$. The construction is geometric and we refer readers to Figure 7 for an illustration.

Since closure($\bar{X}$) is closed, convex and bounded, there exists a strictly separating hyperplane $H(\mathbf{u_1}, \alpha_1)$ such that for some $\delta_1 > 0$,

$$\mathbf{u_1} \cdot \mathbf{y_0} \geq \alpha_1 + \delta_1 > \alpha_1 - \delta_1 \geq \mathbf{u_1} \cdot \mathbf{y} \quad \forall \mathbf{y} \in \text{closure}(\bar{X}).$$

Since every point of $\bar{X}$ is a convex combination of points in X and since X is closed and convex, closure($\bar{X}$) $\subseteq X$, so by construction,

$$\mathbf{u_0} \cdot \mathbf{y_0} \geq \mathbf{u_0} \cdot \mathbf{y} \quad \forall y \in \text{closure}(\bar{X}).$$

Let $\alpha_0 = \mathbf{u_0} \cdot \mathbf{y_0}$. Since C is open, by taking $(\mathbf{u_2}, \alpha_2) = (\mathbf{u_0}, \alpha_0) + \epsilon(\mathbf{u_1}, \alpha_1)$ for some sufficiently small $\epsilon > 0$, we can ensure that $\mathbf{u_2} \in C$, and by construction, $H_2 = H(\mathbf{u_2}, \alpha_2)$ is a strictly separating hyperplane with

$$\mathbf{u_2} \cdot \mathbf{y_0} \geq \alpha_2 + \delta_2 > \alpha_2 - \delta_2 \geq \mathbf{u_2} \cdot \mathbf{y} \quad \forall \mathbf{y} \in \text{closure}(\bar{X}),$$

where $\delta_2 = \epsilon\delta_1$.

Now, let $R = \sup_{\mathbf{y} \in X}\{\|\mathbf{y} - \mathbf{y_0}\|\}$. R is finite because X is bounded. Let H_2^- be the closed half-space on the non-positive side of H_2. If $\mathbf{y} \in X \bigcap H_2^-$, then $\mathbf{y} \in B(\mathbf{y_0}, R) \bigcap H_2^-$, where $B(\mathbf{y_0}, R)$ is the Euclidean closed ball of radius R centered at $\mathbf{y_0}$. Define A to be the open normal cone to $B(\mathbf{y_0}, R) \bigcap H_2^-$, namely,

$$A' = \left\{ \mathbf{u} \in \mathbb{R}^n : \frac{\mathbf{u} \cdot \mathbf{u_2}}{\|\mathbf{u}\|\|\mathbf{u_2}\|} > \frac{\sqrt{R^2 - \delta_2^2}}{R} \right\}.$$

This construction is illustrated in Figure 7. Note that $\mathbf{u_2} \in A'$. Moreover, $\forall \mathbf{u} \in A'$, $\mathbf{x}(\mathbf{u}) \cdot \mathbf{u} \geq \mathbf{y_0} \cdot \mathbf{u} > \mathbf{y} \cdot \mathbf{u} \quad \forall \mathbf{y} \in X \bigcap H_2^-$. Thus $\mathbf{x}(\mathbf{u}) \notin H_2^-$. Let $A = A' \bigcap C$, then A is open since it's the intersection of two open sets, and A is non-empty since $\mathbf{u_2} \in A$ by construction. By the assumption on F, $F(A) > 0$, but $\bar{\mathbf{x}}(A) \notin H_2^-$ (since it's a convex combination of points not in this half-space), so $\bar{\mathbf{x}}(A) \notin \bar{X}$ since $\bar{X} \subseteq H_2^-$. This contradicts the definition of $\bar{X}$. $\square$

Proof of Lemma 3 The goal is to show that we can find a small neighborhood $A \subseteq D$ for which we can perturb the average allocation in direction $\mathbf{d}$ and yield a strict improvement for each $\mathbf{u} \in A$.

Let $\tilde{\mathbf{a_0}} = \texttt{Proj}_D \mathbf{a_0}$, the projection of $\mathbf{a_0}$ onto D. We wish to construct our desired open neighborhood by taking a neighborhood A of $\tilde{\mathbf{a_0}}$ in D such that for every $\mathbf{u} \in U(A)$ (recall that $U(A)$ is the subset of U whose projection on D is in A), the agent prefers the allocation $\mathbf{y_1} = \bar{\mathbf{x}}(U(A)) + \epsilon\mathbf{d}$ rather than $\mathbf{x}(\mathbf{u})$. Moreover, the neighborhood has to be sufficiently small so that $\mathbf{y_1}$ remains feasible, that is, $\mathbf{y_1} \in \Delta$. To do this, we make use of Lemma 3 and construct A from an open cone $C \subseteq D$ that by construction will guarantee the above properties.

