Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Optimal Allocation Without Money: An Engineering Approach

Itai Ashlagi, Peng Shi

To cite this article:
Itai Ashlagi, Peng Shi (2015) Optimal Allocation Without Money: An Engineering Approach. Management Science

Published online in Articles in Advance 19 Aug 2015

. http://dx.doi.org/10.1287/mnsc.2015.2162

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–20
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2015.2162

Optimal Allocation Without Money: An Engineering Approach

Itai Ashlagi, Peng Shi
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 {iashlagi@mit.edu, pengshi@mit.edu}

We study the allocation of heterogeneous services to agents with incomplete information and without monetary transfers. Agents have private, multidimensional utilities over services, drawn from commonly known priors. The social planner's goal is to maximize a potentially 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, in order 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 formulation with more than 868 types of students and 77 schools, and we translate the solution into a finite-market mechanism that significantly outperforms the baseline plan chosen by the city in terms of efficiency, equity, and predictability.

Data, as supplemental material, are available at http://dx.doi.org/10.1287/mnsc.2015.2162.

Keywords: market design; priors; assignment; school choice; optimal design; large-market approximation
History: Received May 22, 2014; accepted November 30, 2014, by Martin Lariviere, operations management.
Published online in *Articles in Advance*.

1. Introduction

In many settings, goods or services are allocated to agents without 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 have multifaceted concerns, such as social welfare, equity, and system costs.

One example of such a problem was faced by the city of Boston in the 2012–2013 school assignment reform. Prior to 2013, seats in Boston Public Schools (BPS) had been allocated as follows. The city was divided into three zones, and each family ranked schools within a menu of choices that consisted of schools within the family's zone and schools within one mile of the family's home. A centralized algorithm then allocated seats based on families' ranking lists, priorities, and lotteries. These large menus of schools led to unsustainably high busing costs, representing approximately 10% of the total school board budget (Russell and Ebbert 2011). In 2012, the city used simulation to choose a new plan from a short list of proposals, to decrease the choice menus while maintaining equity among neighborhoods. Although the proposals in the short list were ad hoc, the techniques in this paper can be used to design the optimal allocation in a systematic way.

When agents' preferences are publicly known, the allocation is merely an optimization problem. This study addresses how to allocate services when agents' preferences are only privately known. In our model, there are multiple kinds of services, and agents have private, multidimensional utilities over the services, drawn from commonly known priors, which may depend on agents' observable information. Since mechanism design with multidimensional valuations has traditionally been difficult, especially for general objectives, we adopt an engineering approach: we first find the optimal mechanism for a large-market formulation, then we convert it into a feasible finite-market mechanism, and we evaluate it using data-driven simulation.

In the large-market formulation, 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 of a certain socioeconomic status. We require mechanisms to be incentive compatible and Pareto optimal among agents of the same type, and we refer to such mechanisms as *valid* mechanisms. Whereas 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 on the utility priors, we show that

any valid mechanism can be described as a collection of competitive equilibria with equal income (CEEI). More precisely, agents of each type are given "virtual prices" for probabilities of each service, and the allocation can be interpreted as giving agents one unit of "virtual money" and allowing them to "purchase" their preferred probabilistic bundle of services (a related mechanism was introduced by Hylland and Zeckhauser 1979). Virtual prices may vary across types, but agents within the same type are given the same prices.

In many contexts, only relative preferences are elicited, but not preference intensities. (For example, in Boston, families submit preference rankings instead of utilities.) Such mechanisms are called *ordinal*, as opposed to *cardinal*, mechanisms, which have no information requirements. Under mild regularity assumptions, we show that any valid ordinal mechanism can be described as "lottery-plus-cutoff": each agent receives a lottery number drawn from the uniform distribution between 0 and 1. 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. Finally, each agent is allocated her most preferred service to which she is admitted.

These structural results provide insights into the types of mechanisms observed in practice. In many business schools, course allocation is made by a bidding process, in which students are given a number of points (virtual currency) and the highest bidders are assigned seats.[1] This is conceptually similar to CEEI. In many public school districts,[2] students submit preference rankings for schools, and a centralized mechanism uses submitted preferences, predefined priorities, and random lottery numbers to determine the assignment.[3] Given ex post lottery number cutoffs at each school for each priority class of students, this is analogous to the lottery-plus-cutoff mechanism.

These theorems also simplify the computation of the optimal mechanism, since we no longer have to optimize over allocation functions, but over a finite number of prices or cutoffs. To illustrate, we compute the optimal large-market ordinal mechanism in an empirically relevant setting. We encode optimal cutoffs by an exponentially sized linear program, and we show that the dual can be solved efficiently when agents' utilities are based on a multinomial-logit discrete choice model.

To demonstrate the relevance of our large-market formulation, we employ it to optimally design school choice in Boston. We use the same amount of busing as in the mechanism chosen by the city, referred to here as the *baseline* mechanism, while optimizing a linear combination of utilitarian welfare and max-min welfare. All the analyses use real data from BPS.

We first define a finite-market formulation of the problem, which resembles one the city faced during the 2012–2013 reform. We define a *large-market approximation* and compute the optimal (large-market) cutoffs. We use these cutoffs to design the corresponding menus and priorities and use the deferred acceptance (DA) algorithm (Gale and Shapley 1962) to produce a feasible finite-market mechanism that can be seen as "asymptotically optimal."[4] We evaluate this mechanism in the finite-market setting and show by simulation that it significantly improves on the baseline in all aspects. It improves utilitarian welfare by an amount equivalent to decreasing students' average distance to school by 0.5 miles, and it improves max-min welfare by about 2.5 miles. These are significant gains, because the baseline improves over the most naïve plan in utilitarian welfare by only 0.6 miles and in max-min welfare by 1.7 miles, so our solution effectively doubles the gains. Furthermore, our mechanism improves students' chances of getting their first choice by 15%.

Our methodology may also be applied to other allocation problems requiring consideration of agents' preferences. Potential examples include housing lotteries, in which the social planner may want to optimize welfare and fairness, subject to various distributional constraints; business school course bidding, in which an optimal cardinal mechanism may be used in place of current bidding schemes; and the assignment of internships or relocation opportunities. We leave detailed explorations for future work.

1.1. Related Literature

Our work connects four strands of previous research. The first is the matching literature, which traditionally focuses on designing mechanisms that satisfy certain properties, such as Pareto efficiency, an appropriate notion of fairness, and strategy-proofness (see, e.g., Roth and Sotomayor 1990, Abdulkadiroğlu and Sönmez 2013). Hylland and Zeckhauser (1979) study cardinal mechanisms that achieve Pareto efficiency and propose the CEEI mechanism, which also arises in our characterization of valid cardinal mechanisms. Bogomolnaia and Moulin (2001) study ordinal mechanisms that satisfy ordinal efficiency, an ordinal notion of Pareto optimality that we adopt (except we consider ordinal efficiency within each type of agent).

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

[2] Examples include Boston, New York City, New Orleans, and San Francisco.

[3] For more information, see Abdulkadiroğlu et al. (2006, 2009).

[4] This solution is asymptotically optimal in the sense that if the market scales up with independent copies of itself, the finite-market formulation converges to the large-market formulation, and the finite-market solution also converges to the large-market optimum.

Abdulkadiroğlu and Sönmez (2003), Abdulkadiroğlu et al. (2009) and (2006) apply matching theory to school choice, and their works have been influential in the adoption of strategy-proof ordinal mechanisms over nonstrategy-proof alternatives in cities such as New York, Boston, Chicago, New Orleans, and San Francisco. More recent works on school choice study nuances in tie breaking (Pathak and Sethuraman 2011, Erdil and Ergin 2008, Ashlagi and Shi 2014), incorporating partial preference intensities (Abdulkadiroğlu et al. 2015), and "controlled school choice," which addresses implementing constraints such as diversity (Kominers and Sönmez 2015, Echenique and Yenmez 2015, Ehlers et al. 2014). Unlike our work, this literature (and the matching literature in general) rarely assumes priors on agent utilities (especially asymmetric priors) and rarely optimizes a global objective. Our work can be viewed as bridging the matching and the mechanism design/auction literature.[5]

Another strand is optimal mechanism design without money in the finite-market framework.[6] Currently, analytical difficulties restrict positive results to single-dimensional valuations. Miralles (2012) solves for the optimal cardinal mechanism for two services and many bidders with a symmetric utility prior. He shows that under certain regularity conditions, the optimal allocation can be described as CEEI. He further shows that this can be implemented using a combination of lottery, insurance, and virtual auction, in which agents use probabilities of the less desirable good as virtual currency to bid for the more desirable good. However, the analysis requires reducing the valuation space to a single dimension by taking the ratio of the utilities for the two services, so it does not generalize to more than two services. Our work shows stronger results by leveraging a large-market approximation. Hartline and Roughgarden (2008), Hoppe et al. (2009), Condorelli (2012), and Chakravarty and Kaplan (2013) study models in which agents cannot pay money but may "burn money" or exert effort to signal their valuation. One insight from their work is that if a utility prior does not have a thick tail—or more precisely, if it satisfies a monotone hazard rate condition—then requiring agents to exert effort is unnecessary, and a lottery will maximize the social welfare. However, their results cannot be easily extended to a multidimensional setting, which is more realistic when multiple types of services are offered.

A third strand is optimal school districting, which studies giving each family a designated school by home location. Representative works include Clarke and Surkis (1968), Sutcliffe et al. (1984), and Caro et al. (2004). These papers use linear or mixed-integer optimization to optimize a global objective, which may take into account distances between students and schools, racial balance, capacity utilization, geographic contiguity, and uncertainties in future student population. We complement this literature by combining optimization with the elicitation of families' private preferences, thus bringing together the best of central planning and the market.

