Optimal Priority-Based Allocation Mechanisms

Peng Shi
USC Marshall School of Business, pengshi@usc.edu

This paper develops a tractable methodology for designing an optimal priority system for assigning agents to heterogeneous items while accounting for agents' choice behavior. The space of mechanisms being optimized includes deferred acceptance and top trading cycles as special cases. In contrast to previous literature, I treat the inputs to these mechanisms, namely the priority distribution of agents and quotas of items, as parameters to be optimized. The methodology is based on analyzing large market models of one-sided matching using techniques from revenue management, and solving a certain assortment planning problem whose objective is social welfare. I apply the methodology to school choice and show that restricting choices may be beneficial to student welfare. Moreover, I compute optimized choice sets and priorities for elementary school choice in Boston, improving upon the results of Ashlagi and Shi (2015).

Key words: market design, one-sided matching, assortment planning, school choice.

1. Introduction

Priorities are used to allocate scarce items in many important contexts, such as public school choice, subsidized housing assignment, and cadaver organ allocation. In these applications, both the agents receiving the items and the items being allocated are heterogeneous, and a good allocation matches each item to an agent who values it highly, while maintaining certain notions of fairness. A challenge is that the value of assigning an item to a particular agent depends both on the observable characteristics of the agent as well as on the agent's private preferences.

An example of an allocation mechanism is the Gale-Shapley deferred acceptance (DA) mechanism, which is used for public school choice in Boston, New York City, Chicago, Denver, New Orleans, Washington DC, among other cities. The mechanism requires the school district to give each student a priority score to each school, which may take into account the student's home location, socio-economic status, test scores, and whether the student has siblings at the school. Each student also submits a ranking of schools, indicating his or her relative preferences among eligible options. Based on these inputs, the mechanism computes a stable matching, which means that no student is rejected by a school that either has leftover seats or has accepted another student with a lower priority score to the school. A competing mechanism that was used in New Orleans in 2012-2013 is top trading cycles (TTC), which interprets the priority scores differently and allows students to trade priorities among themselves. For the allocation of college dormitories, a mechanism that is often used is serial dictatorship (SD), which orders applicants in a list according to

their priorities and allows those at the top of the list to pick their favorite building whose capacity has not been depleted. All of these mechanisms are strategyproof, meaning that agents have no incentives to misreport their preferences.

While there is a large literature that studies the properties of the above mechanisms,[1] almost all previous works treat the priorities as exogenous inputs, whereas priorities in practice are often determined by policy makers, so can be part of the market design. For example, after changing from a non-strategyproof assignment mechanism to deferred acceptance (DA) in 2005, Boston Public Schools (BPS) changed the choice sets and priorities of students under the DA mechanism in 2013, so that students are assigned closer to home while maintaining equity of access (Shi 2015). As another example, the Organ Procurement and Transplantation Network (OPTN), which oversees the allocation of cadaver organs in the US, changed the priorities for allocating kidneys in 2014, so as to improve the longevity matching[2] of patients to kidneys and to improve access for highly sensitized patients, for whom it is difficult to find a compatible kidney (Israni et al. 2014). Both of these reforms were based on simulation analyses that compared several priority systems in terms of their induced outcomes (Pathak and Shi 2013, Thompson et al. 2004).

A natural question is whether a given priority system chosen by policy makers is close to optimal, or whether a similar system with modified parameters can achieve much better performance with respect to the metrics that the policy makers care about. This question cannot be addressed in a computationally tractable way using existing methodology because evaluating a given priority system in a realistic setting requires a complex simulation model, which needs to predict agents' choice behavior under the new priorities and calculate the implied allocations. One might be able to evaluate a few priority systems by simulation, but not all potential systems within any reasonably rich class. For example, the priority systems considered by Boston Public Schools (BPS) in 2013 are allowed to specify for each school a geographic region of home locations from which a student can access the school, and since there are many possible configurations of boundary lines and many schools, the number of possibilities is enormous. Furthermore, there are institutional constraints on what the priority system can depend on: for BPS, the priorities are allowed to depend on students' home location or sibling status, but not on their gender, race, or their preferences for schools. These restrictions are to ensure that students are not discriminated against based on protected characteristics, and that they are incentivized to truthfully report their preferences for schools. Furthermore, the policy makers in Boston preferred to continue to use the DA mechanism, since families are already familiar with such a system after many years of use. Finding the best priority system within such institutional constraints is a difficult problem for which no tractable methodologies existed prior to this paper.

[1] For reviews of the matching markets literature, see Abdulkadiroğlu and Sönmez (2013), and Vulkan et al. (2013).

[2] Longevity matching means to match the best quality organs to patients with the longest predicted survival.

1.1. Contributions

This paper develops a tractable framework for solving the aforementioned problem of optimizing priority systems in realistic settings. Concretely, I propose efficient algorithms for computing the optimal priority system to be used within the deferred acceptance (DA) mechanism, such that the priorities may depend only on a limited set of observable characteristics as determined by policy makers. I apply these techniques to real data from BPS, and quantify the optimality gap between the priority system chosen by the policy makers in 2013 and the best possible priority system that depends only on a student's neighborhood and random tie-breakers. The optimized priority system has a simple structure similar to the actually implemented system: each neighborhood is associated with a set of eligible schools, and the priority score of a student to a school is equal to a publicly announced constant that depends on the student's neighborhood and the school, plus a randomly generated tie-breaker for each student. The techniques are based on extending ideas and algorithms from revenue management (Liu and van Ryzin 2008, Rusmevichientong et al. 2010, Gallego and Topaloglu 2014, Li et al. 2015, Feldman and Topaloglu 2017) to analyze large market models of one-sided[3] matching (Bogomolnaia and Moulin 2001, Azevedo and Leshno 2016, Abdulkadiroğlu et al. 2015, Leshno and Lo 2018). While I focus on the DA mechanism for the empirical exercise, the framework can also be used to find optimal priorities in other mechanisms such as top trading cycles (TTC) or serial dictatorship (SD), although the conditions for tractability are more restrictive.

The first contribution is characterizing the set of allocation outcomes that can be achieved using DA, TTC or SD when the priorities can be arbitrarily chosen as long as they depend only on a given set of observable characteristics. For tractability, the theoretical results are derived in the context of a large market model with a continuum of agents, who are to be matched to one of several possible items, each of which has a fixed capacity. A priority-based allocation mechanism is a function that maps an agent's action and priorities to an assigned item. To model institutional constraints on what priorities can depend on, I assume that agents are classified into a finite number of segments based on observable characteristics, and the priorities of an agent can depend on the segment of the agent as well as on random tie-breakers, but not on anything else. Formally, a priority system specifies for each segment of agents a probability distribution from which priorities are drawn. This setup ensures that any mechanism based on such a priority system offers agents from the same segment the same opportunities from an ex-ante perspective. In addition, the priority system is allowed to restrict the amount of each item that can be allocated by setting a quota lower than the item's capacity. Theorem 1 shows that under mild regularity conditions, the set of allocation outcomes that can be achieved using the DA mechanism can be precisely described

[3] The matching markets studied in this paper are called one-sided because only agents may behave strategically based on their private preferences, whereas the items are under the control of the social planner.

as the feasible region of a linear program with $m2^n$ non-negative decision variables and $m+n$ additional constraints, where m is the number of segments and n is the number of items. The same feasible region is also the set of outcomes that can be achieved using any priority-based allocation mechanism, implying that DA is the most flexible mechanism within this class. Theorem 1 also shows that the set of outcomes achievable using TTC or SD is much more limited, being equal to the intersection of the above feasible region with certain non-linear constraints that restrict the subset of decision variables that can be non-zero.[4] To make the characterization practically useful, the proof of Theorem 1 constructs priority distributions and quotas that can be used to implement any feasible outcome under DA, TTC, and SD.

The second contribution is developing efficient algorithms to compute the optimal priority distributions and quotas under a wide class of objective functions. This optimization assumes that the social planner knows the distribution of preferences of each segment, as represented by an $(n+1)$-dimensional distribution of cardinal utilities for each of the n items plus an outside option. In the school choice example, each segment may correspond to a local neighborhood, and the utility distribution can be estimated from the preference rankings that students submitted in previous years.[5] When maximizing any objective function that is concave in the expected utilities and assignment probabilities of agents, an optimal priority system under DA can be found by solving a convex program, which is directly tractable if the number of items is small. When the number of items is large, the convex program can still be efficiently solved by column generation, as long as one can efficiently solve a certain assortment planning problem whose objective is social welfare. In school choice, this problem can be interpreted as finding an optimal set of schools for students from a given neighborhood, so as to maximize the total utility of these students minus the opportunity cost they impose on others for occupying limited resources. Theorem 2 derives efficient algorithms for solving this assortment planning problem under a multinomial logit (MNL) utility distribution, under a d-level nested logit utility distribution, and under a variant of the Markov chain based model of Blanchet et al. (2016). The algorithms can also handle certain cardinality constraints, which is important for the school choice application. These algorithms generalize those of Rusmevichientong et al. (2010), Gallego and Topaloglu (2014), Li et al. (2015), Xie (2016), and Feldman and Topaloglu (2017) from maximizing revenue to maximizing social welfare.

The third contribution is applying the above methodology to optimize choice sets and priorities in school choice. In Section 5, I illustrate using stylized examples that having more choices is not

[4] The presence of the non-linear constraints in characterizing the feasible outcomes under TTC and SD is why it is more difficult to compute the optimal priority system under these mechanisms compared to under DA.