Let $\mathbf{y_0} \in X \cap H(\mathbf{a_0}, 1) \cap \texttt{int}(\Delta)$. ($\Delta$ is the feasibility simplex.) Let γ be the distance from $\mathbf{y_0}$ to the boundary of Δ, then $\gamma > 0$ since $\mathbf{y_0}$ is in the interior of Δ. Define $\epsilon = \frac{3}{4}\gamma$. Define $r = \frac{\tilde{\mathbf{a_0}} \cdot \mathbf{d}}{6\|\tilde{\mathbf{a_0}}\|}\epsilon$. Define the cones

$$C_1 = \{\mathbf{a} \in D : \frac{\mathbf{a}}{\|\mathbf{a}\|} \cdot (\mathbf{y_0} - \mathbf{x}(\mathbf{a})) > -r\}$$
$$C_2 = \{\mathbf{a} \in D : \frac{\mathbf{a}}{\|\mathbf{a}\|} \cdot \mathbf{d} > \frac{\tilde{\mathbf{a_0}}}{2\|\tilde{\mathbf{a_0}}\|} \cdot \mathbf{d} = 3\frac{r}{\epsilon}\}$$

Note that $\tilde{\mathbf{a_0}} \in C_1$ and $\tilde{\mathbf{a_0}} \in C_2$. C_1 and C_2 are cones because the expressions that define them depend only on $\frac{\mathbf{a}}{\|\mathbf{a}\|}$. (Note also that $\mathbf{a} \cdot \mathbf{x}(\mathbf{a}) = \mathbf{a} \cdot \mathbf{x}(\lambda\mathbf{a})$ for any $\lambda > 0$ by incentive compatibility.) They are open because

the LHS of the inequalities that define them are continuous functions of $\mathbf{a}$. (Note that $g(\mathbf{a}) = \mathbf{a} \cdot \mathbf{x}(\mathbf{a})$ is a continuous function of $\mathbf{a}$ as this is the objective of the linear maximizer over convex set X.)

Therefore, the set

$$C = C_1 \bigcap C_2$$

is a non-empty open cone. Moreover, by continuity and full relative support of F, $F(U(\cdot))$ is an atomless measure on C such that $F(U(C)) > 0$ and for every open cone $A \subseteq C$, $F(U(A)) > 0$. Finally X and $\mathbf{x}(\cdot)$ satisfy the assumptions of Lemma 4, so by the lemma, $\exists A \subseteq C$, $F(U(A)) > 0$, such that $\|\bar{\mathbf{x}}(U(A)) - \mathbf{y_0}\| \leq r$. Now, define $\mathbf{y_1} = \bar{\mathbf{x}}(U(A)) + \epsilon \mathbf{d}$, then $\mathbf{y_1} \in \Delta$ since $\|\mathbf{y_1} - \mathbf{y_0}\| \leq \epsilon + \delta \leq \frac{7}{8}\gamma$. Consider the alternative allocation rule

$$\mathbf{y}(\mathbf{u}) = \begin{cases} \mathbf{y_1} & \text{if } \mathbf{u} \in U(A) \\ \mathbf{x}(\mathbf{u}) & \text{otherwise.} \end{cases}$$

Then $\mathbf{y}$ strictly Pareto improves over $\mathbf{x}$ because $\forall \mathbf{a} \in A$,

$$\begin{aligned} & \mathbf{a} \cdot \mathbf{y_1} - \mathbf{a} \cdot \mathbf{x}(\mathbf{a}) \\ =& \mathbf{a} \cdot (\bar{\mathbf{x}}(U(A)) + \epsilon \mathbf{d}) - \mathbf{a} \cdot \mathbf{x}(\mathbf{a}) \\ =& \mathbf{a} \cdot (\bar{\mathbf{x}}(U(A)) - \mathbf{y_0}) + \mathbf{a} \cdot (\mathbf{y_0} - \mathbf{x}(\mathbf{a})) + \epsilon \mathbf{a} \cdot \mathbf{d} \\ \geq& -r\|\mathbf{a}\| - r\|\mathbf{a}\| + 3r\|\mathbf{a}\| \\ >& 0 \end{aligned}$$