In terms of techniques, our paper builds on previous large-market models with a continuum of agents. Continuum agent models often simplify the analysis over a finite model, leading to stronger and cleaner results. Such models are common in the industrial organization literature (Tirole 1988). Azevedo and Leshno (2015) study matching markets where each side of the market has preferences over the other (in our case, services do not have preferences over individual agents). They focus on characterizing stable matchings and do not consider global optimization. In the operations management literature, our engineering approach of solving for a large-market optimum and using it to construct a feasible solution has appeared before, but most previous applications have been related to queuing.[7] Our paper illustrates the power of this approach when applied to the allocation problem.

2. The Large-Market Environment

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 the convention in mechanism design, our notion of "type" does not denote agents' private information, but rather their public information.[8] There is a finite set S of services, and every agent must be allocated exactly one service (outside options can be represented by an additional "null service"). Allocations may be probabilistic, so the set of possible allocations for each agent is the probability simplex,[9]

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

Agents of type t have private utilities for services, which are distributed according to a commonly known continuous measure F_t over utility space $U = \mathbb{R}^{|S|}$. Each $\mathbf{u} \in U$ is a utility vector in which

[5] 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 global objectives in the matching literature.

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

[7] See, e.g., Maglaras and Zeevi (2005) and Perry and Whitt (2009).

[8] For example, in school choice, the type may correspond to the student's neighborhood; in course allocation, the type may correspond to the student's major.

[9] It suffices to study assignment probabilities because, by the Birkhoff-von Neumann theorem, assignment probabilities can be decomposed as a random lottery over deterministic assignments.

each component denotes utility for a service. For any measurable subset $A \subseteq U$, $F_t(A)$ denotes the mass of type t agents who have utilities in A. Since the total mass for type t is n_t, the total measure $F_t(U)$ equals n_t.

The social planner must design a mechanism to elicit agents' preferences and to decide on the allocation. The goal is to maximize a certain objective. Mechanisms can either be cardinal or ordinal: cardinal mechanisms allow unrestricted preference elicitation, whereas ordinal mechanisms allow only elicitation of relative rankings but not preference intensities. We study these separately in §§3 and 4.

3. Cardinal Mechanisms

Stated formally, a cardinal *mechanism* $\mathbf{x}$ is a collection of *allocation rules* $\mathbf{x}_t$, one for each type t, where each $\mathbf{x}_t$ is a mapping from reported utilities to a possible allocation, $\mathbf{x}_t\colon U \to \Delta$, and is measurable with respect to F_t.[10]

For agents to submit their true preferences safely, we require the allocation rule to be incentive compatible.[11] An allocation rule is *incentive compatible* if it is in the agent's best interest to report her true utility vector:

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

We further require allocation rules to be *Pareto efficient within type*, which implies that agents of the same type cannot benefit from trading allocation probabilities. Formally stated, $\mathbf{x}_t$ is Pareto efficient within type if no other function $\mathbf{x}'_t\colon U \to \Delta$ exists that has the same average allocation

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

is weakly preferred by all agents

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

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

Pareto efficiency within type can be viewed as a "stability" criterion: our formulation implicitly assumes that the social planner treats agents within a given type symmetrically and cannot discriminate based on the exact identity of the agent. Therefore, it may be unreasonable to enforce no trading of allocations for agents of the same type. This may also be interpreted as foreseeing possible Pareto-improving trades and internalizing them in the mechanism.

Observe that our requirement of Pareto efficiency within type still allows agents of different types to benefit from trading allocations.[12] This allows our mechanism to differentiate freely with respect to types if this improves the objective. For school choice, this corresponds to differentiating students with respect to home location. Unless otherwise noted, all subsequent mentions of Pareto efficiency refer to within type.

We call an allocation rule *valid* if it is both incentive compatible and Pareto efficient within type, and a mechanism is *valid* if all its allocation rules are valid. For now, we allow the objective function $W(\mathbf{x})$ to arbitrarily depend on all the allocation rules $\mathbf{x}_t$'s. This generality allows the objective to incorporate many considerations, such as agents' welfare, capacity constraints, and differential costs. To illustrate how such considerations can be modeled, observe that the expected utility of an agent of type t is

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

Thus one 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 $q_s = \sum_t \int_U x_{ts}(\mathbf{u})\, dF_t$. We can model a hard capacity limit m_s for service s by setting $W(\mathbf{x})$ to be negative infinity when $q_s > m_s$. Alternatively, one can model a smooth penalty for exceeding capacity by subtracting a penalty term $C(\max\{0, q_s - m_s\})$ from the objective where $C(\cdot)$ is a convex cost function.

3.1. 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 CEEI of an artificial currency. This means that for each service s, there exists a "price" $a_s \in (0, \infty]$ (possibly infinite) for buying units of probability using units of the artificial currency, and the allocation is what agents would buy with one unit of artificial currency when offered probabilities of various services at these prices:

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

Figure 1 illustrates a CEEI with three services. The price vector $\mathbf{a}$ is the same for all agents of the same type, but it may differ across types.

This result implies that the search for the optimal mechanism can be restricted to a search over the set of price vectors for each type. 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 $\mathbf{x}_t\colon U \to \Delta$.

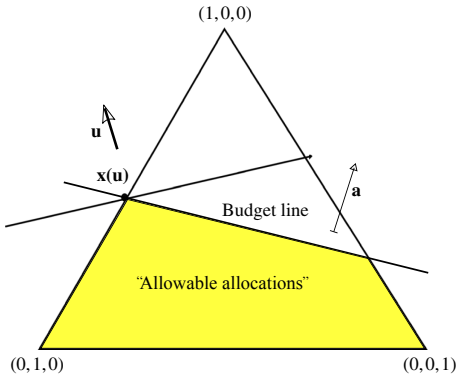
Since every agent must be assigned, we gain no useful information if an agent's utilities were all

[10] If we did not work within a large-market environment, the allocation rule would also be a function of others' reports.

[11] If one assumes agents play in a Bayesian Nash equilibrium, assuming incentive compatibility is without loss of generality.

[12] If full Pareto efficiency were desired, one could model the situation with one type or use weighted social welfare as the objective.

Figure 1 (Color online) Illustration of CEEI with $|S| = 3$



Notes. 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 $\{\mathbf{x}_t(\mathbf{u})\}$. An agent of this type could obtain allocations in the interior of this region by randomizing over several reports. For utility vector $\mathbf{u}$, the agent receives an allocation $\mathbf{p}$ that maximizes expected utility $\mathbf{u} \cdot \mathbf{p}$ subject to the price of $\mathbf{p}$, $\mathbf{a} \cdot \mathbf{p}$, not exceeding the budget of one.

shifted by the same additive constant. To eliminate this degree of freedom, let $D = \{\mathbf{u} \in \mathbb{R}^n: \mathbf{u} \cdot \mathbf{1} = 0\}$. This is the normal subspace for the all-ones 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 $A \subseteq D$, we define $U(A)$ to be the subset of U whose projections onto D are in A.

The regularity condition we impose on F_t is that, a priori, an agent's relative preference could, with positive probability, take any direction in D. This can be interpreted as the central planner not being able to rule out any relative preferences a priori.

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[13] $C \subseteq D$ 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 nonempty open cone $C \subseteq D$, $F_t(U(C)) > 0$.

Theorem 1. *Consider a given type and suppose that 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 CEEI with some price vector* $\mathbf{a} \in (0, \infty]^{|S|}$.

Note that because of our large-market environment, this theorem implicitly assumes that the allocation rule treats agents symmetrically and is nonatomic, meaning that no individual agent can affect the market, because in evaluating Pareto efficiency we consider only positive masses of agents.

The full proof of this theorem contains fairly technical steps and is deferred to the appendix. But we provide the main idea here: incentive compatibility in our setting implies that the set $X_t = \{\mathbf{x}_t(\mathbf{u})\}$ lies on the boundary of its convex hull and 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 allocation rule $\mathbf{x}_t$ 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. Denote by X_t 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 a unique supporting hyperplane that intersects X_t in the interior of the unit simplex, we are done, since that hyperplane can be represented by a price vector. If there are at least two such hyperplanes, then we show that there is a "trading direction" $\mathbf{d}$ by which some positive measure of agents may prefer allocations in the direction $\mathbf{d}$ and others would prefer allocations in the direction $-\mathbf{d}$, thus producing a Pareto-improving trade, 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.

Similar results have previously appeared in the literature but with different assumptions. They do not imply our result. Aumann (1964) shows conditions in which any Pareto-efficient allocation is supported by equilibrium prices, although not necessarily from equal incomes. His analysis crucially depends on the unboundedness of the space of allocations, which in our cases is the bounded unit simplex. Zhou (1992) and Thomson and Zhou (1993) show that under certain regularity conditions, any "strictly envy-free" and Pareto-efficient allocation is supported as CEEI. However, their strictly envy-free condition is stronger than our incentive compatibility condition. Their analysis also requires the allocation space to not contain its boundary, which is not true in our case since our unit simplex is closed.

Our result has similarities to the second welfare theorem, which states that any Pareto-efficient allocation can be sustained by a competitive equilibrium. An important difference is that in the second welfare theorem,[14] the budgets of agents may need to be different

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

[14] The second welfare theorem is traditionally stated with locally nonsatiated utilities, which is not satisfied when agents can be allocated only one good and utilities are not transferable. See Miralles Asensio and Pycia (2014) for a version of the welfare theorem in this setting.

and possibly depend on their preferences. However, in our case, we can give all agents a common budget, independent of their utilities.

4. Ordinal Mechanisms

Ordinal mechanisms cannot elicit preference intensities but only relative rankings. Let Π be the set of permutations of S, corresponding to preferences rankings over services. An *ordinal mechanism* is a collection of *ordinal allocation rules* $\mathbf{x}_t$, one for each type t, each of which is a mapping between preference rankings and allocations, $\mathbf{x}_t\colon \Pi \to \Delta$.