[5] There is a large literature on how to estimate students' utility distributions using revealed preference data. See Agarwal and Somaini (2019) for a recent review.

necessarily better for student welfare: in some cases limiting students to their own neighborhood schools may maximize utilitarian welfare, even if one assumes perfect rationality, symmetric utility distributions, no peer effects, and no transportation costs. The reason is that a student's choice to go to a school does not account for the externality imposed on the student who is displaced. The results can also be interpreted as giving examples in which the DA mechanism under neighborhood priorities outperforms the TTC mechanism under the same priorities. In Section 6, I compute the optimal choice sets and priority distributions for elementary school assignment in Boston. While the optimization is based on the continuum model, the performance is demonstrated in a simulation model with discrete agents and stochastic demand, similar to that used by the school board during the 2013 reform (Pathak and Shi 2013, Shi 2015). The results improve upon those of Ashlagi and Shi (2015) by obtaining similarly good performance in student welfare, assignment predictability, and average travel distance, while vastly reducing the cost of school busing, as proxied by the areas schools have to cover to pick up students and the average number of schools in the choice sets requiring busing. These metrics were salient in the 2013 reform and their tractable optimization is now possible as a result of the new algorithms in Theorem 2.

1.2. Relationship to Literature

This paper contributes to the growing literature applying optimization to the design of matching markets without monetary transfers. The most related works are Su and Zenios (2006) and Ashlagi and Shi (2015), which also model the choice behavior of agents and derive an optimal allocation mechanism based on the solution to a linear program (LP). The distinction is that the space of mechanisms being optimized is different in those papers. Su and Zenios (2006) study the allocation of cadaver kidneys of heterogeneous quality to patients who are differentiated by a privately known willingness-to-wait for higher quality organs. The mechanisms they consider are those that pool kidneys of various qualities into queues and allow patients to choose a preferred queue based on the desired trade-off between waiting time and quality. Ashlagi and Shi (2015) study the allocation of heterogeneous items to agents who are differentiated both by a publicly known type and a privately known utility vector. They optimize over the space of ordinal mechanisms[6] that are incentive compatible and ordinal efficient within type,[7] which in their model is equivalent to the

[6] An ordinal mechanism is one that only elicits agents' preference rankings but not their preference intensities. Ashlagi and Shi (2015) also study cardinal mechanisms, which are allowed to elicit preference intensities, but they only optimize over ordinal mechanisms.

[7] Ordinal efficiency within type means that no coalition of agents within a given type can trade probabilities and all improve in a first-order stochastic dominance sense. This is a form of a Pareto efficiency assumption that is only required to hold within each agent type.

space of DA mechanisms with a single tie-breaker (DA-STB).[8] In contrast, the space of mechanisms considered in this paper includes DA mechanisms with arbitrary tie-breaking rules, and is thus strictly larger than the space of ordinal mechanisms considered in Ashlagi and Shi (2015), and disjoint with the space of queuing mechanisms studied in Su and Zenios (2006). Nevertheless, I use the same data as in Ashlagi and Shi (2015) in the empirical portions and the same discrete simulation engine, so as to be able to compare results. Furthermore, Theorem 2 generalizes the computational efficiency results in Ashlagi and Shi (2015) for the MNL utility distribution to much richer utility distributions and constraints.

Other papers that apply optimization in school choice include Ashlagi and Shi (2014), Feigenbaum et al. (2020), and Bodoh-Creed (2020), all of which optimize another policy lever than what is considered here. Ashlagi and Shi (2014) optimize the assignment of students to schools so as to maximize the chance students from the same neighborhood attend the same school, while keeping fixed everyone's assignment probability to every school. Feigenbaum et al. (2020) optimize the assignment of tie-breakers across successive rounds of school assignment so as to minimize the change in assignment as the DA mechanism is rerun after previously assigned students drop out to opt for private schools. Bodoh-Creed (2020) optimizes the assignment of students to schools given the submitted preference rankings of every student, subject to certain constraints imposed by incentive compatibility and exogenously given priorities. To implement his framework, the school board would need to change the assignment algorithm to one that solves a sophisticated linear program. In contrast, I allow the school board to continue using the DA mechanism, but optimize the priority system that is used as input.

Previous papers that study the design of priority distributions in the DA mechanism focus on comparing two specific alternatives: using a single tie-breaker (STB) for each student at all schools, or using multiple tie-breakers (MTB) that are drawn independently for each student at each school. Abdulkadiroğlu et al. (2009) conduct simulations using data from New York City and observe that DA-STB gives more students their top choice, while DA-MTB leaves fewer students unassigned. Arnosti (2016) and Ashlagi et al. (2019) give theoretical explanations of this observation by analyzing large market models under different asymptotic assumptions.[9] Jeong (2018) argues that DA-MTB is better than DA-STB in promoting school diversity.

[8] The characterization result in Ashlagi and Shi (2015) on ordinal mechanisms is closely related to an earlier insight due to Liu and Pycia (2016) that all symmetric, asymptotically Pareto efficient, and asymptotically strategy-proof mechanisms lead to the same allocation in large markets. Pycia (2019) extends these ideas and show that if one evaluates mechanisms only using anonymous summary statistics, then even non-symmetric mechanisms would look nearly identical, as long as the mechanisms are all Pareto efficient and strategy-proof. This paper does not assume any notion of Pareto efficiency, and thus the mechanisms studied are outside the characterizations in Liu and Pycia (2016), Ashlagi and Shi (2015) and Pycia (2019).

[9] Ashlagi and Nikzad (2017) refine the story by relating the comparison of DA-STB and DA-MTB to the popularity of schools: for students assigned to popular schools that are over-demanded, DA-STB dominates DA-MTB. On the other hand, for students assigned to non-popular schools, neither dominates the other.

Another strand of literature studies how to implement quotas for various types of students in order to more effectively implement affirmative action and enforce social-economic diversity (Kojima 2012, Hafalir et al. 2013, Ehlers et al. 2014, Kominers and Sönmez 2016, Dur et al. 2018). In contrast, this paper does not consider quotas that apply only to a certain subset of students: such policy tools are redundant in a large market model with highly flexible priorities, since one can always manipulate the distributions of priorities to accomplish the same goals.

2. Model: Priority-Based Allocation Mechanisms

There is a unit mass of infinitesimal agents, each of whom is to be matched to one of n possible items. Let the items be indexed by $j \in [n] := \{1, 2, \cdots, n\}$. Multiple agents may be assigned to the same item, and the maximum mass of agents that can be assigned to item j is given by a capacity c_j. Unmatched agents receive an outside option, which I refer to as item 0, with capacity $c_0 = \infty$. Let $J = [n] \cup \{0\}$ denote the set of items including the outside option. For succinctness, I refer to a particular agent using the pronoun "he," and the social planner using the pronoun "she."

Based on institutional constraints on what priorities can depend on, the social planner classifies agents into m possible market segments based on observable characteristics, with $\lambda_t > 0$ denoting the mass of agents of segment $t \in [m]$. While agents from one segment can receive a systematically different treatment than agents of another segment, agents from the same segment must be treated in an identical way from an ex-ante perspective.

Each agent i has a $(n+1)$-dimensional utility vector $u_i \in \Theta$, where the component u_{ij} denotes his cardinal utility for being matched to item $j \in J$. Assume that preferences are always strict, meaning that Θ is the subset of $\mathbb{R}^{n+1}$ in which no two components are equal. For an agent of segment t, his utility vector u_i is drawn from a segment-dependent utility distribution F_t, which is a probability measure on Θ. A market is summarized by the tuple $M = (m, n, \lambda, c, F)$.

A matching μ for a market M specifies a function $\mu_t : \Theta \times J \to [0, 1]$ for each agent segment t, where $\mu_t(u, j)$ specifies the probability that an agent with utility vector $u \in \Theta$ matches to item $j \in J$. The function must satisfy the following constraints: the probabilities must add up to one, $\sum_{j \in J} \mu_t(u, j) = 1$ for each $u \in \Theta$, and the item capacities must be respected.

$$\text{(Capacity constraint)} \qquad \sum_{t \in [m]} \lambda_t p_{tj}(\mu) \leq c_j \qquad \text{for each } j \in [n], \tag{1}$$

$$\text{where} \qquad p_{tj}(\mu) := \int_u \mu_t(u, j)\, dF_t. \tag{2}$$

$p_{tj}(\mu)$ is the proportion of segment t agents assigned to item $j \in J$ under the matching μ.

One innovation of the model in this paper is the notion of a priority-based allocation mechanism, which I formally define in this paragraph and give an intuitive explanation in the next: A priority-based allocation mechanism for a market M is characterized by a tuple $X = (A, \Pi, G, x)$, where A is

a finite set of possible actions, Π is an arbitrary set of possible priorities, G_t is a probability measure on Π that is indexed by the agent segment t, and $x : A \times \Pi \to J$ is an allocation function that maps each action $a \in A$ and priority $\pi \in \Pi$ to an assignment outcome $x(a, \pi) \in J$. The allocation function x must satisfy the following two constraints: First, an agent must always be allowed to take his outside option: for any priority $\pi \in \Pi$, there exists an action $a \in A$ such that $x(a, \pi) = 0$. Second, the allocation must not violate item capacities. Formally, let the matching induced by market M and mechanism X be defined as $\mu^{M,X}$, where

$$\mu_t^{M,X}(u, j) := \mathbb{P}_{\pi \sim G_t}\left(u_j = \max_{a \in A} u_{x(a,\pi)}\right). \tag{3}$$

The allocation function x is such that $\mu^{M,X}$ satisfies the capacity constraint (1). Note that the probability expression in (3) uses the assumption that utilities for different items are never equal.