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

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

Then $\mathbf{x}'$ still strictly Pareto improves over $\mathbf{x}$ but $\bar{\mathbf{x}}'(U) = \bar{\mathbf{x}}(U) + \delta \mathbf{d}$, which is what we needed. □

C.2. Characterization for Ordinal Mechanisms

Proof of Theorem 2. The proof is similar to that of Theorem 1 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 1, we have the base polytope of a polymatroid. The exchange argument is also simpler because the space of permutations Π is discrete and every member has positive probability due to full relative support.

As before, we first apply a series of lemmas and prove them later.

LEMMA 5. *An ordinal allocation rule $\mathbf{x}(\pi)$ is incentive compatible if and only if there exists monotone submodular set function $f : 2^{|S|} \to [0, 1]$ s.t. for every permutation $\pi \in \Pi$ and for every k ($1 \leq k \leq |S|$),*

$$x_{\pi(k)}(\pi) = f(\{\pi(1), \cdots, \pi(k)\}) - f(\{\pi(1), \cdots, \pi(k-1)\}).$$

We call f the monotone submodular set function that corresponds to $\mathbf{x}$.

If X is the range of $\mathbf{x}$, then the above lemma says that $\mathbf{x}$ is incentive compatible if and only if X is the vertex set of the base polytope of polymatroid defined by f:

$$\begin{aligned}\sum_{s\in M} x_s &\le f(M) \quad \forall M \subseteq S\\ \sum_{s\in S} x_s &= 1\\ x &\ge 0\end{aligned}$$

The following lemma embodies the exchange argument.

LEMMA 6. *Let f be the monotone submodular set function that corresponds to incentive compatible allocation rule $\mathbf{x}$. If $\mathbf{x}$ is ordinal efficient, then for any $M_1, M_2 \subseteq S$,*

$$f(M_1 \cup M_2) = \max\{f(M_1), f(M_2)\}.$$

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

Proof of Lemma 5. If $\mathbf{x}(\pi)$ is an incentive compatible ordinal allocation rule, then for any $M \subseteq S$, define

$$f(M) = \sum_{j=1}^{|M|} \mathbf{x}_{\pi(j)}(\pi), \quad \text{where } \{\pi(1), \pi(2), \cdots, \pi(|M|)\} = M.$$

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 $S\backslash M$), to be fixed, regardless of the relative rank between services in M and betewen services in $S\backslash M$. If this were not the case, then for some large $b > 0$ and small $\epsilon > 0$, consider an agent with utilities $u_s = \mathbb{1}(s \in M)b + \epsilon_s$, where $\mathbb{1}(s \in M)$ equals one if $s \in M$ and zero otherwise, and ϵ_s's are distinct numbers to be defined later, with $|\epsilon_s| \le \epsilon$. If the agent's chance of getting one of the service in M can be altered by changing relative order in M and the relative order in $S\backslash M$, while she ranks M before $S\backslash M$, 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, let $u_s = \mathbb{1}(s \in M_1 \cup \{s\})b$ with some $b > 0$ to be specified later. Her true ranking is $M_1 \cup \{s\}$ before $M_2\backslash M_1$. But reporting this true ranking gives her expected utility of $bf(M_1 \cup \{s\})$. However, if she instead ranked M_1, then $M_2\backslash M_1$, then s, she would get $b(f(M_1) + f(M_2 \cup \{s\}) - f(M_2)) > bf(M_1 \cup \{s\})$. So she has incentive to misreport. This contradicts incentive compatibility.

Now, the construction of f implies f is monotone, and by the definition of f, we have that $\forall \pi \in \Pi$ and $1 \le k \le |S|$, $x_{\pi(k)}(\pi) = f(\{\pi(1), \cdots, \pi(k)\}) - f(\{\pi(1), \cdots, \pi(k-1)\})$.