Denote by $U(\pi) \subseteq U$ the set of utility vectors that are consistent with the ranking π in the sense that utilities for services 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 allocation rule $\mathbf{x}_t$ is *incentive compatible* if truth-telling maximizes utility. That is, for every $\mathbf{u} \in U(\pi)$,

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

An allocation rule $\mathbf{x}_t$ is *ordinal efficient within type* if no coalition of agents of this type can trade probabilities and all improve in the sense of first-order stochastic dominance. Precisely speaking, $\mathbf{x}_t$ is ordinal efficient if no another function $\mathbf{x}'_t\colon \Pi \to \Delta$ exists 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 for every $\pi \in \Pi$ and for every $1 \le k \le |S|$,

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

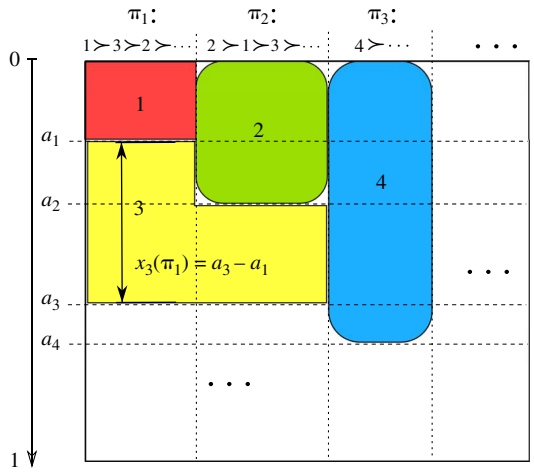
and the inequality is strict for some k and some π of positive measure, $F_t(\pi) > 0$.

The social planner's goal is to optimize an arbitrary function $W(\mathbf{x})$ subject to each allocation rule $\mathbf{x}_t$ being valid. Note again that we require only ordinal efficiency within type, so the mechanism has the flexibility to tolerate potential Pareto-improving trades between different types if this improves the objective. All subsequent mentions of ordinal efficiency refer to within type.

4.1. Characterization of Valid Ordinal Allocation Rules

We show that any valid ordinal allocation rule can be represented as a lottery-plus-cutoff mechanism: agents are given lottery numbers distributed as

Figure 2 (Color online) Illustration of Lottery-Plus-Cutoff



Notes. 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.

`Uniform`$[0, 1]$, and each service has a lottery cutoff for each type. An agent is "admitted" to a service if and only if her lottery number does not exceed the cutoff and the agent chooses her most preferred service among those that she is admitted to. This is illustrated in Figure 2. In mathematical terms, this can be stated as follows.

Definition 2. An ordinal allocation rule $\mathbf{x}\colon \Pi \to \Delta$ is *lottery-plus-cutoff* if there exist "cutoffs" $a_s \in [0, 1]$ such that

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

For additional insights, we provide here an equivalent formulation of lottery-plus-cutoff, which we call *randomized-menus-with-nested-menus*: define probabilities $p_1, \ldots, p_{|S|}$ and menus of services, $M_1, \ldots, M_{|S|}$, such that $S = M_1 \supseteq M_2 \supseteq \cdots \supseteq M_{|S|}$. An ordinal allocation rule is randomized-menus-with-nested-menus if each agent is offered menu M_k with probability p_k. To see that this is equivalent to lottery-plus-cutoff, relabel the services in increasing order of cutoffs: $a_1 \le a_2 \le \cdots \le a_{|S|}$. Define $a_0 = 0$. The one-to-one mapping is $M_k = \{k, \ldots, |S|\}$, and $p_k = a_k - a_{k-1}$.

Analogous to the full relative support assumption in the cardinal setting, we require a regularity condition on the utility distribution, which says that every ordinal ranking $\pi \in \Pi$ is possible.

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

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

As with the cardinal case, because of how we set up the large-market environment, this theorem implicitly assumes that the allocation rule is symmetric and nonatomic.

The proof is given in Appendix C.[15] The intuition behind it is similar to Theorem 1. In the cardinal case, incentive-compatible allocation rules are associated with arbitrary convex subsets of Δ. In the ordinal case, instead of a convex set, we have an associated 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 a lottery-plus-cutoff, some agents can trade and obtain allocations that first-order stochastically dominate their current allocations.

This result can be seen as a generalization of previous characterizations of ordinal efficiency in large markets. Bogomolnaia and Moulin (2001) prove that in a finite market, every ordinal efficient mechanism is a variant of their "probabilistic serial" mechanism. (Their formulation only has one type of agent.) Che and Kojima (2011) show that random serial dictatorship, the simple mechanism of ordering agents randomly and having them pick their best remaining choices iteratively, is asymptotically equivalent to the probabilistic serial in a large market. Liu and Pycia (2013) further show that in a large market, all asymptotically efficient, symmetric, and asymptotically strategy-proof ordinal mechanisms coincide with the probabilistic serial in the limit. One can view our lottery-plus-cutoff mechanism (or randomized-menus-with-nested-menus) as a generalization of the probabilistic serial but allowing arbitrary prioritization based on agents' types, so our result can be viewed as a generalization to allow heterogeneous agent types.

4.2. Comparing Cardinal and Ordinal Mechanisms

Intuitively, in a cardinal setting, agents of the same type can trade probabilities for various services at different ratios, hence expressing their preferences for not only *which services* they value but also *how much* they value each. In an ordinal setting, agents of the same type can also trade probabilities, but they must trade services one-for-one; hence they can express only preference rankings. The value of a cardinal mechanism over an ordinal mechanism lies in its ability to differentiate agents with extreme preference intensities. Thus, if agents' preferences for various services are of similar intensities, then we would expect ordinal mechanisms to perform well compared with cardinal ones. If preferences, however, exhibit extreme differences in intensities, then one would expect cardinal mechanisms to outperform ordinal ones.

In Appendix C.3 we give an example showing that extreme differences in preference intensities may lead to an arbitrarily large ratio between the optimal social welfare achievable from a valid cardinal mechanism and that achievable from a valid ordinal mechanism.[16]

5. Empirical Application: Public School Assignment

Now we demonstrate how our techniques can be applied to a real-world problem and yield empirically relevant results. The problem is based on the 2012–2013 Boston school assignment reform, which considered a list of potential plans and used simulation to evaluate them on a portfolio of metrics. In this section, we ask the reverse question: Given the objective and constraints, what is the *optimal* plan? More precisely, we seek a plan that uses the same transportation budget as the baseline plan chosen by the city but is optimized for efficiency and equity as measured by utilitarian welfare and max-min welfare. We will also improve predictability (chances of getting a top choice) in the process.

Although we use real data, the problem presented here has been simplified for conceptual clarity.[17] We recognize that to produce implementable recommendations, the precise objective and constraints must be scrutinized and debated by all stakeholders and constituents, which has not yet taken place. Hence, the purpose of this section is not to provide concrete policy recommendations for Boston but to provide a proof of concept of our methodology applied to a real-world setting.

We first give a finite-market formulation of the problem. This is an asymmetric, multidimensional mechanism design problem with complex objective and constraints, for which an exact optimum remains

[15] A similar theorem appears in Ashlagi and Shi (2014), but the setting is slightly different here, and we give a different proof.

[16] Independently, Pycia (2014) provides another example.

[17] In the actual problem faced by the city committee, one needs to consider continuing students, specialized programs for English Language Learners and disabled students, and special preference for students who have older siblings already attending a school. In this paper, we ignore these complications. However, all of them could be accommodated in principle.

elusive. As a baseline, we describe the actual plan adopted by BPS, which can be seen as an intuitive heuristic solution to the original problem. To apply the techniques in this paper, we define a large-market approximation of the finite-market formulation and solve for the large-market optimum. We then define a finite-market analog to this large-market solution, which is a feasible solution to the finite-market formulation and is asymptotically optimal in the sense that it becomes the large-market optimum as the finite-market formulation is scaled up. We compare this "asymptotically optimal" solution to the baseline, quantify the improvements, and discuss insights.

5.1. Finite-Market Formulation

5.1.1. Student Population and School Capacities. Approximately 4,000 students each year apply to BPS for kindergarten 2 (K2), the main entry grade for elementary school. The social planner (the city, in this case) is charged with designing an assignment system for K2 that is efficient; equitable; and respectful of certain capacity, budget, and institutional constraints. The social planner partitions Boston into 14 neighborhoods. Based on historical data, the social planner models the number of K2 applicants from each neighborhood as the product of two normally distributed random variables. The first represents the overall number of applications and has mean 4,294 and standard deviation 115, reflecting the historical distribution in years 2010–2013 (four years of data).[18] The second represents the proportion of applicants from each neighborhood. The mean and standard deviation for each neighborhood is estimated using historical data and is shown in Table B.1 in Appendix B.

Each of the 14 neighborhoods is broken down further into geocodes, which partition the city into 868 small contiguous blocks. Since home location is the only allowable way to differentiate students, we use geocodes as agent 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 averages from years 2010–2013.

Each student is to be assigned to one of 77 schools. For each school s, there is a capacity limit m_s, which is the number of seats available 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 geocode for which s is the closest capacity school. For capacity school s, the limit m_s applies only 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, every student can at least be assigned to the closest capacity school.[19] There are 19 capacity schools distributed across the city.

The distribution of students across geocodes and the capacities of schools are plotted in Figure B.1, panel (a). The locations of the capacity schools are also shown. The capacities are the actual numbers from 2013.

5.1.2. Utility Prior for Students. Using historical choice data, the social planner estimates the following utility prior for student i in geocode t and school s:

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

where Distance_{ts} and Walk_{ts} are from the data and the coefficients κ_s, ω, and β are to be estimated from historical choices using maximum likelihood.