A mechanism $X = (A, \Pi, G, x)$ can be interpreted as follows. Each agent i of segment t is given a priority π_i drawn independently from the distribution G_t, and chooses an action $a \in A$. His assignment is a deterministic function of his action and priority. Equation (3) says that each agent always chooses the utility-maximizing action given his priority realization. This property would be satisfied if the agent knows his priority realization π at the time of choosing his action, or if the function x is such that the utility maximizing action a is independent of the priority realization π, such as in the deferred acceptance or top trading cycles mechanisms as described in Section 2.1.[10]

Conceptually, one can think of the assignment for each agent as also dependent on the aggregate distribution of actions and priorities of other agents. However, in the continuum model, the aggregate behavior of the market is deterministic, so can be built into the allocation function x. Nevertheless, it is convenient to describe preferences and priorities using probability distributions as this implies certain anonymity and fairness constraints: the social planner does not know the preferences of individual agents when designing the mechanism, and agents from the same segment must be given the same opportunities from an ex-ante perspective.

The social planner's goal is to find a priority-based allocation mechanism $X = (A, \Pi, G, x)$ for a given market $M = (m, n, \lambda, c, F)$ in order to maximize a certain function of the induced matching $\mu^{M,X}$. Before Section 4, the analysis does not depend on any specific objective function. Section 5 focuses on maximizing the utilitarian welfare, which is the aggregate expected utility of all agents. In Section 6, the school choice optimization for Boston incorporates not only utilitarian welfare, but measures of equity and transportation cost. Appendix C shows how to efficiently maximize any objective function that is jointly concave in the expected utility of each agent segment and the assignment probabilities $p_{tj}(\mu)$.

[10] An example of a mechanism outside of the framework in this paper the Boston mechanism used for school choice (Abdulkadiroğlu and Sönmez 2003, Abdulkadiroğlu et al. 2011), as the optimal action for an agent depends on his priority realization, which he does not know when choosing his action. However, if the Boston mechanism were to reveal to agents their priority realizations before they choose their actions, then it would be within the framework.

2.1. Examples of Mechanisms

2.1.1. Serial Dictatorship (SD) The mechanism is parameterized by segment-dependent priority distributions G_t and by a n-dimensional quota vector q. Each G_t is a continuous probability measure on $\Pi = [0,1]$. The quota q_j specifies the mass of item j to allocate, and may be lower than the capacity c_j, thus allowing the social planner to withhold supply if desired.

Agents are selected in order of decreasing priority and each selected agent picks his favorite item whose quota has not been depleted by previously selected agents. Formally, the set of actions is $A = J$, and for each action $a \in J$ and priority $\pi \in [0,1]$, the allocation $x(a,\pi)$ depends on a n-dimensional cutoff vector $z^{SD(M,G,q)}$, whose jth component represents the lowest priority needed to access item j:

$$x(a,\pi) = \begin{cases} a & \text{if } a \in [n] \text{ and } \pi \geq z_a^{SD(M,G,q)}, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The cutoff vector $z^{SD(M,G,q)}$ is uniquely determined by the utility and priority distributions, the mass of each segment, and the quota of each item. Appendix A gives an algorithmic description of $z^{SD(M,G,q)}$ based on Bogomolnaia and Moulin (2001).

Random serial dictatorship (RSD) is the special case in which all agent segments share the same priority distribution, which implies that agents are ordered uniformly randomly regardless of their segment. Thus, RSD only needs to be parameterized by a quota vector q.[11]

2.1.2. Agent-Proposing Deferred Acceptance (DA) The mechanism is parameterized by a n-dimensional quota vector $q \leq c$, and by segment-dependent priority distributions G_t, each of which is a probability measure over the space of priorities Π. Unlike in SD, Π is not one-dimensional, but is a bounded subset of $\mathbb{R}^n$. The priority realization of an agent i is a vector $\pi_i \in \Pi$, where the component π_{ij} denotes his priority score for item j, with higher values being better.

The action space A is the set of permutations over $J = [n] \cup \{0\}$, so that each action $a \in A$ specifies the agent's complete preference ranking over items, with the understanding that relative rankings among items ranked worse than the outside option will not be considered. To denote a permutation, let $a_j = 1$ if item j is the agent's first choice, $a_{j'} = 2$ if j' is the agent's second choice, and so on. Hence, $\arg\min_{j \in S}\{a_j\}$ denotes the agent's favorite item among the set $S \subseteq J$.

Given the market M and the priority distributions G, the (agent-proposing) DA mechanism specifies a n-dimensional cutoff vector $z^{DA(M,G,q)}$, with the jth component denoting the minimum priority needed to access item j. The allocation function is simply to match each agent to his favorite item for which his priority score is at least equal to the cutoff:

$$x(a,\pi) = \arg\min_{j \in J} \left\{ a_j : j = 0 \text{ or } \pi_j \geq z_j^{DA(M,G,q)} \right\}. \tag{5}$$

[11] In random serial dictatorship, since everyone shares the same priority distribution G, we can without loss of generality assume that $G = \text{Uniform}(0,1)$, since only relative priorities matter.

The cutoff vector $z^{DA(M,G,q)}$ is the result of the following iterative procedure: For each $j \in [n]$, initialize the priority cutoff z_j to be the lowest possible priority score for the item.

1. Each agent i applies to his favorite item among the set specified in (5), which includes his outside option 0 and any item j for which his priority score meets the cutoff, $\pi_{ij} \geq z_j$.
2. For each item $j \in [n]$, if the mass of current applicants exceeds the quota q_j, then increase the cutoff z_j by the smallest amount so that the mass of applicants whose priority score meets the cutoff is exactly q_j. If any cutoff is updated, go back to Step 1.

The algorithm terminates when the cutoff vector is unchanged between two rounds of iteration. In a discrete market, the algorithm always terminates after finitely many iterations. In a continuous market, the algorithm might not terminate, but Azevedo and Leshno (2016) and Abdulkadiroğlu et al. (2015) show that the cutoff vector z converges to a limit, denoted by $z^{DA(M,G,q)}$. A formal definition of the cutoff vector is given in Appendix A.

A special case of the mechanism is DA with single tie-breakers (DA-STB), which is parametrized by the quota vector q and by a deterministic $m \times n$ matrix of priority boosts b. The priority score π_{ij} of an agent i of segment t for item j is defined as $\pi_{ij} = b_{tj} + \delta_i$, where b_{tj} is denotes the priority boost of segment t agents for item j and $\delta_i \sim \text{Uniform}(0, 1)$ is a random tie-breaker that is common for the same agent across all items.

The agent-proposing DA mechanism is strategyproof for the agents (Dubins and Freedman 1981, Roth 1982, Abdulkadiroğlu and Sönmez 2003), which implies regardless of an agent's priority realization π, the optimal action is to submit his true preference ranking. Hence, the mechanism fits within the framework of priority-based allocation mechanisms.

2.1.3. Top Trading Cycles (TTC) As with DA, this mechanism is parameterized by a quota vector $q \leq c$ and segment-dependent priority distributions G_t, each of which is a measure on the space of priorities $\Pi \subseteq \mathbb{R}^n$. The action space A is the set of permutations of J. The difference with DA is that agents can trade priorities among themselves, so an agent with a good priority score to a very popular item can trade in his priority to obtain almost any item. For technical reasons, Π is assumed to satisfy additional regularity conditions as given in Assumption 1 of Appendix A.

In a model with discrete agents, the TTC mechanism repeatedly iterates the following procedure until every agent is matched:

1. Each agent points to his favorite item among those whose quota has not yet been depleted. If an agent's favorite item is his outside option, he is immediately matched to it.
2. Each non-depleted item points to the unmatched agent with the highest priority for the item. Within the set of directed arrows between agents and items, there must be at least one cycle. Match each agent in this cycle to the item he points to, and decrease the quota of every item in this cycle by 1. If not every agent has been matched, go back to the above step.

The version of the TTC mechanism considered in this paper is the one derived by Leshno and Lo (2018) for a continuum model, which they show approximates the outcome of the discrete TTC mechanism in large finite markets. The allocation function for this mechanism can be written as

$$x(a,\pi) = \arg\min_{j\in J} \left\{ a_j : j = 0 \text{ or } \pi_k \geq z_{jk}^{TTC(M,G,q)} \text{ for some } k \in [n] \right\}, \tag{6}$$

where $z_{jk}^{TTC(M,G,q)}$ specifies the minimum priority score for item k that an agent needs in order to trade in that priority to obtain item j. I give a precise definition of the priority cutoffs $z^{TTC(M,G,q)}$ in Appendix A. As with DA, the TTC mechanism is strategyproof for agents (Abdulkadiroğlu and Sönmez 2003), which implies that agents do not need to know their priority realizations to choose an optimal action.

3. Solution Technique for Optimal Mechanism Design

3.1. Linear Program for Optimal Budget Set Probabilities

Given a priority-based allocation mechanism $X = (A, \Pi, G, x)$, define the budget set B_π^X corresponding to a priority realization $\pi \in \Pi$ to be the set of possible items an agent with priority π can be matched to under some action $B_\pi^X := \{x(a,\pi) : a \in A\}$.

Definition 1. Given a market M and a mechanism X, define the mechanism's budget set probabilities as a $m \times 2^{n+1}$ non-negative matrix y^X, in which the entry $y_{tS}^X \in [0,1]$ denotes the probability that an agent of segment $t \in [m]$ receives a priority that yields the budget set $S \subseteq J$: $y_{tS}^X := \mathbb{P}_{\pi\sim G_t}(S = B_\pi^X)$.