Conversely, if f is a monotone submodular set function. We show that if we define $\mathbf{x}$ so that $x_{\pi(k)}(\pi) = f(\{\pi(1),\cdots,\pi(k)\}) - f(\{\pi(1),\cdots,\pi(k-1)\})$, 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:

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

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

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

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. □

Proof of Lemma 6. By monotonicity of f, $f(M_1 \cup M_2) \ge \max\{f(M_1), f(M_2)\}$. What we need to show is that $f(M_1 \cup M_2) \le \max\{f(M_1), f(M_2)\}$. 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, π_1 and π_2: π_1 ranks services in M_1 first, followed by M_2, followed by other services in arbitrary order; π_2 ranks services in M_2 first, followed by M_1, followed by others. By Lemma 5, since $\mathbf{x}$ is incentive compatible, $\sum_{s\in M_2} \mathbf{x}(\pi_1) = f(M_1 \cup M_2) - f(M_1) > 0$, and $\sum_{s\in M_1} \mathbf{x}(\pi_2) = f(M_1 \cup M_2) - f(M_2) > 0$. Thus, agents with preference ranking π_1 can trade probabilities with agents with preference ranking π_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 ordinal support, there exist positive measures of both kinds of agents, so $\mathbf{x}$ is not ordinal efficient, contradiction. □

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 mostly in its ability to distinguish between agents that have an extremely large relative preferences for a services over another and agents that have only a weak preference.

Let M and N be two positive real numbers with $M \ge \min(3, N^3)$ and $N \ge 1$. Suppose that there are three services. Service 1 has capacity $\frac{1}{N^2}$, while services 2 and 3 have capacity 1 each. Suppose there is only one type of mass 1, and $\frac{1}{N}$ of the agents have utilities $(M, 1, 0)$ and the remaining agents have utilities $(2, 1, 0)$.[31] The objective is to maximize the social welfare. An optimal cardinal mechanism charges price vector $(p, 1, 0)$, where $p > 2$ to differentiate the two types of agents. Agents have virtual budget 1. The $\frac{1}{N}$ of agents would

[31] Although this does not satisfy full relative support, we can trivially modify it to satisfy by having ϵ mass of agents with utilities (u_1, u_2, u_3), where the u_j's are distributed i.i.d. standard Normal.

purchase the bundle $(\frac{1}{N}, 0, 1-\frac{1}{N})$, while the other agents will opt for $(0,1,0)$. Hence, the social welfare is $\frac{M}{N^2}+(1-\frac{1}{N}) > \frac{M}{N^2}$. However, with an ordinal mechanism, one cannot distinguish between the two groups of agents, and the best a lottery-plus-cutoff mechanism can do is to have cutoffs $(\frac{1}{N^2}, 1, 1)$ for the services, and every gets allocation $(\frac{1}{N^2}, 1-\frac{1}{N^2}, 0)$. The social welfare is $\frac{1}{N}(\frac{M}{N^2}+1-\frac{1}{N^2})+(1-\frac{1}{N})(\frac{2}{N^2}+1-\frac{1}{N^2}) \leq \frac{3M}{N^3}$. So the ratio between the best cardinal and the best ordinal in this example is at least $\frac{N}{3}$, which we can make arbitrarily large.

C.4. Computation Results

Proof of Theorem 3 Define dual variables for (LargeMarketLP) as follows: let γ be the dual variable for the cost constraint, λ_s for the capacity constraint of school s, μ_t for the constraint of menu probabilities summing to one for type t, and ν_t be the constraint enforcing the minimum constraint for type t. The dual is as follows.

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

Label the right hand side of the first inequality as $f_t(\gamma, \boldsymbol{\lambda}, \boldsymbol{\nu}, M)$. 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 $\mathbb{1}(t\in T_s)\lambda_s + \gamma B_{ts}$ credits, then $f_t(\gamma, \boldsymbol{\lambda}, \boldsymbol{\nu}, M)$ 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 an "optimal-menu" problem.

DEFINITION 5. Given γ, $\boldsymbol{\lambda}$, $\boldsymbol{\nu}$, the *optimal menu sub-problem* is to find the solution set

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

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

LEMMA 7. *$\mu_t(\gamma, \boldsymbol{\lambda}, \boldsymbol{\nu})$ is convex.*

Proof of Lemma 7. This follows from $f_t(\gamma, \boldsymbol{\lambda}, \boldsymbol{\nu}, M)$ being linear in γ, $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ for fixed M. So μ_t is the upper envelope of a family of linear functions, and is therefore convex. □

Therefore, the dual can be written as a convex program with $2|T|+|S|+1$ non-negative variables, with objective $C\gamma + \mathbf{m}\cdot\boldsymbol{\lambda} + \sum_{t\in T}\mu_t(\gamma, \boldsymbol{\lambda}, \boldsymbol{\nu})$ and a single linear constraint $\sum_{t\in T}\nu_t \geq 1-\alpha$. One difficulty is that the optimal menu sub-problem needs to optimize over all possible exponentially many menus $M\subseteq S$. However, when preferences are multinomial-logit, we can efficiently solve the sub-problem.