The coefficient κ_s represents "school quality," which encapsulates all common propensities that families observe before choosing a school (e.g., facilities, test scores, teacher quality, school environment). The variable Distance_{ts} is the distance from geocode t to school s in miles, estimated using walking distance according to Google Maps. The model is normalized so that the coefficient of Distance_{ts} is -1, which allows us to measure utility in the unit of miles, so one additional unit of utility can be interpreted as the equivalent of moving a school one mile closer. The variable Walk_{ts} is an indicator for whether geocode t is within the *walk zone* of school s, which is an approximately one-mile radius around. The variable ω represents additional utility for walk-zone schools, as these schools are in the immediate neighborhood and students can avoid busing; ϵ_{is} represents unobservable, idiosyncratic preferences, assumed to be independent and identically distributed (i.i.d.) standard Gumbel distributed;[20] and β represents the strength of the idiosyncratic preference.

[18] These statistics are for the first round of BPS applications only. This is the main round and represents over 80% of all applicants.

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

[20] 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 and K2 Noncontinuing, Regular Education Students

Parameter	Value	Interpretation
κ_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 farther. These values are graphically displayed in Figure B.1, panel (b). We normalize the smallest value to be 0.
ω	0.86	Additional utility for going to a school within the walk zone.
β	1.88	Standard deviation of idiosyncratic preference.

Notes. We use K1 data in fitting the utility model as well, since families face similar choices in the two grades, and more data allow 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).

We plot the school qualities in Figure B.1, panel (b) and tabulate the other coefficients in Table 1. Although more sophisticated demand models are possible, in this exercise, we treat the above as the "true" utility prior.

5.1.3. Objective and Constraints. The social planner's objective, W, is to maximize a linear combination of the average utility and the utility of the worst-off neighborhood (we also call these terms the *utilitarian welfare* and the *max-min welfare*, respectively):

$$W = \alpha \sum_t \left(w_t v_t + (1-\alpha) \min_{t'} v_{t'} \right).$$

Here, v_t is the expected utility of a student from geocode t, w_t is the proportion of students who live in geocode t (taking the expected number of students from each geocode and normalizing so that the weights sum to 1), and α is a parameter specifying the desired trade-off between efficiency and equity. Note that $\alpha = 1$ represents maximizing the average expected utility, $\alpha = 0$ represents maximizing the expected utility for the worst-off geocode, and $\alpha = 0.5$ represents an equal weighting of the two.

In addition to capacity constraints, the social planner faces a *busing constraint*. Since busing is needed only for students outside the one-mile walk zone, the amount of busing needed for a student from geocode t to school s is

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

Suppose that the social planner budgets C miles of busing per student in expectation; then the busing constraint[21] 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.[22]

The institutional constraint is that the social planner must design an ordinal mechanism that is incentive compatible, is "ex post Pareto efficient" within a geocode,[23] and requires only eliciting preference rankings, rather than preference intensities. The reason we consider only ordinal mechanisms is that the system has been ordinal from 1988 to 2012, so families are used to gathering and submitting preference rankings. The reason we require incentive compatibility is that in 2006, a nonincentive-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 concern that it might negatively affect their chances.[24] The reason we require ex post Pareto efficiency is to mitigate public discontent, as students in the same geocodes are likely to compare their assignments with one another.

These objectives and constraints, along with the above assumptions on the student population, the school capacities, and the utility prior, define a concrete allocation problem. We call this the *finite-market formulation.*

5.2. Baseline: Actual Implementation

One feasible solution to the finite-market formulation is the actual plan adopted by BPS after the reform, which is called the *Home Based* plan. It is based on a proposal by Shi (2013), although there are significant deviations. An input to this plan is a classification of BPS schools into four tiers according to standardized test scores, with Tier I being the best and Tier IV being the worst. Each student's choice menu consists of any school within one mile of his or her residence, plus a certain number of the 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*.

[21] In reality, busing cost is much more complicated, involving routing, the number of buses used, the kinds of buses used, and legal requirements for more expensive door-to-door busing for certain special education students. We leave finer modeling of transportation costs in Boston for future work.

[22] Note that this is a "soft" budget constraint in that it needs to be satisfied only in expectation. We use this rather than a hard budget constraint because typically, in school board operations, the initial budget is only a projection but may be revised later if needed.

[23] This means that students in the same geocode should not be able to trade *assignments* with one another, and all improve in utility. Note the difference from the definition of Pareto efficiency in our large-market formulation in §3, which considers the trading of probabilities, not just assignments. As the market is scaled up, the distinction between these definitions disappears.

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

The priority structure is as follows.[25] One of the 14 neighborhoods is East Boston. In the assignment plan, East Boston students get priority for East Boston schools, and non-East Boston students get priority for non-East Boston schools.[26] 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.

Having defined the menus and priorities, the plan computes the assignment by the algorithm presented below.

5.2.1. Deferred Acceptance Algorithm with Menus and Priorities. Given menus for each student and priority h_{ts} for geocode (type) t to school s, define a student's *score* to school s as $\sigma_{is} = r_i - h_{ts}$, where r_i is a Uniform$[0, 1]$ random variable, independently drawn across the students. Compute the assignment as follows:

Step 1. An arbitrary student i applies to her top choice s within her menu.

Step 2. School s tentatively accepts the student.

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

Step 4. Iterate Steps 1–3 until all unassigned students have applied to all their choices.[27]

It is well known that this algorithm does not depend on the order of students' application in Step 1 and that the induced mechanism is strategy-proof, which means that it is a dominant strategy for all students to report their truthful preference rankings.[28]

We simulated the baseline plan 10,000 times according to the assumptions described in §5.1, and we tabulated the plan's transportation burden, average expected utility, expected utility of the worst-off geocode, and predictability in Table 3, under the column "Baseline."

5.3. Defining the Large-Market Approximation

To apply our methodology, we first define a *large-market approximation* to the finite-market formulation. We do this by replacing each student with a continuum of infinitesimal students of mass 1. Instead of a stochastic mass of students from each geocode, we approximate the scenario with a deterministic mass n_t of students, setting n_t to be the expected value.

For the transportation budget, we use 0.6 miles, which is just under the 0.63 used in the plan chosen by the city, the baseline (see Table 3).

This yields exactly the environment in §4, since the capacity constraints and the busing constraint can be incorporated into the objective function by setting regions of constraint violation to negative infinity. Moreover, ex post Pareto efficiency is equivalent to ordinal efficiency in the large-market setting.[29]

5.4. Computing the Large-Market Optimum

In this section, we show how to compute the large-market optimum in the ordinal case when the utility priors take a multinomial-logit form. Our technique can be applied more generally, but we maintain the school choice language for ease of exposition.

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} values are i.i.d. standard Gumbel distributed.

This structure arises from multinomial logit discrete choice models, which in practice are widely used because of their tractability. Our utility prior in §5.1.2 takes this form.

The lottery-plus-cutoff characterization in §4.1 implies that we can optimize on the $|T||S|$ cutoffs. However, if one were to formulate the optimization in terms of the cutoffs, the resulting optimization is nonlinear. Moreover, the expected utilities have discontinuous first derivatives in the cutoffs at points where the cutoffs cross. It turns out that it is simpler to work with the equivalent characterization of randomized-menus-with-nested-menus from §4.1. This offers menu $M \subseteq S$ to an agent of geocode t with probability $z_t(M)$, and agents pick their most preferred school from their menus. For each type, the menus that are offered with positive probability are all nested within one another. (If M_1 and M_2 are both offered to agents of geocode t with positive probability, then either $M_1 \subseteq M_2$ or $M_2 \subseteq M_1$.)

For a menu of schools $M \subseteq S$, we abuse notation slightly and let $v_t(M)$ denote the expected utility of the best school in this menu for a student from geocode t,

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

[25] The actual plan also contains priorities for continuing students, students with siblings in a school, and students on a wait list from previous rounds. Since we do not model these complexities, our priority structure is simpler.

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

[27] Note that if all students include the closest capacity school in their rankings, then in the end, all students will be assigned.

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

[29] For works that study this equivalence, see Che and Kojima (2011) and Liu and Pycia (2013).

Let $p_t(s, M)$ denote the probability that a student from geocode t would choose school s to be the most preferred in menu $M \subseteq S$:

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

Let $z_t(M)$ denote the probability a student from geocode t is shown menu $M \subseteq S$. Let T_s denote the set of geocodes for which the capacity limit for school s applies.[30] If F_t is multinomial logit, then we have $v_t(M) = b_t \log(\sum_{s \in M} \exp(\bar{u}_{ts}/b_t))$, and $p_t(s, M) = \exp(\bar{u}_{ts}/b_t)/\sum_{s' \in M} \exp(\bar{u}_{ts'}/b_t)$.

By Theorem 2, to have ordinal efficiency within type, the menus must be nested. However, it is difficult to formulate this as a constraint. We will for now consider the following relaxed linear program (LP), which ignores nestedness. It will turn out that the nested structure will automatically be satisfied at optimality:

(LargeMarketLP)

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

There are $2^{|S|} - 1$ possible menus $M \subseteq S$; the number of variables of this LP is exponential in $|S|$. However, because of the multinomial-logit structure, the dual can be solved efficiently, and the menus given positive probabilities will be nested. This implies an efficient algorithm to solve for the optimal cutoffs.

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 exponentially sized* (LargeMarketLP) *can be found in time, polynomial in $|T|$ and $|S|$. Moreover, at any optimum, if $z_t(M_1) > 0$ and $z_t(M_2) > 0$, then either $M_1 \subseteq M_2$ or $M_2 \subseteq M_1$.*

The full proof is in Appendix C, but we give an overview here. In the dual, there are shadow prices for the budget and capacity constraints, which together define an "allocation cost" from geocode t to school s. The multinomial logit structure implies that if we put a school s in a menu, we would always include all the schools with a lower allocation cost as well. Given the shadow prices, one can find the optimal menu of each geocode simply by deciding how many schools to include. The larger the menu, the greater the expected utility, but the greater the expected allocation cost. The importance we put on the utility of a geocode t depends on its weight w_t in the utilitarian welfare and its dual variable in the equity constraint, which is higher for the worst-off geocodes. Since the menus are easy to compute given the prices, we only have to solve a convex optimization problem involving the prices, which can be done efficiently. Moreover, since all the possible optimal menus for each type are nested within one another by the above structure, the menus that are given positive probability in the optimal primal solution will be nested.