Note that all relevant information for determining the matching $\mu^{M,X}$ defined in (3) is encoded in the budget set probabilities y^X. Moreover, the same matching can be implemented using the following reduced-form mechanism: offer each agent of segment t a budget set S with probability y_{tS}, and let him choose his favorite item within the offered set. I call this the random assortment mechanism with assortment probabilities y.

The problem of finding the optimal priority-based allocation mechanism is equivalent to finding the optimal budget set probability matrix y, which can be formulated as a finite-dimensional mathematical program. Define $U_t(S)$ to be the expected utility of a segment t agent under budget set $S \subseteq J$, and $P_t(j,S)$ is the probability that the agent's favorite item within S is j:

$$U_t(S) := \mathbb{E}_{u\sim F_t}\left[\max_{j\in S} u_j\right], \tag{7}$$

and

$$P_t(j,S) := \mathbb{1}(j \in S)\mathbb{P}_{u\sim F_t}\left(u_j = \max_{j'\in S} u_{j'}\right). \tag{8}$$

The following linear program (LP) finds the budget set probability matrix y that maximizes utilitarian welfare. All summations of segment t are over $[m] := \{1, 2, \cdots, m\}$ and all summations of set S are over the power set 2^J.

$$\text{Maximize}_y \quad \sum_{t,S} \lambda_t U_t(S) y_{tS} \tag{9}$$

$$\text{s.t.} \quad y_{tS} \geq 0 \tag{10}$$

$$\text{(Capacity)} \quad \sum_{t,S} \lambda_t P_t(j,S) y_{tS} \leq c_j \qquad \text{for each item } j \in [n]. \tag{11}$$

$$\text{(Valid probabilities)} \quad \sum_{S} y_{tS} = 1 \qquad \text{for each segment } t \in [m]. \tag{12}$$

$$\text{(Outside option)} \quad y_{tS} = 0 \qquad \text{if } S \not\ni 0. \tag{13}$$

Definition 2. Define the set Y^M of feasible budget set probabilities to be the feasible region of the above LP. This is the set of $m \times 2^{n+1}$ matrices satisfying Constraints (10)-(13).

The above LP is analogous to the choice-based linear program from the network revenue management literature (Gallego et al. 2004, Liu and van Ryzin 2008), which optimizes the probability of offering a customer each subset S of products, so as to not violate the capacity constraint of each item. The difference is that the objective function (9) represents not revenue but social welfare.

Ignoring the variables that are always zero by (13), the number of decision variables in the above LP is $m2^n$, which is manageable when n is small. I show in Section 4 that even when n is large, the LP can be efficiently solved if the utility distribution satisfies certain parametric assumptions.

3.2. Implementation of Desired Budget Set Probabilities

Definition 3. A budget set probability matrix $y \in Y^M$ is said to be implemented by a mechanism X if $y = y^X$, where y^X is defined in Definition 1 and Y^M in Definition 2. Moreover, y is said to be implementable using a given class of mechanisms if there exists a mechanism within the class that implements y.

For example, a budget set probability matrix $y \in Y^M$ is implementable using the class of DA mechanisms if there exists a priority distribution G and a quota vector q that gives the budget set S to a segment t agent with probability y_{tS}. Every $y \in Y^M$ is implementable using the class of random assortment mechanisms. One obstacle to implementability using DA is the multiplicity of stable matchings, as DA only finds the agent-optimal stable matching. However, Azevedo and Leshno (2016) have shown that multiplicity of stable matchings almost never occurs in a large market model, and one technical assumption that rules them out is given in Definition 4. This assumption is satisfied for example if for each segment, any preference ranking of items is possible.

Definition 4. A market $M = (m, n, \lambda, c, F)$ is said to be regular if in each segment, the probability that the outside option is an agent's favorite item among a given set S is strictly decreasing in the set S: if $\{0\} \subseteq S \subsetneq S'$, then $P_t(0,S) > P_t(0,S')$ for every $t \in [m]$.

The following theorem characterizes whether a budget set probability matrix y is implementable using each of the classes of mechanisms described in Section 2.1, depending on whether y satisfies certain easily verifiable conditions stated in Definition 5.

DEFINITION 5. A budget set probability matrix $y \in Y^M$ is said to be

a) nested within segment if for all t, $y_{tS} > 0$ and $y_{tS'} > 0$ implies that either $S \subseteq S'$ or $S \supseteq S'$;

b) nested if for all t and t', $y_{tS} > 0$ and $y_{t'S'} > 0$ implies that either $S \subseteq S'$ or $S \supseteq S'$;

c) non-degenerate if each item that is allocated with zero probability is either present in all budget sets or absent from all: let $\mathcal{A} = \{S : y_{tS} > 0 \text{ for some } t\}$, then if item $j \in [n]$ is such that the left hand side of (11) is zero, then either $j \in S$ for all $S \in \mathcal{A}$, or $j \notin S$ for all $S \in \mathcal{A}$.

THEOREM 1 **(Characterization of Mechanisms)**. *Let Y^M be the set of feasible budget set probabilities as in Definition 2.*

a) If a market M is regular, then

i) any budget set probability matrix $y \in Y^M$ is implementable using the class of deferred acceptance (DA) mechanisms.

ii) a budget set probability matrix $y \in Y^M$ is implementable using the class of DA-STB mechanisms if and only if y is nested within segment.

b) For any market M, a budget set probability matrix $y \in Y^M$ is implementable using the class of top trading cycles (TTC) mechanisms if and only if y is nested and non-degenerate. The same holds if TTC is replaced by serial dictatorship (SD).

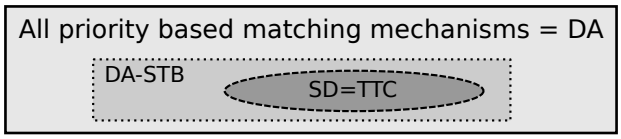


Figure 1 Hierarchy of mechanisms according to Theorem 1 for regular markets. Deferred acceptance (DA) with arbitrary quotas and priorities is flexible enough to implement any feasible budget set probability matrix. DA with single tie-breakers (DA-STB) is strictly more flexible than serial dictatorship (SD) or top trading cycles (TTC), which are equivalent to each other. Intuitively speaking, TTC is less flexible than DA because of its ex-post Pareto efficiency, which substantially limits the possible budget sets.

Part a-i) of the above theorem implies that deferred acceptance (DA) is a very flexible class of mechanisms: for regular markets, optimizing over the space of priority distributions and quotas under DA is equivalent to optimizing over the space of all priority-based allocation mechanisms.[12]

[12] Note that the definition of the set Y^M in (10)-(13) is based on the definition of segments, so the flexibility of DA depends on the policy makers' choice of what observable information the priorities can depend on. If the priorities must be identically distributed across all agents, then the polyhedron Y^M of implementable outcomes has $2^n - (1+n)$ dimensions. If priorities can systematically differ across m segments, then the polyhedron is larger, with $m2^n - (m+n)$ dimensions.

I show in Appendix H.1.1 that the assumption that the market is regular is necessary for the result to hold, but the assumption can be removed if we enrich the class of DA mechanisms to allow for more complex policy levers, such as having a minimum threshold on an agent's priority score for him to be eligible for an item.

Parts a-ii) and b) characterize whether DA with single tie-breakers (DA-STB), serial dictatorship (SD), or top trading cycles (TTC) can be used to implement the desired budget set probabilities. For example, if the market is regular, then any outcome that can be implemented using TTC can also be implemented using DA-STB, but not vice versa. This is because in a regular market M, every $y \in Y^M$ is non-degenerate,[13] and the condition of being nested within segment is strictly weaker than being nested. The proof of Theorem 1 in Appendix H.1 is constructive and specifies in each case the parameters that can be used to implement the desired budget set probabilities.

The equivalence in part b) between TTC and SD is closely related to Lemma 1 of Abdulkadiroğlu and Sönmez (1998), which states that for one-to-one matching markets, the set of outcomes implementable by either of these classes of mechanisms is exactly equal to the set of ex-post Pareto efficient matchings. The condition of nested budget sets is related to Pareto efficiency as non-nested budget sets imply an opportunity for trade: if sets S and S' are non-nested, with $j \in S \backslash S'$ and $j' \in S' \backslash S$, then an agent who is given budget set S but prefers j' may trade with another agent who is given budget set S' but prefers j. However, the proof in Abdulkadiroğlu and Sönmez (1998) is in a one-to-one matching model and does not directly carry over to the setting in this paper.

3.3. Relationship to Models with Discrete Agents

All the results so far pertain to the large market model with a continuum of agents. However, the LP in Section 3.1 can also be used to derive an upper-bound to the welfare achievable in a model with discrete agents for the DA mechanism under arbitrary priorities and quotas. This is because in a discrete model, the agent-proposing DA mechanism still matches each agent to his favorite item within a certain budget set, which is independent of the agent's own preferences. (This follows from strategyproofness and is formally shown in Appendix B.) The difference in the discrete model is that the budget sets may be correlated across agents, whereas in the continuum model they are independent across agents. Regardless, if one defines y_{tS} in the discrete model to be the probability that a type t agent receives the budget set S, and λ_t to be the number of segment t agents, then the matrix y satisfies the constraints of the LP given by (10)-(13), so the optimal LP objective is an upper bound to the utilitarian welfare. The same statement also holds for SD and TTC, as budget sets are also well defined in the discrete versions of these mechanisms.[14]

[13] In a regular market, $P_t(j,S) \geq P_t(j,J) \geq P_t(0, J\backslash\{j\}) - P_t(0,J) > 0$. Hence, if the LHS of (11) is zero, then $y_{tS} = 0$ for every $S \ni j$, so every $y \in Y^M$ is non-degenerate.