LEMMA 8. *Under multinomial-logit utility priors, if $\alpha w_t + \nu_t > 0$, then the number of optimal solutions for the optimal menu sub-problem is at most $|S|$, and can all be found in time $|S|\log|S|$.*

Proof of Lemma 8. Recall that multinomial-logit utilities means that $u_{is} = \bar{u}_t s + b_t \epsilon_{is}$ where the ϵ_{is}'s are standard Gumbel distributed idiosyncratic shocks. Fix a type t. Let $h_s = \frac{n_t(\mathbb{1}(t \in T-s)\lambda_s + \gamma B_{ts})}{(\alpha w_t + \nu_t) b_t}$, $z_s = exp(\frac{\bar{u}_{ts}}{b_t})$. The optimal menu sub-problem is equivalent to finding all solutions to

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

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:

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

Now if $h_s < h_{s'}$ and $y_{s'} > 0$ but $y_s < z_s$, then by decreasing $y_{s'}$ by δ and by increasing y_s by δ, for small $\delta > 0$, 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 $1 \leq k \leq |S|$.

$$y_s = z_s \quad \forall s < k, \quad y_k \in [0, z_k], \quad y_s = 0 \quad \forall s > k.$$

We show that at an optimal solution, it must be that $y_k \in \{0, z_k\}$. Suppose on the contrary that $y_k \in (0, z_k)$. Now, let $d_1 = \sum_{s<k} z_s$, $d_2 = h_k d_1 - \sum_{s<k} h_s z_s$. As a function of y_k, the objective and its first and second derivatives are

$$\begin{aligned} g(y_k) &= \log(d_1 + y_k) + \frac{d_2}{d_1 + y_k} - h_k, \\ g'(y_k) &= \frac{1}{d_1 + y_k} \left(1 - \frac{d_2}{d_1 + y_k}\right), \\ g''(y_k) &= \frac{1}{(d_1 + y_k)^2} \left(\frac{2d_2}{d_1 + y_k} - 1\right). \end{aligned}$$

Since y_k is an interior optimum, $\frac{d_2}{d_1 + y_k} = 1$, and so the second derivative $g''(y_k) = \frac{1}{(d_1 + y_k)^2} > 0$, which implies that y_k is a strict local minimum, which contradicts our assumption. Therefore, the objective is maximized when $y_k \in \{0, z_k\}$.

This implies that all optimal solutions are restricted to be of the form $M_k = \{1, \cdots, k\}$ (the services are sorted in increasing order of h_s), and so we only need to search through $1 \leq k \leq |S|$. 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. This also implies that the number of optimal solutions is at most $|S|$.

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_k's for $k \in \{1, \cdots, |S|\}$ still yields all optimal solutions. $\square$

The proof of Lemma 8 reveals insight on what the optimal solution looks like with a multinomial-logit utility model: based on the shadow cost vector γ for the budgets and shadow cost λ_s for capacity of service s, the algorithm places a virtual "allocation cost" $\mathbb{1}(t \in T_s)\lambda_s + \gamma B_{ts}$ 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" k number of services to include, and this is chosen by balancing expected allocation costs with expected utility, with the weight on expected utility $\frac{\alpha w_t + \nu_t}{n_t}$ 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 $\mathcal{M}_t$ 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:

$$\begin{aligned}
\sum_{M \in \mathcal{M}_t} v_t(M) z_t(M) &\geq y \quad \forall t \in T \text{ with equality if } \nu_t > 0 \\
\sum_{M \in \mathcal{M}_t} z_t(M) &= 1 \quad \forall t \in T \\
\sum_{t \in T_s} \sum_{M \in \mathcal{M}_t} n_t p_t(s, M) z_t(M) &\leq m_s \; \forall s \in S \text{ with equality if } \lambda_s > 0 \\
\sum_{s,t} \sum_{M \in \mathcal{M}_t} n_t p_t(s, M) B_{ts} z_t(M) &\leq C \quad \text{with equality if } \gamma > 0 \\
z_t(M) &\geq 0 \quad \forall t \in T, M \in \mathcal{M}_t
\end{aligned}$$

□