[30] For capacity schools, this is all geocodes not in the catchment region; for other schools, this is all geocodes.

Table 2 **Performance in the Large-Market Formulation 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

From the optimal solution to (LargeMarketLP), we get the optimal cutoffs for each school s to each geocode t: $a_{ts} = \sum_{M \ni s} z_t(M)$.

We evaluate the optimal plan in the large-market formulation using three settings of α in the objective: $\alpha = 1$ (utilitarian welfare), $\alpha = 0$ (max-min welfare), and $\alpha = 0.5$ (equal weighting of the two). We tabulate the results in Table 2. As can be seen, setting $\alpha = 0.5$ yields near optimal utilitarian welfare and max-min welfare, so for the remainder of this paper, we use $\alpha = 0.5$.

5.5. 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 in §5.2.1 and use the optimal large-market cutoffs a_{ts} estimated using previous years' data to guide the menus and priorities: for students of geocode t, define their menu to be schools for which the cutoff a_{ts} is positive and their priority for school s to be simply $h_{ts} = a_{ts}$. We then assign using these menus and priorities as in §5.2.1. We call the resulting mechanism the *DA analog* to the optimal large-market mechanism.

The DA analog is "asymptotically optimal" in the following sense: in the limit in which the students and school capacities are duplicated with many independent copies, the finite-market environment converges to the large-market approximation, and the DA analog converges to the large-market optimum.

Table 3 **Evaluating a Variety of Plans in the Finite-Market Formulation Using 10,000 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 of the worst-off geocode	2.86	4.53	7.05	7.02
% getting top choice	0.66	0.64	0.80	0.79
% getting top-three choice	0.88	0.85	0.94	0.93

This is because the independent copies wash away the stochasticity in the number of students from each geocode. Moreover, in this limit, a student i is assigned to school s in the DA analog if and only if her score is negative, which implies the same assignment probabilities as in the large-market optimum.

Observe that the DA analog is incentive compatible and ex post Pareto efficient within each geocode. It is incentive compatible because deferred acceptance is strategy-proof for the students when the menus and scores are exogenous (in our case, they are calculated from previous years' preference submissions). Ex post Pareto efficiency within geocode follows from priorities a_{ts} being the same for students of each geocode and from our using for each student the same random number r_i at all schools.[31]

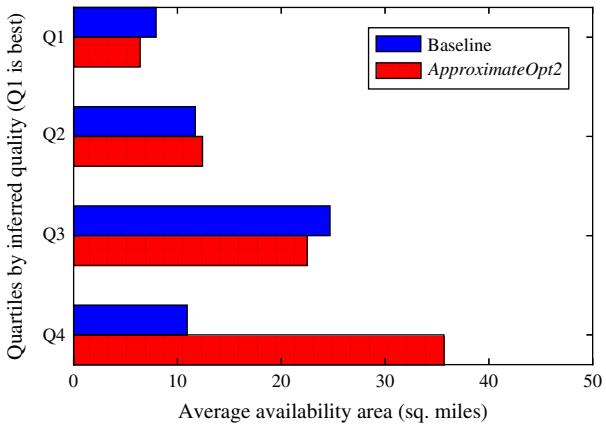
5.6. Numerical Results

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 10,000 times and compare its performance to the baseline in Table 3. It turns out that this plan evaluated in the finite-market formulation uses 0.71 miles of busing, which exceeds the 0.63 miles of the baseline. Thus, we let *ApproximateOpt2* denote the DA analog with $\alpha = 0.5$ and a busing budget of 0.50. In the finite-market simulations, *ApproximateOpt2* stays within the busing budget of the baseline and significantly improves it in terms of the average utility, the utility of the worst-off geocode, and the probability of students getting into their top or top-three choices. To give a frame of reference for the magnitude of the improvement, we also evaluate what we call the *most naïve* plan, which has no priorities and includes in the menu for each geocode only the capacity school and schools with zero transportation costs (schools in the walk zone).

As shown in Table 3, whereas the baseline uses 0.29 miles of additional busing per student over the most naïve, it improves the average utility by 0.64 miles and the utility of the worst-off geocode by 1.67 miles. However, *ApproximateOpt2* uses fewer miles of transportation while improving the average utility over the most naïve by 1.18 miles (almost doubling the gains) and improving the expected utility of the worst-off geocode by 4.16 miles (more than doubling the gains). It also significantly improves students' chances of getting into their top or top three choices.

To gain further insight, we compute for each school its "availability area" in each plan. This is the total area of the geocodes for which this school shows up in the menu. We then divide schools into quartiles by their quality κ_s, with Q1 being the best and Q4 being the worst. We compare the average availability areas for different quality quartiles in the baseline and *ApproximateOpt2* in Figure 3. As shown in the table, *ApproximateOpt2* promotes lower-quality schools by offering them 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

Figure 3 **(Color online) Comparison of Availability Areas for Schools of Various Quality in the Baseline and *ApproximateOpt2***



Note. Here, Q1 is the best quartile in quality; Q4 is the worst.

[31] To see this, consider all students of a given geocode and all the schools they are assigned to (counting multiplicity of seats) from deferred acceptance. Consider the student with the best (smallest) random number r_i out of these. Of these schools, this student must be assigned to her most preferred. (This follows from the student having a lower score than all other students from this geocode to all schools.) Hence, this student would not take part in any improvement cycle with other students from this geocode. Considering the partial market without this student and the assigned seat and proceeding by induction, we rule out all improvement cycles with students from this geocode.

areas. However, to compensate the students who do not live near such schools, the plan offers them many lower-quality schools, in hopes that the students will have high idiosyncratic personal affinity for them. Interestingly, the baseline mimics the same behavior for Q1, Q2, and Q3 schools but not for Q4 schools, which makes sense in retrospect, because some of the Q4 schools may be at risk of closure.

6. Discussion

This paper studies the allocation of goods and services to agents without monetary transfers. The social planner has priors over agents' utilities and is interested in maximizing a public objective. The approach in this paper sacrifices exact analysis by taking a continuum approximation to gain the analytical tractability needed to handle large-scale applications with complex objectives and many types of agents and 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.

In the large-market environment, we characterize incentive-compatible mechanisms that are Pareto optimal within each type, and we show how to compute the optimal ordinal mechanism when the utility prior follows a multinomial-logit structure. One open problem is the efficient computation of the optimal ordinal mechanism with other utility priors, as well as the efficient computation of the optimal cardinal mechanism.[32]

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. (Limits on busing are needed because the costs are borne by tax payers.) 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 the predictability of choice options for families.

One difficulty in implementing optimal mechanism design without money is in estimating utility distributions. With transfers, estimation becomes simpler because 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, but the underlying utilities are not observable, so one may question the model's validity. A fundamental question is whether behavior can indeed be captured with such models.[33] Nevertheless, *if* a utility model can be estimated, the methods used in this paper can be used to compute the optimal mechanism.

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

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mnsc.2015.2162.

Acknowledgments

The authors thank Itay Fainmesser, Steve Graves, and Özalp Özer for helpful discussions. I. Ashlagi acknowledges the research support of the National Science Foundation [Grant SES-1254768].

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 one mile straight-line distance;
- the closest two Tier I schools;
- the closest four Tier I or II schools;
- the closest six Tier I, II, or III schools;
- the closest school with Advanced Work Class (AWC);[34]
- the closest Early Learning Center (ELC);[35]
- the three closest capacity schools;[36]
- the three *citywide 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 K8 School in Allston–Brighton.

Appendix B. Additional Tables and Figures

Table B.1 shows the forecasted proportion of students applying from each neighborhood. A big picture view of the distribution of supply and demand for schools and of inferred school quality in Boston is given in Figure B.1, panel (a) and Figure B.1, panel (b), respectively.

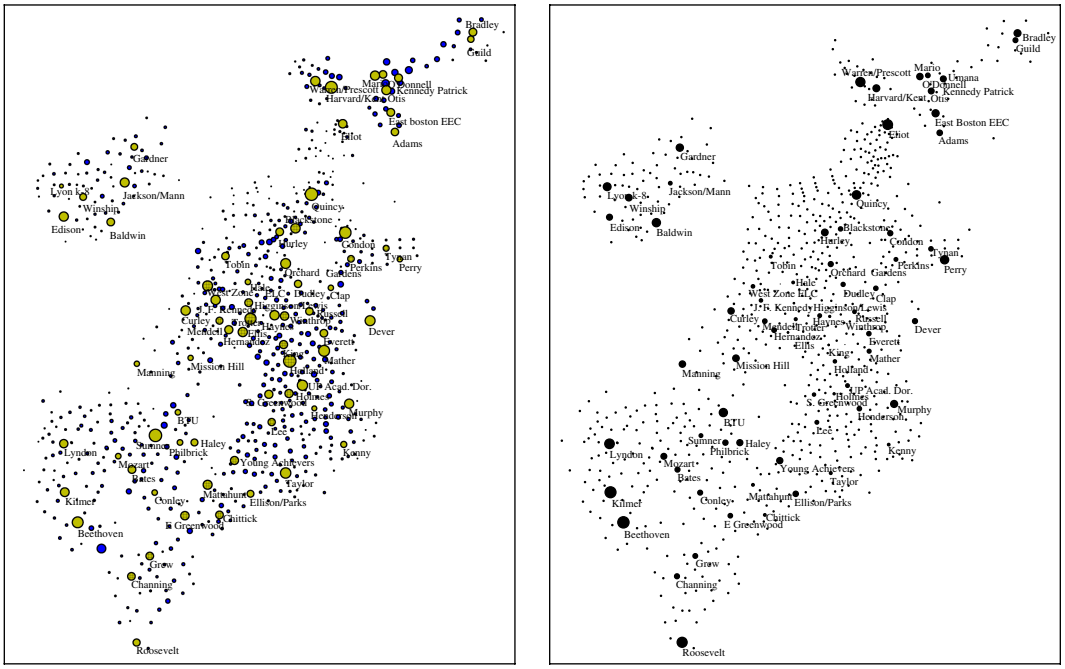
[33] One project that examines the validity of utility models in Boston school choice, compared with an alternative model based on marketing or salience, is Pathak and Shi (2015), in which the authors use various methods to predict how families choose schools after the 2012–2013 reform and evaluate the prediction accuracy of utility-based models.