[14] In the discrete version of SD, an agent's budget set is the set of items that are not yet depleted. In the discrete version of TTC, the existence and definition of budget sets are shown in Leshno and Lo (2018). The concept of budget

4. Efficient Computation

The LP in Section 3.1 for computing the optimal budget set probabilities may be difficult to solve directly when the number of items n is large, since the number of decision variables is exponential in n. Nevertheless, Appendix C uses a standard column generation argument to show that the LP can be efficiently solved provided that the following assortment planning sub-problem can be efficiently solved. Throughout this section, the subscript t is omitted for simplicity, since the sub-problem always applies to one agent segment at a time.

Definition 6 (Socially Optimal Assortment Planning). Given a parameter $\alpha \geq 0$, a revenue vector $r \in \mathbb{R}^n$, a utility distribution F, which is a probability measure on $\Theta \subseteq \mathbb{R}^{n+1}$, and a constraint set $\Psi \subseteq 2^J$, define the socially optimal assortment planning problem as finding a set $S \in \Psi$ that maximizes the weighted sum of expected utility and expected revenue:

$$\max_{S \in \Psi} \alpha U(S) + \sum_{j=1}^{n} r_j P(j, S), \tag{14}$$

where $U(S) = \mathbb{E}_{u \sim F}\left[\max_{j \in S} u_j\right]$ and $P(j, S) = \mathbb{1}(j \in S)\mathbb{P}_{u \sim F}(u_j = \max_{j' \in S} u_{j'})$ are as in (7) and (8).

When applied to solve the LP in Section 3.1, the quantity $-r_j$ is equal to the shadow price of the capacity constraint (11) for item j, so the objective in (14) can be interpreted as maximizing social welfare, as it trades off the utilities for one segment with the negative externalities imposed on others. When $\alpha = 0$, the above problem is identical to the revenue-maximizing assortment planning problem that has been well-studied in the revenue management literature. When $\alpha > 0$, one also has to consider the utility term $U(S)$, so the problem has a different mathematical structure.

Theorem 2 below shows that even with a large number of items, the socially-optimal assortment planning problem can be efficiently solved under the following families of utility distributions.

- **Multinomial Logit (MNL)**: The utility of an agent i for item $j \in J$ is distributed as

$$u_{ij} = \overline{u}_j + \epsilon_{ij}, \tag{15}$$

  where $\overline{u}_j$ is a constant representing the average utility of the agent segment for item j, and ϵ_{ij} represents the idiosyncratic component of agent preferences and is i.i.d. drawn from a Gumbel distribution, which is also known as the type-I extreme value distribution.
- **2-Level Nested Logit**: This generalizes the MNL utility distribution to allow for positive correlations in an agent's utilities for similar items. The set of items J is partitioned into nests,

set also appears in an earlier analysis in Pycia and Ünver (2017), but there the meaning is different as it refers to the set of items that can be accessed within a particular round of TTC.

$J = \dot{\bigcup}_s J_s$, with similar items in the same nest s. Moreover, the outside option 0 is always in its own nest by itself. The utility of agent i for an item j in nest s is

$$u_{ij} = \overline{u}_j + \delta_{is} + \epsilon_{ij}, \tag{16}$$

where $\overline{u}_j$ and ϵ_{ij} are analogous to the MNL utility distribution, and the additional term δ_{is} is common for all items within the same nest s. The terms δ_{is} and ϵ_{ij} are assumed to be independent, with the sum $\delta_{is} + \epsilon_{ij}$ being Gumbel distributed with scale parameter 1, and ϵ_{ij} being Gumbel distributed with scale parameter $0 < \eta_s \leq 1$.

- **d-level Nested Logit**: This generalizes the 2-level nested logit to allow each nest to be further partitioned into sub-nests and so on, with d being the maximum length of a chain of sets $J \supsetneq J_{s_1} \supsetneq J_{s_2} \supsetneq \cdots \supsetneq J_{s_{d-1}}$. A precise description is given in Appendix H.2.2.
- **Markov Chain Based Choice Model**: A variant of a model proposed by Blanchet et al. (2016) that is designed to be a tractable approximation of a general random utility model. A precise description is given in Appendix H.2.5.

The constraint sets Ψ referred to in Theorem 2 are as follows.

- **Trivial**: The outside option must always be accessible: $\Psi = \{S \subseteq J : 0 \in S\}$.
- **Cardinality**: In addition to the outside option being included, there exists a constant $k \leq n$ and a subset of items $S_0 \subseteq [n]$ such no more than k items from S_0 can be included: $\Psi = \{S \subseteq J : 0 \in S, |S \cap S_0| \leq k\}$.
- **Cardinality within nest**: For the 2-level nested logit utility distribution, the number of items that can be included from nest s is at most k_s: $\Psi = \{S \subseteq J : 0 \in S, |S \cap J_s| \leq k_s \text{ for each nest } s\}$.

Only trivial constraints are needed for the LP in Section 3.1, but cardinality constraints are used to limit school busing costs in Section 6. Given functions $f(n)$ and $g(n)$, recall that $f(n) = O(g(n))$ if there exists a constant $C > 0$ such that $f(n) \leq Cg(n)$ for all sufficiently large n.

THEOREM 2 **(Efficient Computation)**. *For each of the following combinations of utility distribution F and constraint set Ψ, there exists an algorithm to compute an optimal solution to the socially optimal assortment planning problem* (14) *with the following runtime guarantees:*

a) multinomial logit (MNL) utilities and cardinality constraint: $O(n^2 \log n)$;

b) d-level nested logit utilities and trivial constraint: $O(dn \log n)$;

c) 2-level nested logit utilities and cardinality constraint within nest: $O(n^2 \log n)$;

d) the Markov chain based choice model and trivial constraint: solving a linear program with $O(n)$ variables and $O(n)$ constraints.

The guarantees in Theorem 2 are asymptotically the same as the best known for the revenue maximizing special case with $\alpha = 0$, and the theorem shows that the general case with $\alpha \geq 0$

can also be solved in similar numbers of operations. The efficient algorithms and their proofs of correctness are presented in Appendix H.2. For parts a) and b), the algorithms are modifications of those of Rusmevichientong et al. (2010), and Li et al. (2015), which are both based on identifying a small candidate set out of which an optimal assortment can be found. A key lemma shows that the same candidate set for the $\alpha = 0$ case also contains an optimal assortment for any $\alpha \geq 0$, and this is true for any Generalized Extreme Value (GEV) utility distribution, which includes the MNL and nested logit utility distributions as special cases. For part c), the proof combines ideas from parts a) and b). The overall algorithm is similar to that in Xie (2016), but generalizes it from $\alpha = 0$ to $\alpha \geq 0$. For part d), the proof is based on modifying the LP-based solution approach of Feldman and Topaloglu (2017). A consequence of the proof of part b) is as follows:

PROPOSITION 1 **(Setting in which RSD is Optimal)**. *Suppose there is a single segment, whose utilities follow a d-level nested logit distribution. Random serial dictatorship with quotas equal to capacities maximizes utilitarian welfare among all priority-based allocation mechanisms.*

When priorities cannot depend on agent characteristics but only one random tie-breakers, then Proposition 1 implies that RSD is optimal if the distribution of utilities follow a d-level nested-logit structure. When there are multiple agent segments, each of whose utilities follow a d-level nested logit utility distribution, Proposition 1 implies that there exists a priority boost matrix such that DA-STB with quotas equal to capacities maximizes utilitarian welfare among all priority-based allocation mechanisms. The proof of Proposition 1 is in Appendix H.3.

5. More Choices is Not Necessarily Better

An ongoing debate in public school policy is whether to limit students to a neighborhood school, or to provide additional options under a so-called open enrollment policy (Mikulecky 2013). Among school districts with open enrollment, another question is how much choice should be provided. For example, Boston Public Schools (BPS) implemented in 1988 a 3-Zone Plan with about 30 school options for each elementary school student, but there were attempts by the district to cut down the number of choices in 2004 and in 2009, both of which failed due to resistance by the public. After much debate, the city finally changed to an assignment plan in 2014 with about half as many choices as before (Shi 2015). Reasons for limiting choice includes reducing busing costs for the city, increasing community cohesion, and simplifying the assignment system. Before analyzing the Boston context in detail in Section 6, I illustrate here using stylized examples that it may be beneficial to limit choice even if the only consideration is to maximize student welfare.

A potential benefit of open enrollment is that students can be matched to a school that best tailors to their individual preferences. However, when school capacities are limited, a student's

choice to go to a school does not incorporate the externality imposed on the student who is displaced from the school.[15] Therefore, it is possible that implementing open enrollment would attract new applicants who benefit less from the school than the neighborhood students they displace. Proposition 2 illustrates this argument using the following two-school example.