[34] AWC is a full-time, invited program that provides an accelerated academic curriculum.

[35] ELCs are extended-day kindergartens.

[36] 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 IV schools.

Figure B.1 **(Color online) Diagrams Showing the (a) Distribution of Students and Capacities of Schools and (b) Estimates of κ_s (Inferred Quality of Schools)**



(a) Supply and demand (b) School quality

Notes. In (a), 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 capacity schools are shaded. The distribution of students is based on the 2010–2013 average. The capacities are based on data from 2013. In (b), the size of the circle is proportional to the estimated κ_s, based on the 2013 data, with higher-quality schools having larger circles.

Table B.1 **Means and Standard Deviations of the Proportion of K2 Applicants from Each Neighborhood, Estimated Using Four 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

Appendix C. Omitted Proofs

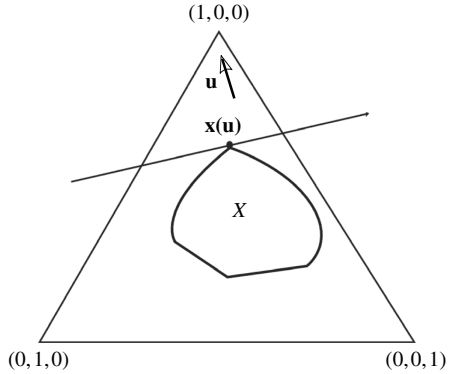
C.1. Characterization for Cardinal Mechanisms

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.

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 $\mathbf{x}$. This is illustrated in Figure C.1.*

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 allocation rule $\mathbf{x}$. Suppose X does

Figure C.1 **An Incentive-Compatible Allocation Rule with $|S|=3$**



Note. X is an arbitrary closed convex subset of the feasibility simplex Δ; $\mathbf{y}=\mathbf{x}(\mathbf{u})$ is a maximizer of the linear objective $\mathbf{u}\cdot\mathbf{y}$ with $\mathbf{y}\in X$.

not intersect the relative interior of Δ, $\text{int}(\Delta) = \{\mathbf{y} \in \mathbb{R}^{|S|}: \mathbf{y} > 0, \sum_s y_s = 1\}$; 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 $X \cap \text{int}(\Delta) \neq \varnothing$.

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

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

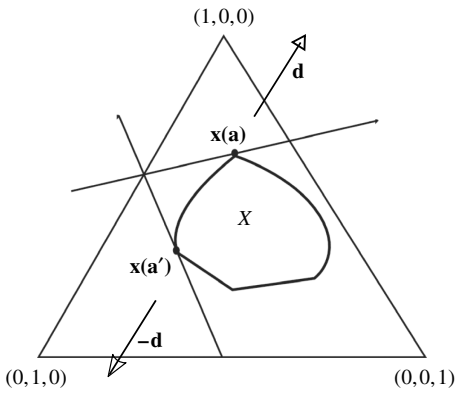
Recall that $D = \{\mathbf{u} \in U: \mathbf{u} \cdot \mathbf{1} = 0\}$ is the space of relative utilities. For any set $A \subseteq D$ (subset of the relative utilities), let $U(A) = \{\mathbf{u} \in U: \text{Proj}_D(\mathbf{u}) \in A\}$. This is the set of utilities for which the projection onto D 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 CEEI 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 $\text{int}(\Delta)$. Intuitively, if it has two different tangents $H(\mathbf{a}, 1)$ and $H(\mathbf{a}', 1)$, with nonzero 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 C.2. However, defining a feasible move with positive measure in all cases is nontrivial, 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, $\text{int}(\Delta)$. 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 C.2 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 $\text{int}(\Delta)$



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 $\text{int}(\Delta)$, with different nonzero unit projections onto U, $\tilde{\mathbf{a}} \neq \tilde{\mathbf{a}}'$. Take any unit vector $\mathbf{d} \in \text{int}(H^+(\mathbf{a}, 0) \cap H^-(\mathbf{a}', 0)) \cap D$ (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 Pareto efficiency of $\mathbf{x}(\cdot)$. Therefore, X has only one supporting hyperplane in Δ that intersects it in the interior $\text{int}(\Delta)$. □

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

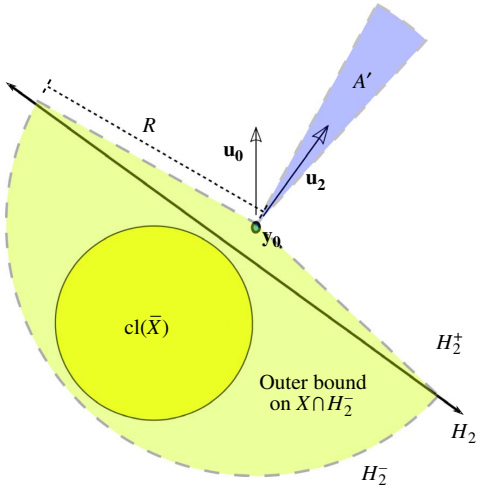
Conversely, if for some closed convex set X, for any $\mathbf{u} \in U$, $\mathbf{x}(\mathbf{u}) \in \arg\max_{\mathbf{y} \in X} \{\mathbf{u} \cdot \mathbf{y}\}$, then $\forall\, \mathbf{u}' \in U$, $\mathbf{x}(\mathbf{u}') \in X$, so $\mathbf{u} \cdot \mathbf{x}(\mathbf{u}) \geq \mathbf{u} \cdot \mathbf{x}(\mathbf{u}')$.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

Proof of Lemma 2. Any tangent hyperplane Y of X in $\text{aff}(\Delta)$ is a $(|S|-2)$-dimensional affine subset of the $(|S|-1)$-dimensional affine set $\text{aff}(\Delta)$. 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 $H = \{\mathbf{y}: \mathbf{a} \cdot \mathbf{y} = 1\}$ for some $\mathbf{a} > 0$, and by construction, $\mathbf{a}$ is not colinear with $\mathbf{1}$. Now, $\mathbf{a} \cdot \mathbf{z} = 1/(1+\epsilon) < 1$, so $\mathbf{a}$ points outward from X. □

In carrying out the exchange argument in Figure C.2, we need to guarantee that a positive measure of agents benefits 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 obtains 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 $\mathbf{d}$ may no longer be a strictly improving direction. 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 δ close to $\mathbf{x}(\mathbf{u}_0)$. 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 $\mathbf{x}(\mathbf{u}_0)$.

Figure C.3 **(Color online) 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 $\mathbf{x}(\mathbf{u})$ Cannot Be in $\mathrm{cl}(\bar{X})$**



Lemma 4. *Given any bounded closed convex set $X \subseteq \mathbb{R}^n$, any nonempty open cone $C \subseteq \mathbb{R}^n$ and any measurable function $\mathbf{x}\colon C \to \mathbb{R}^n$ 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 nonempty open cone $A \subseteq C$, we have $F(A) > 0$. Now define*

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

We have that for every $\mathbf{u} \in C$, $\arg\max_{\mathbf{y}\in X}\{\mathbf{u}\cdot\mathbf{y}\} \subseteq \mathrm{cl}(\bar{X})$.

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

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

By Lyapunov's convexity theorem, since F is atomless, the range of this measure, denoted by M, is convex. Therefore, the cone generated by M, $\mathrm{cone}(M) = \{\lambda\mathbf{x}\colon \mathbf{x} \in M, \lambda > 0\}$, is convex, and so its intersection with the hyperplane $(\cdot, 1)$ is convex (last component restricted to 1). This intersection is nonempty since $F(C) > 0$. Moreover, this intersection, restricted to the 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 $\mathbf{y}_0 \notin \mathrm{cl}(\bar{X})$. We will exhibit an open subset $A \subseteq C$ such that $\bar{\mathbf{x}}(A) \notin \bar{X}$. But $F(A) > 0$ by the openness of A, so $\bar{\mathbf{x}}(A) \in \bar{X}$ by definition of $\bar{X}$. This is a contradiction. The construction is given below. Because of its geometric nature, we refer readers to Figure C.3 for an illustration.

Since $\mathrm{cl}(\bar{X})$ is closed, convex, and bounded, there exists a strictly separating hyperplane $H(\mathbf{u}_1, \alpha_1)$ such that for some $\delta_1 > 0$,

$$\mathbf{u}_1\cdot\mathbf{y}_0 \geq \alpha_1 + \delta_1 > \alpha_1 - \delta_1 \geq \mathbf{u}_1\cdot\mathbf{y} \quad \forall \mathbf{y} \in \mathrm{cl}(\bar{X}).$$

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

$$\mathbf{u}_0\cdot\mathbf{y}_0 \geq \mathbf{u}_0\cdot\mathbf{y} \quad \forall \mathbf{y} \in \mathrm{cl}(\bar{X}).$$

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

$$\mathbf{u}_2\cdot\mathbf{y}_0 \geq \alpha_2 + \delta_2 > \alpha_2 - \delta_2 \geq \mathbf{u}_2\cdot\mathbf{y} \quad \forall \mathbf{y} \in \mathrm{cl}(\bar{X}),$$

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 nonpositive side of H_2. If $\mathbf{y} \in X \cap H_2^-$, then $\mathbf{y} \in B(\mathbf{y}_0, R) \cap H_2^-$, 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 $B(\mathbf{y}_0, R) \cap H_2^-$; namely,

$$A' = \left\{\mathbf{u} \in \mathbb{R}^n\colon \frac{\mathbf{u}\cdot\mathbf{u}_2}{\|\mathbf{u}\|\|\mathbf{u}_2\|} > \frac{\sqrt{R^2\|\mathbf{u}_2\|^2 - \delta_2^2}}{R\|\mathbf{u}_2\|}\right\}.$$

This construction is illustrated in Figure C.3. Note that $\mathbf{u}_2 \in A'$. Moreover, $\forall \mathbf{u} \in A'$, $\mathbf{x}(\mathbf{u})\cdot\mathbf{u} \geq \mathbf{y}_0\cdot\mathbf{u} > \mathbf{y}\cdot\mathbf{u}$ for every $\mathbf{y} \in X \cap H_2^-$. Thus $\mathbf{x}(\mathbf{u}) \notin H_2^-$. Let $A = A' \cap C$; then A is open since it is the intersection of two open sets, and A is nonempty 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 is 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 $\bar{X}$. □

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 $\tilde{\mathbf{a}}_0 = \mathrm{Proj}_D \mathbf{a}_0$, 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 4.

Let $\mathbf{y}_0 \in X \cap H(\mathbf{a}_0, 1) \cap \mathrm{int}(\Delta)$. ($\Delta$ 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 $r = (\tilde{\mathbf{a}}_0\cdot\mathbf{d}/(6\|\tilde{\mathbf{a}}_0\|))\epsilon$. Define the following cones:

$$C_1 = \left\{\mathbf{a} \in D\colon \frac{\mathbf{a}}{\|\mathbf{a}\|}\cdot(\mathbf{y}_0 - \mathbf{x}(\mathbf{a})) > -r\right\},$$

$$C_2 = \left\{\mathbf{a} \in D\colon \frac{\mathbf{a}}{\|\mathbf{a}\|}\cdot\mathbf{d} > \frac{\tilde{\mathbf{a}}_0}{2\|\tilde{\mathbf{a}}_0\|}\cdot\mathbf{d} = 3\frac{r}{\epsilon}\right\}.$$

The first represents relative utilities for which the allocation of $\mathbf{y}_0$ is "not too bad," and the second represents relative utilities that are "roughly" in the direction of $\mathbf{d}$. We want to show that $C = C_1 \cap C_2$ is a nonempty open cone, so that we can apply Lemma 4.

To see that $C = C_1 \cap C_2$ is a nonempty open cone, note that $\tilde{\mathbf{a}}_0 \in C_1$ and $\tilde{\mathbf{a}}_0 \in C_2$, so C is nonempty; C_1 and C_2 are cones because the expressions that define them depend only on $\mathbf{a}/\|\mathbf{a}\|$, so C is a cone. (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.) The two cones are open because the left-hand side 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 the bounded convex set X.) Therefore, C is open.

Now we apply Lemma 4. 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 $A \subseteq C$, $F(U(A)) > 0$. Moreover, 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 $\|\bar{\mathbf{x}}(U(A)) - \mathbf{y}_0\| \le r$. Now, define $\mathbf{y}_1 = \bar{\mathbf{x}}(U(A)) + \epsilon\mathbf{d}$; then $\mathbf{y}_1 \in \Delta$ since $\|\mathbf{y}_1 - \mathbf{y}_0\| \le \epsilon + r \le \frac{7}{8}\gamma$. 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 $\forall \mathbf{a} \in A$,

$$\begin{aligned} &\mathbf{a}\cdot\mathbf{y}_1 - \mathbf{a}\cdot\mathbf{x}(\mathbf{a}) \\ &\quad = \mathbf{a}\cdot(\bar{\mathbf{x}}(U(A)) + \epsilon\mathbf{d}) - \mathbf{a}\cdot\mathbf{x}(\mathbf{a}) \\ &\quad = \mathbf{a}\cdot(\bar{\mathbf{x}}(U(A)) - \mathbf{y}_0) + \mathbf{a}\cdot(\mathbf{y}_0 - \mathbf{x}(\mathbf{a})) + \epsilon\mathbf{a}\cdot\mathbf{d} \\ &\quad \ge -r\|\mathbf{a}\| - r\|\mathbf{a}\| + 3r\|\mathbf{a}\| \\ &\quad > 0. \end{aligned}$$

Now, let $\delta_0 = \epsilon F(U(A))$ for any $\delta \in (0, \delta_0)$ if we set $\mathbf{x}'(\mathbf{u}) = (\delta/\delta_0)\mathbf{y}(\mathbf{u}) + (1 - \delta/\delta_0)\mathbf{x}(\mathbf{u})$. 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 as a result 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 $f\colon 2^{|S|} \to [0,1]$ such that for every permutation $\pi \in \Pi$ and for every k $(1 \le k \le |S|)$,*

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

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

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:

$$\sum_{s\in M} x_s \le f(M) \quad \forall M \subseteq S, \tag{C1}$$

$$\sum_{s\in S} x_s = 1, \tag{C2}$$

$$x \ge 0. \tag{C3}$$

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 $M_1, M_2 \subseteq S$,*

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

Let $a_s = f(\{s\})$. An easy induction using the above lemma yields $\forall M \subseteq S$, $f(M) = \max_{s\in M} a_s$, which together with the expression in 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

$$f(M) = \sum_{j=1}^{|M|} x_{\pi(j)}(\pi), \quad \text{where } \{\pi(1), \pi(2), \ldots, \pi(|M|)\} = M.$$

This is well defined because incentive compatibility requires each agent's chances of getting a service in M, conditional on ranking these first in some order (ranking all of M before all of $S\backslash M$), to be fixed, regardless of the relative ranking between services in M and between 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 1 if $s \in M$ and 0 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 services 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 b/ϵ, she has incentive to misreport.

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. She prefers all of $M_1 \cup \{s\}$ before any of $M_2\backslash M_1$. Reporting this true ranking gives an expected utility of $bf(M_1 \cup \{s\})$. However, if she instead ranked M_1, then $M_2\backslash M_1$, then s, and 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 that f is monotone, and by the definition of f, we have that $\forall \pi \in \Pi$ and $1 \le k \le |S|$, $x_{\pi(k)}(\pi) = f(\{\pi(1), \ldots, \pi(k)\}) - f(\{\pi(1), \ldots, \pi(k-1)\})$. This proves the forward direction.

Conversely, if f is a monotone submodular set function, we show that if we define $\mathbf{x}$ so that $x_{\pi(k)}(\pi) = f(\{\pi(1), \ldots, \pi(k)\}) - f(\{\pi(1), \ldots, \pi(k-1)\})$, then $\mathbf{x}$ is incentive compatible. Note that the range of $\mathbf{x}$ defined this way is simply the vertex set of the base polytope of the polymatroid defined by f (see Equations (C1)–(C3)).

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 relabel S so that

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

then an optimal point of the base polytope has x_1 set to $f(\{1\})$, x_2 set to $f(\{1,2\}) - f(\{1\})$, and so on, which is

exactly our expression for the allocation rule $\mathbf{x}$. 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, $f(M_1 \cup M_2) \geq \max\{f(M_1), f(M_2)\}$. What we need to show is that $f(M_1 \cup M_2) \leq \max\{f(M_1), f(M_2)\}$. By monotonicity, it suffices to show this for the case in which $M_1 \cap M_2 = \varnothing$.

Suppose that, on the contrary, $f(M_1 \cup M_2) > \max\{f(M_1), f(M_2)\}$, $M_1 \cap M_2 = \varnothing$. 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; and 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, whereas 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, which contradicts the ordinal efficiency of $\mathbf{x}$. □

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 example uses the intuition that the value of a cardinal mechanism lies mostly in its ability to distinguish between an agent who has an extremely large relative preference for a service over another who has only a weak preference.

Let M and N be two positive real numbers with $M \geq \min(3, N^3)$ and $N \geq 1$. Suppose that there are three services. Service 1 has capacity $1/N^2$, whereas services 2 and 3 have capacity 1 each. Suppose there is only one type of mass 1, and $1/N$ of the agents have utilities $(M, 1, 0)$ and the remaining agents have utilities $(2, 1, 0)$.[37] The objective is to maximize the social welfare. An optimal cardinal mechanism charges price vector $(p, 1, 0)$, where price $p > 2$ differentiates the two types of agents. Agents have a virtual budget 1. The $1/N$ of agents would purchase the bundle $(1/N, 0, 1 - 1/N)$, whereas the other agents will opt for $(0, 1, 0)$. Hence, the social welfare is $M/N^2 + (1 - 1/N) > M/N^2$. 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 $(1/N^2, 1, 1)$ for the services, and every agent gets allocation $(1/N^2, 1 - 1/N^2, 0)$. The social welfare is $(1/N)(M/N^2 + 1 - 1/N^2) + (1 - 1/N)(2/N^2 + 1 - 1/N^2) \leq 3M/N^3$. So the ratio between the best cardinal and the best ordinal in this example is at least $N/3$, 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 be the capacity constraint of school s, μ_t be the constraint of menu probabilities summing to 1 for geocode t, and ν_t be the constraint enforcing the minimum constraint for geocode t. The dual is as follows:

$$\text{(Dual)} \quad \min\left\{C\gamma + \mathbf{m} \cdot \boldsymbol{\lambda} + \sum_{t \in T} \mu_t\right\},$$

$$\mu_t \geq (\alpha w_t + \nu_t) v_t(M) - n_t \sum_s p_t(s, M)(\mathbb{1}(t \in T_s)\lambda_s + \gamma B_{ts}) \quad \forall\, t \in T,\ M \subseteq S,$$

$$\sum_{t \in T} \nu_t \geq 1 - \alpha,$$

$$\gamma, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu} \geq 0.$$

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 student from geocode t contributes $\alpha w_t + \nu_t$ "credits" to the city, whereas assigning the student 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 a student from geocode 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" subproblem.