EXAMPLE 1. There are two schools $j \in \{1,2\}$ with equal capacities, $c_1 = c_2$, located respectively in two neighborhoods $t \in \{1,2\}$. There is a unit mass of students in each neighborhood, each of whose utility for the neighborhood school is drawn i.i.d. from a distribution F_0, and for the other school from a distribution F_1. Each student also has an outside option drawn i.i.d. from a distribution H. (One can interpret the outside option assumption as every student being eligible to at least one other school with excess capacity; these other schools may be private options, or under-demanded public schools not explicitly modeled here.) Assume that the demand of each neighborhood for each school is strictly positive, and that the expected gain for being assigned to the neighborhood school is weakly higher than to the other school: if $u_0 \sim F_0$, $u_1 \sim F_1$ and $\alpha \sim H$, then $\mathbb{P}(u_0 \geq \alpha) > 0$, $\mathbb{P}(u_1 \geq \alpha) > 0$, and $\mathbb{E}[u_0 - \alpha | u_0 \geq \alpha] > \mathbb{E}[u_1 - \alpha | u_1 \geq \alpha]$. Moreover, assume that capacities are scarce so that neighborhood applicants alone can fill all seats: $c_1 = c_2 \leq \mathbb{P}(u_0 \geq \alpha)$.

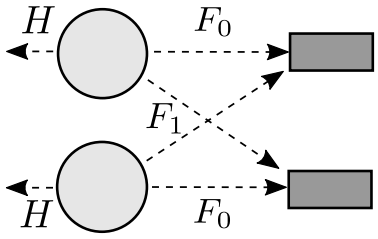


Figure 2 Illustration of Example 1: the circles represent the two neighborhoods and the rectangles the two schools. A student's utility for his neighborhood school is distributed according to F_0 and for the other school according to F_1. Outside options are distributed according to H. School capacities are scarce.

Define the neighborhood assignment plan as offering each student from neighborhood t the budget set $\{0,t\}$ with probability p and $\{0\}$ otherwise, where p is the maximum probability that does not violate the capacity constraint (11). Define the open enrollment plan as running RSD with quotas equal to capacities. By symmetry and by Theorem 1 b), serial dictatorship or top trading cycles under any priority distribution yields the same utilitarian welfare as RSD in this example. Unlike with neighborhood assignment, open enrollment is guaranteed to be ex-post Pareto efficient.

[15] See Appendix H.6 for an one-item example in which it is easy to quantify the externality one student's choice imposes on another student.

PROPOSITION 2 **(Analysis of Example 1)**. *In Example 1, the neighborhood assignment plan maximizes utilitarian welfare among all priority-based allocation mechanisms if and only if the following inequality holds:*

$$\mathbb{E}[u_0 - \alpha | u_0 \geq \alpha] \geq E[\max(u_0, u_1) - \alpha | \max(u_0, u_1) \geq \alpha], \tag{17}$$

where $u_0 \sim F_0$, $u_1 \sim F_1$ *and* $\alpha \sim H$ *and the random variables are independent. On the other hand, the open enrollment plan (RSD) is optimal if and only if the above inequality is reversed.*

Inequality (17) can be interpreted as follows. The left hand side (LHS) is the value of being assigned to one's neighborhood school conditional on preferring it over one's outside option. The right hand side (RHS) is the value of being assigned to one's favorite school conditional on preferring it over one's outside option. Every student by definition weakly prefers his favorite school over his neighborhood school, but the RHS also includes students with utilities $u_1 > \alpha > u_0$, who would settle for their outside option under a neighborhood assignment plan but would compete for space under the open enrollment plan. If the proportion of such students is large and their expected gain $\mathbb{E}[u_1 - \alpha | u_1 > \alpha > u_0]$ is small compared to the LHS, then the neighborhood assignment plan may achieve higher aggregate welfare.

For a concrete example in which neighborhood assignment is optimal, consider the following parameters: the outside option is normalized to zero, $\alpha = 0$. Every student's utility for his own neighborhood school follows a two point distribution, in which $u_0 = H$ with probability p and $u_0 = L$ otherwise; every student's utility for the other school is $u_1 = \epsilon$. The values are such that $L < 0 < \epsilon < H$. As a result, the LHS of (17) is H and the RHS is $pH + (1-p)\epsilon < H$. In stylized language, students either love or hate their neighborhood school, and those who hate their neighborhood school would be happy with their outside option. However, everyone marginally prefers the other school to his outside option, so those who hate their neighborhood school would still apply to the other school if given the opportunity. Hence, it is optimal for each school to save its seats for the neighborhood applicants, as these students benefit the most from being assigned there.

The previous example relies on students having higher expected gains from being assigned to their neighborhood school. Proposition 3 shows that this assumption is not necessary, as a similar result holds even with ex-ante symmetric utilities if capacities remain scarce and the outside option distribution is sufficiently left skewed.

EXAMPLE 2. There are n schools located in n neighborhoods respectively. School capacities are possibly unequal but the supply-demand ratio is uniform across neighborhoods: $c_j/\lambda_j = c_{j'}/\lambda_{j'}$ for all $j, j' \in [n]$. The utility of a student for each school is drawn i.i.d. from a continuous distribution F. Outside options are drawn i.i.d. from a continuous distribution H, which is assumed to have a

weakly larger upper support: if $F(x)$ and $H(x)$ are the CDFs, then $F(x) < 1$ implies that $H(x) < 1$.[16] As in Example 1, there is sufficient demand from within the neighborhood to fill school capacity: $c_1/\lambda_1 \leq \mathbb{P}_{u\sim F,\alpha\sim H}(u \geq \alpha)$.

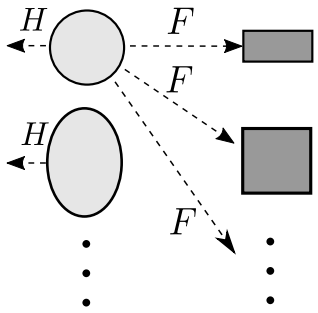


Figure 3 Illustration of Example 2: the circles represent the n neighborhoods and the rectangles the n schools. School sizes may be unequal, but are assumed to be proportional to respective neighborhood populations. Student preferences for the n schools are i.i.d. drawn from F, which is the same across schools and neighborhoods. Outside options are distributed according to H, and capacities are scarce as in Example 1.

Definition 7. A distribution with CDF $H(x)$ and density $h(x)$ has a heavy left-tail if the left hazard rate $h(x)/H(x)$ is weakly increasing in $x \in (\underline{x}, \overline{x})$, where $\underline{x} := \inf\{x : H(x) > 0\}$ and $\overline{x} := \sup\{x : H(x) < 1\}$. (It is possible that $\underline{x} = -\infty$ or $\overline{x} = \infty$.) Conversely, the distribution has a light left-tail if $h(x)/H(x)$ is weakly decreasing in $x \in (\underline{x}, \overline{x})$.[17]

Distributions with a heavy left-tail are left skewed, such that the density $h(x)$ increases at a rate higher than the CDF $H(x)$. Examples include the negated Pareto distribution, and the negated Weibull distribution with shape parameter less than or equal to one. Examples of distributions with light left-tails include the uniform, the normal, or the Gumbel distributions. The negated exponential distribution is in both classes and represents the borderline between the two.

Proposition 3 **(Analysis of Example 2)**. *In Example 2, the neighborhood assignment plan maximizes utilitarian welfare among all priority-based allocation mechanisms if the outside option distribution H has a heavy left-tail. The open enrollment plan (RSD) maximizes utilitarian welfare if H has a light left-tail.*

The intuition behind Proposition 3 is as follows. When the school district changes from a neighborhood assignment plan to an open enrollment plan, the students who are assigned to one of the schools benefit from the better matches, while more students are assigned to an outside option not

[16] The proof of Proposition 3 only requires the assumption that H has a weakly larger upper support when H has a heavy left-tail, but not when H has a light left-tail.

[17] An equivalent characterization of a heavy left-tail is that the CDF is log-convex: $\log(H(x))$ is convex in $x \in (\underline{x}, \overline{x})$. Similarly, a distribution has a light left-tail if and only if its CDF is log-concave: $\log(H(x))$ is concave in $x \in (\underline{x}, \overline{x})$.

because they like it but because they have no other alternative. If the outside option distribution has a heavy left-tail and its upper support is larger than a student's utility for any school, then most students are okay with their outside option, but a few are extremely unhappy about it. Under these assumptions, it is more important to offer as many students as possible at least one alternative to their outside option, rather than to provide a better match for the assigned students, so the neighborhood assignment plan is better. On the other hand, when outside options are light left-tailed, the benefit of better matches for the assigned students outweighs the possible harm of leaving more students with only their outside option, so open enrollment is better.[18]

The proofs of Propositions 2 and 3 are in Appendices H.4 and H.5, and are based on analytically solving the LP in Section 3.1 to reveal the structure of the optimal budget sets.

5.1. Comparison of DA and TTC under Neighborhood or Sibling Priorities

The results in the previous section can also be interpreted as stylized comparisons of the DA and TTC mechanisms. This is because the neighborhood assignment plan in Examples 1 and 2 can be interpreted as implementing the DA mechanism while prioritizing students from the neighborhood over non-neighborhood students and breaking remaining ties randomly. Similarly, the open enrollment plan can be interpreted as implementing the TTC mechanism with the same priorities. The notion of neighborhood can also be replaced by any other affinity between students and schools, such as having a sibling currently assigned at the school.

As documented in Abdulkadiroğlu et al. (2006), policy makers in Boston were recommended in 2005 to adopt TTC due to its ex-post Pareto efficiency, but decided to choose DA, which is not ex-post Pareto efficient under neighborhood priorities. Similarly, school officials in New Orleans chose to migrate to DA after using TTC for one year (Abdulkadiroğlu et al. 2017). Rationales for choosing DA over TTC that are provided in the literature include the complexity of explaining TTC and not eliminating justified envy (Abdulkadiroğlu et al. 2006, Pathak 2017, Abdulkadiroğlu et al. 2017). Propositions 2 and 3 suggest an additional rationale based on maximizing aggregate welfare: Since students who are given neighborhood or sibling priorities to a school on average have high values for that school, it may be better for aggregate welfare to respect these priorities.

[18] The idea that unequal outside options can disrupt the welfare properties of a seemingly efficient allocation mechanism appears also in Akbarpour and Van Dijk (2018) and Calsamiglia et al. (2019), who argue that moving from an ordinal allocation mechanism to a cardinal may harm agents with poor outside options, while prior analysis had suggested that such a move can be Pareto improving if outside options are equal (Abdulkadiroğlu et al. 2011). However, the driving force behind Proposition 3 is different because the mechanisms considered in this paper are all ordinal, as an agent's optimal action depends only on his relative rankings of items but not on the intensities of his preferences.

6. Empirical Application: School Choice in Boston

I now apply the methodologies developed in this paper to real data from Boston Public Schools (BPS) to compute an optimal priority-based allocation mechanism for elementary school assignment that satisfies certain institutional constraints. Since 2006, BPS has been using the DA mechanism to assign students, whose eligibility to schools and distribution of priorities depend on their home locations. In 2013, policy makers in Boston decided to migrate from a 3-Zone assignment plan to a Home-Based plan, being guided by a simulation analysis described in Pathak and Shi (2013) and Shi (2015), which estimated a MNL utility distribution of student preferences using past data and used it to compare various assignment plans. (See Appendix D for details of the 3-Zone and Home-Based plans.) In this context, an assignment plan specifies a mapping from each neighborhood t to a certain choice menu $J_t \subseteq J$, which is the set of eligible schools that students from that location can include in the preference rankings they submit to the DA mechanism. Moreover, an assignment plan specifies the multivariate distribution G_t of priorities for students of each neighborhood t. Another important piece of background information is that every neighborhood has a certain walk-zone, which includes every school within a 1-mile radius. The city is required to provide school busing only for students assigned to a school outside of their walk-zone.

Table 1 compares various assignment plans based on a simplified version of the simulation model used in the Boston reform, which is also used in Ashlagi and Shi (2015). As seen in the first two columns, the Home-Based plan decreases the amount of school busing needed by more than a factor of two compared to the 3-Zone plan, as measured by all three metrics shown in rows 3 through 5 of Table 1. However, the Home-Based plan also reduces the expected utilities of students as implied by the MNL utility distribution. Despite this trade-off, policy makers decided to adopt the Home-Based plan, as it provided other benefits as documented in Shi (2015).

After the policy change, Ashlagi and Shi (2015) proposed an optimized plan that uses the same average miles of busing as in the Home-Based plan, while achieving better expected utilities for students and better predictability[19] (see the third column of Table 1). However, their methodology is unable to bound other measures of busing costs, such as the average area from which a school needs to bus students (row 4 of Table 1), which is a metric that was salient during the Boston policy reform (Pathak and Shi 2013, Shi 2015). In contrast, the optimized plan in this paper (see the rightmost column of Table 1) dominates both the 3-Zone plan and the Home-Based plan in all of the metrics shown, which makes it more viable for implementation.

[19] Predictability of the assignment is measured by the chance a student is assigned to one of his top submitted choices among the schools in his choice menu. Note that a neighborhood assignment plan has 100% predictability according to this metric as the choice menu is a singleton.

	3-Zone	Home-Based	Ashlagi & Shi (2015)	**This paper**
Descriptive statistics				
(1) Av. # of choices	29.24	14.77	21.12	14.61
(2) Av. miles to assigned school	1.79	1.30	1.32	1.29
Busing requirement				
(3) Miles bused per student	1.26	0.64	0.63	0.63
(4) Av. bus coverage area	22.63	8.51	13.47	7.77
(5) Av. # of busing choices	22.29	8.17	14.64	8.15
Expected utilities of neighborhoods				
(6) Weighted average	7.20	6.95	7.49	7.45
(7) 10th percentile	6.41	6.10	7.31	7.27
(8) Lowest	4.95	4.58	7.03	6.96
% getting top choices in menu				
(9) Top 1	61.2%	64.1%	78.7%	78.7%
(10) Top 3	80.4%	84.9%	93.4%	92.8%

Table 1 Comparison of assignment plans for Boston, using the MNL utility distribution and simulation framework described in Appendix D. Each column corresponds to an assignment plan, which specifies the set of school options each student can rank in the DA mechanism as well as his priority distribution. The results are for Kindergarten-2, which is the main entry grade to elementary schools. The rows correspond to metrics of interest for policy makers during the 2013 Boston student assignment reform. The metrics are respectively (1) the average number of schools in the choice menu, which is the set of eligible schools each student can rank; (2) the average distance from a student's home to his/her assigned school; (3) the total distance from home to school for students assigned to a school outside of the 1-mile walk-zone, divided by the total number of students; (4) the average area in square miles from which a school needs to pick up children by school bus; (5) the average number of schools within the choice menu for which students are eligible for busing (i.e. outside of the 1-mile walk-zone); (6) the average expected utility of all students; (7) the expected utility of a neighborhood in the bottom 10th percentile; (8) the minimum expected utility of any neighborhood; (9) the proportion of students assigned to their top choice within their choice menu; and (10) the proportion of students assigned to one of their top three choices within their choice menu. All estimates are based on the average of 100,000 independent simulations.

Despite the use of real data and policy relevant metrics, the analysis in this section is not intended to argue for a policy change, but only to illustrate how to handle the difficult non-linear constraints that may arise in practice via appropriate approximations. Influencing policy would require extensive deliberations across stakeholders to decide on the metrics and parameters, as well as extensive robustness tests. The optimized choice sets and priorities also need to be simplified so that the reasoning behind them is understandable to parents in Boston, who might not trust a black box system. Section 7 discusses how these remaining issues can be addressed.

6.1. Optimization Formulation

The optimized plan in the rightmost column of Table 1 is based on finding an exact solution to the following LP, which at a high level is to find a budget set probability matrix that yields high expected utilities for the students, subject to staying within the busing requirements of the Home-

Based plan. As in Ashlagi and Shi (2015), the objective is a weighted sum of the utilitarian welfare and the expected utility of the worst-off neighborhood, so that students on average are matched to schools they like and no neighborhood is very badly off. The constraints (25), (26) and (27) bound the amount of busing by the city according to the three metrics in rows 3 to 5 of Table 1. Constraint (25) corresponds to the average miles of busing per student, and is also used in Ashlagi and Shi (2015). Constraints (26) and (27) are linearizations of the constraints corresponding to rows 4 and 5 of Table 1, and they are what differentiates the optimization here from the one in Ashlagi and Shi (2015). These two constraints can only be tractably included due to the efficient algorithm developed in Theorem 2 for solving the socially optimal assortment planning problem under MNL utilities and cardinality constraints. The input data are as follows:

- m: the number of neighborhoods; $m = 868$ in the dataset.
- n: the number of schools; $n = 77$ in the dataset.
- λ_t: the expected number of applicants from neighborhood $t \in [m]$. Define $\Lambda := \sum_{t\in[m]} \lambda_t$.
- $U_t(S), P_t(j,S)$: expected utilities and choice probabilities for a student of neighborhood t given budget set S, as defined in (7) and (8). They are based on a MNL utility distribution estimated from past choice data. Their parameters and closed form expressions are given in Appendix D.2.
- j_t: the default school for neighborhood t. (See Appendix D.3 for more details.)
- a_t: the area of neighborhood t in square miles.
- c_j: the number of seats of school j available to students for whom j is not their default school.
- d_{tj}: the Google Map walking distance from the centroid of neighborhood t to school j.
- S_t^{walk}: the set of schools within the one-mile walk-zone of the neighborhood t.
- A, B_1, B_2, B_3: Tuning parameters of the optimization. $A \in [0,1]$ corresponds to the relative weight in the objective function between the average utility of all neighborhoods and the expected utility of the worst-off neighborhood. B_1, B_2 and B_3 correspond to various forms of busing budget, and are the right hand sides of constraints (25), (26) and (27).

In the formulation below, all of the summations of t are over $[m]$, of j are over $[n]$ and of S are over the power set $2^{[n]}$.

$$\text{Maximize}_y \quad A\left(\frac{1}{\Lambda}\sum_t \lambda_t w_t\right) + (1-A)\underline{w} \tag{18}$$

$$\text{s.t.} \qquad y_{tS} \geq 0$$

$$\text{(Neighborhood utilities)} \qquad \sum_S U_t(S) y_{tS} = w_t \quad \text{for each neighborhood } t. \tag{19}$$

$$\text{(Assignment probabilities)} \qquad \sum_S P_t(j,S) y_{tS} = p_{tj} \quad \text{for each } t \text{ and school } j. \tag{20}$$

$$\text{(Default school)} \qquad y_{tS} = 0 \quad \text{if } j_t \notin S \tag{21}$$

$$\text{(Valid probabilities)} \qquad \sum_S y_{tS} = 1 \quad \text{for each } t. \tag{22}$$

$$\text{(Lower bound on utilities)} \qquad \underline{w} \le w_t \quad \text{for each } t. \tag{23}$$

$$\text{(School capacity)} \qquad \sum_t \lambda_t \mathbb{1}(j \neq j_t) p_{tj} \le c_j \quad \text{for each school } j. \tag{24}$$

$$\text{(Miles bused)} \qquad \frac{1}{\Lambda} \sum_{t,j} \lambda_t p_{tj} d_{tj} \mathbb{1}(j \notin S_t^{walk}) \le B_1 \tag{25}$$

$$\text{(Bus coverage area)} \qquad \frac{1}{n} \sum_{t,S} a_t |S \backslash S_t^{walk}| y_{tS} \le B_2 \tag{26}$$

$$\text{(\# of busing choices)} \qquad \frac{1}{\Lambda} \sum_{t,S} \lambda_t |S \backslash S_t^{walk}| y_{tS} \le B_3 \tag{27}$$

Note that the school capacity in (24) only applies to students not assigned to their default school. This can be interpreted as the school district subsequently adding enough capacity to the default schools so as to take in all unassigned students, which it is required to do by mandatory schooling laws. Moreover, the constraints related to busing (25)-(27) only apply to schools outside of the walk-zone, as only students assigned to a school outside of their walk-zone are bused.