Definition 5. Given $\gamma, \boldsymbol{\lambda}, \boldsymbol{\nu}$, the *optimal menu subproblem* 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. *The function $\mu_t(\gamma, \boldsymbol{\lambda}, \boldsymbol{\nu})$ 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 $|T| + |S| + 1$ nonnegative 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 subproblem needs to optimize over exponentially many menus $M \subseteq S$. However, when preferences are multinomial logit, we can efficiently solve the subproblem. Moreover, the optimum has the nested structure that we need.

Lemma 8. *Under multinomial-logit utility priors, if $\alpha w_t + \nu_t > 0$, then the number of optimal solutions for the optimal menu subproblem is at most $|S|$ and can all be found in time $|S| \log |S|$. Moreover, the optimal menus will all be nested within one another.*

Proof of Lemma 8. Recall that having multinomial-logit utilities means that $u_{is} = \bar{u}_{ts} + b_t \epsilon_{is}$, where the $\bar{u}_{ts}$'s are averages for this school and geocode, and the ϵ_{is}'s are unobservable, idiosyncratic preferences, which are standard Gumbel distributed. Fix a geocode t. Let $h_s = n_t(\mathbb{1}(t \in T_s)\lambda_s + \gamma B_{ts})/((\alpha w_t + \nu_t) b_t)$, $z_s = \exp(\bar{u}_{ts}/b_t)$. Substituting in the formula for $v_t(M)$ and $p_t(s, M)$ from the multinomial-logit

[37] 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 standard normals drawn independently for each agent.

structure and simplifying, the optimal menu subproblem is equivalent to finding all solutions to

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

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:

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

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

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

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

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

We show that at an optimal solution, it must be that $y_k \in \{0, z_k\}$. Suppose, on the contrary, that $y_k \in (0, z_k)$. 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 is $g(y_k) = \log(d_1 + y_k) + d_2/(d_1 + y_k) - h_k$. The first and second derivatives are $g'(y_k) = (1/(d_1 + y_k))(1 - d_2/(d_1 + y_k))$ and $g''(y_k) = (1/(d_1 + y_k)^2)(2d_2/(d_1 + y_k) - 1)$, respectively.

Since y_k is an interior optimum, $d_2/(d_1 + y_k) = 1$, and so the second derivative is $g''(y_k) = 1/(d_1 + y_k)^2 > 0$, 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 $M_k = \{1, \ldots, k\}$ (the services are sorted in increasing order of h_s), and so we only need to search through $1 \le k \le |S|$. This can be done in $|S| \log |S|$ time as it is a linear search after sorting services in nondecreasing 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 we can 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 nondecreasing order and searching through the M_k's for $k \in \{1, \ldots, |S|\}$ still yields all optimal solutions. Moreover, because the M_k's are nested by definition, all the optimal menus will be nested within one another. □

Since the subproblems are efficiently solvable, we can efficiently solve the dual to arbitrary precision. If $\mathcal{M}_t$ is the solution set to the optimal menu subproblem for geocode t using the optimal dual variables, then an optimal primal feasible solution can be recovered using complementary slackness[38] by finding a feasible solution to the polynomial-sized LP:

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

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

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

$$\sum_{s,t} \sum_{M\in\mathcal{M}_t} n_t p_t(s, M) B_{ts} z_t(M) \le C \quad \text{with equality if } \gamma > 0,$$

$$z_t(M) \ge 0 \quad \forall t \in T, M \in \mathcal{M}_t. \quad \square$$

References

Abdulkadiroğlu A, Sönmez T (2003) School choice: A mechanism design approach. *Amer. Econom. Rev.* 93(3):729–747.

Abdulkadiroğlu A, Sönmez T (2013) Matching markets: Theory and practice. Acemoglu D, Arellano M, Dekel E, eds. *Advances in Economics and Econometrics: 10th World Congress, Volume 1: Economic Theory* (Cambridge University Press, New York), 3–47.

Abdulkadiroğlu A, Che Y-K, Yasuda Y (2015) Expanding "choice" in school choice. *Amer. Econom. J.: Microeconomics* 7(1):1–42.

Abdulkadiroğlu A, Pathak PA, Roth AE (2009) Strategy-proofness versus efficiency in matching with indifferences: Redesigning the NYC high school match. *Amer. Econom. Rev.* 99(5): 1954–1978.

Abdulkadiroğlu A, Pathak PA, Roth AE, Sönmez T (2006) Changing the Boston school choice mechanism. Boston College Working Papers in Economics 639, Boston College, Chestnut Hill, MA.

Ashlagi I, Shi P (2014) Improving community cohesion in school choice via correlated-lottery implementation. *Oper. Res.* 62(6): 1247–1264.

Aumann RJ (1964) Markets with a continuum of traders. *Econometrica* 32(1):39–50.

Azevedo E, Leshno J (2015) A supply and demand framework for two-sided matching markets. *J. Political Econom.* Forthcoming.

Bogomolnaia A, Moulin H (2001) A new solution to the random assignment problem. *J. Econom. Theory* 100(2):295–328.

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

Budish E, Cantillon E (2012) The multi-unit assignment problem: Theory and evidence from course allocation at Harvard. *Amer. Econom. Rev.* 102(5):2237–2271.

Caro F, Shirabe T, Guignard M, Weintraub A (2004) School redistricting: Embedding GIS tools with integer programming. *J. Oper. Res. Soc.* 55(8):836–849.

Chakravarty S, Kaplan TR (2013) Optimal allocation without transfer payments. *Games Econom. Behav.* 77(1):1–20.

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

Clarke S, Surkis J (1968) An operations research approach to racial desegregation of school systems. *Socio-Econom. Planning Sci.* 1(3):259–272.

Condorelli D (2012) What money can't buy: Efficient mechanism design with costly signals. *Games Econom. Behav.* 75(2):613–624.

[38] Because the convex primal problem is induced by an LP, it suffices to solve to a finite level of precision to guarantee that the right dual variables are positive when we stop the optimization. In practice, we may stop the optimization earlier and look for primal solutions that approximately satisfy complementary slackness.

Echenique F, Yenmez MB (2015) How to control controlled school choice. *Amer. Econom. Rev.* 105(8):2679–2694.

Ehlers L, Hafalir IE, Yenmez MB, Yildirim MA (2014) School choice with controlled choice constraints: Hard bounds versus soft bounds. *J. Econom. Theory* 153(September):648–683.

Erdil A, Ergin H (2008) What's the matter with tie-breaking? Improving efficiency in school choice. *Amer. Econom. Rev.* 98(3):669–689.

Gale D, Shapley LS (1962) College admissions and the stability of marriage. *Amer. Math. Monthly* 69(1):9–15.

Hartline JD, Roughgarden T (2008) Optimal mechanism design and money burning. *Proc. 40th Annual ACM Sympos. Theory Comput.* (*STOC '08*) (ACM, New York), 75–84.

Hoppe HC, Moldovanu B, Sela A (2009) The theory of assortative matching based on costly signals. *Rev. Econom. Stud.* 76(1): 253–281.

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

Kominers SD, Sönmez T (2015) Matching with slot-specific priorities: Theory. *Theoret. Econom.* Forthcoming.

Liu Q, Pycia M (2013) Ordinal efficiency, fairness, and incentives in large markets. Working paper, University of California, Los Angeles, Los Angeles.

Maglaras C, Zeevi A (2005) Pricing and design of differentiated services: Approximate analysis and structural insights. *Oper. Res.* 53(2):242–262.

Miralles A (2012) Cardinal Bayesian allocation mechanisms without transfers. *J. Econom. Theory* 147(1):179–206.

Miralles Asensio A, Pycia M (2014) Prices and efficient assignments without transfers. Working paper, Boston University, Boston. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2505241.

Myerson RB (1981) Optimal auction design. *Math. Oper. Res.* 6(1):58–73.

Pathak PA, Sethuraman J (2011) Lotteries in student assignment: An equivalence result. *Theor. Econom.* 6(1):1–17.

Pathak PA, Shi P (2015) Demand modeling, forecasting, and counterfactuals, part I. Working paper, Massachusetts Institute of Technology, Cambridge. http://arxiv.org/abs/1401.7359.

Perry O, Whitt W (2009) Responding to unexpected overloads in large-scale service systems. *Management Sci.* 55(8):1353–1367.

Pycia M (2014) The cost of ordinality. Working paper, University of California, Los Angeles, Los Angeles. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2460511.

Roth AE, Sotomayor M (1990) *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis* (Cambridge University Press, Cambridge, UK).

Russell J, Ebbert S (2011) The high price of school assignment. *Boston Globe* (June 12), http://www.boston.com/news/education/k_12/articles/2011/06/12/the_high_price_of_school_assignment/.

Schummer J, Vohra R (2007) Mechanism design without money. Nisan N, Roughgarden T, Tardos E, Vazirani VV, 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. Memo, Massachusetts Institute of Technology, Cambridge. http://www.mit.edu/~pengshi/papers/closest-types.pdf.

Sönmez T, Ünver MU (2010) Course bidding at business schools. *Internat. Econom. Rev.* 51(1):99–123.

Sutcliffe C, Board J, Cheshire P (1984) Goal programming and allocating children to secondary schools in reading. *J. Oper. Res. Soc.* 35(8):719–730.

Thomson W, Zhou L (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. *J. Econom. Theory* 57(1):160–175.