Constraints (26) and (27) are linearized versions of the following non-linear constraints, which exactly correspond to rows 4 and 5 of Table 1 but are difficult to optimize directly due to their non-linear dependence on y.

$$\text{Exact constraint for bus coverage area:} \qquad \frac{1}{n} \sum_t a_t |J_t(y) \backslash S_t^{walk}| \le B_2, \tag{28}$$

$$\text{Exact constraint for \# of busing choices:} \qquad \frac{1}{\Lambda} \sum_t \lambda_t |J_t(y) \backslash S_t^{walk}| \le B_3, \tag{29}$$

$$\text{where} \qquad J_t(y) := \bigcup \{S : y_{tS} > 0\}. \tag{30}$$

$J_t(y)$ is the set of schools to which a student from neighborhood t has a non-zero chance of being assigned under budget set probabilities y. Therefore, the left hand side (LHS) of (28) is equal to the average across schools of the area each school needs to cover to pick up students. Similarly, the LHS of (29) is the average across students of the number of schools outside of the walk-zone that students can potentially be assigned to. If y_{tS} were binary, then due to (22), constraint (26) would be equivalent to (28), and (27) would be equivalent to (29). For continuous y_{tS}, the LHS of (26) is a lower bound on the LHS of (28), and the LHS of (27) is a lower bound on the LHS of (29), so constraints (26) and (27) in the LP are linear relaxations of the exact constraints (28) and (29).

6.2. Solving the Optimization by Column Generation

The LP in (18)-(27) can be efficiently solved as follows. For each neighborhood t, maintain a set of assortments $\mathcal{A}_t$, which is a small subset of the power set of $2^{[n]}$. Define $Opt(\mathcal{A})$ to be the LP in

which all summations of S are over $\mathcal{A}_t$ instead of $2^{[n]}$, so that only decision variables y_{tS} with $S \in \mathcal{A}_t$ need to be included. Given an optimal solution to $Opt(\mathcal{A})$, let the shadow prices of constraints (23)-(27) be ν_t, γ_j, ξ_1, ξ_2 and ξ_3 respectively. For each neighborhood t, define the column generation sub-problem given (ν, γ, ξ) as the following assortment optimization problem:

$$\max_{S \in \Psi_t} \left\{ \alpha U_t(S) + \left(\sum_{j \in [n]} r_j P_t(j, S) \right) - \zeta |S \backslash S_t^{walk}| \right\}, \tag{31}$$

$$\text{where } \Psi_t := \{S \subseteq [n] : j_t \in S\}, \tag{32}$$

$$\alpha := A \frac{\lambda_t}{\Lambda} + \nu_t, \tag{33}$$

$$r_j := -\lambda_t \left[\mathbb{1}(j \neq j_t)\gamma_j + \frac{d_{tj}}{n} \mathbb{1}(j \in S_t^{walk})\xi_1 \right], \tag{34}$$

$$\zeta := \frac{a_t}{n}\xi_2 + \frac{\lambda_t}{\Lambda}\xi_3. \tag{35}$$

The optimization (31) can be solved using the algorithm in Appendix H.2.1 for socially optimal assortment planning with MNL utilities and cardinality constraints by constraining the cardinality of $|S \backslash S_t^{walk}|$ to be no more than k, penalizing the objective by $-\zeta k$, and searching through all $k \in \{0, 1, \cdots, n - |S_t^{walk}|\}$. While the algorithm in Appendix H.2.1 runs in $O(n^2 \log n)$ time for a given cardinality k, one can modify it so that it optimizes over all possible k's simultaneously, so that the whole column generation sub-problem (31) can be solved in $O(n^2 \log n)$ time. The optimized budget sets have an intuitive structure that is explained in Appendix H.7.

The full algorithm for solving the LP is as follows.

1. Initialize $\mathcal{A}$ arbitrarily so that $\mathcal{A}_t \subseteq \Psi_t$ and $Opt(\mathcal{A})$ is feasible. For the Boston dataset, I initialize $\mathcal{A}_t$ to be all assortments of the form $\{j_t, j\}$, where $j \in [n]$.
2. Solve $Opt(\mathcal{A})$ and let the shadow price of (22) be ϕ_t. Solve the column generation sub-problem (31) for each neighborhood t and let the optimal objective value be ϕ'_t. This is guaranteed to be at least ϕ_t, because by duality, ϕ_t is equal to the optimal objective value of the sub-problem (31) with the constraint set Ψ_t replaced by $\mathcal{A}_t \subseteq \Psi_t$.
3. If $\phi'_t = \phi_t$ for all t, then terminate and return an optimal solution to $Opt(\mathcal{A})$. Otherwise, for each neighborhood t such that $\phi'_t > \phi_t$, add an optimal solution S^* to the sub-problem (31) to the set $\mathcal{A}_t$, and go back to step 2.

The correctness of the above algorithm follows from the discussion in Appendix C. For the Boston dataset, it yields an exact optimal solution y^* to the LP in (18)-(27) in about 7 minutes, when implemented using Python 3.7 and Gurobi 8.1 and run using one core of a Intel 2.70GHz CPU.[20]

[20] The majority of the runtime is spent on solving the sub-problem (31) rather than the LP $Opt(\mathcal{A})$.

6.3. Constructing an Assignment Plan Satisfying Institutional Constraints

Given an optimal solution y^* to the LP in Section 6.1, I construct an assignment plan using the following heuristics. (Recall that an assignment plan in the Boston context specifies for each neighborhood a choice menu of eligible schools, and a distribution of priorities.)

1. Re-solve the LP in Section 6.1 with the following modification: remove constraints (26) and (27), and alter every summation in S to be over the set $J_t(y^*)$, which is defined in (30). Let the optimal solution be y^{**}. Since y^* is a feasible solution to the modified LP, y^{**} is guaranteed to achieve a weakly higher objective value. By construction, it does not increase the actual bus coverage area or the number of busing choices as defined by the LHS of the exact constraints (28) and (29). Moreover, Proposition 1 implies that y^{**} is guaranteed to be nested within segment, which implies by Theorem 1 that it can be implemented using DA-STB. This is convenient as DA-STB is the version of DA implemented in Boston.
2. Define the choice menu of neighborhood t to be $J_t(y^{**})$, which is the set of schools that a student from neighborhood t has a non-zero probability of being assigned to under y^{**}.
3. Define the priority distribution by constructing the priority boost matrix b corresponding to y^{**} following the proof of Theorem 1, so that the priority boost of neighborhood t for school j is equal to $b_{tj} := \sum_{S \ni j} y^{**}_{tS}$. For a student i from neighborhood t, his priority score for school j is defined as $\pi_{ij} := b_{tj} + \delta_i$, where $\delta_i \sim \text{Uniform}(0,1)$ is the random tie-breaker for student i, and higher priorities scores are preferred.

For the final optimized plan, which corresponds to the last column of Table 1, the parameter A (defined in Section 6.1) is set to 0.5, as this simultaneously yields near optimal utilitarian welfare and max-min welfare on the dataset. The parameters (B_1, B_2, B_3) are tuned so that the assignment plan constructed above, when evaluated under the discrete simulation model, yields a lower busing requirement than the Home-Based Plan as measured by the metrics in rows 3 through 5 of Table 1.[21]

Appendix E expands on the arguments described in Section 3.3 and uses the LP in Section 6.1 to derive a provable upper bound to the maximum objective value that can be achieved by any assignment plan under the discrete simulation model. With $A = 0.5$, the LP objective function is the unweighted average of utilitarian welfare (average expected utility across students) and max-min welfare (worst expected utility of any neighborhood). With respect to this metric, the optimized plan achieves 82% of the possible improvement over the Home-Based plan, subject to not using more busing resources. This estimate is conservative as the upper bound may not be achievable.

[21] The final parameters used are $B_1 = 0.6$, $B_2 = 8.5$ and $B_3 = 6.2$. The bus coverage constraint (26) is not tight at optimality, as the constraint (27) is the bottleneck. This explains why in Table 1, the average bus coverage area of the optimized plan is only 7.77, which is significantly below the budget of 8.51 from the Home-Based plan.

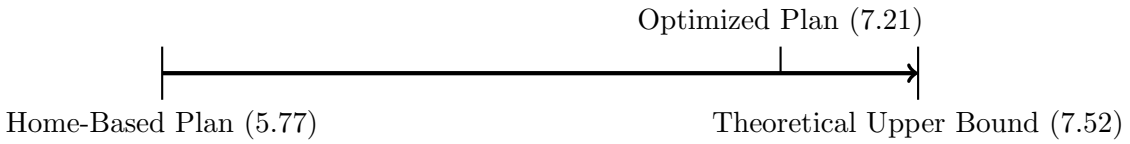


Figure 4 Comparison in the finite market stochastic model of the objective value achieved by the optimized plan from Section 6.3 with that achieved by the Home-Based plan and of the theoretical upper bound. The objective value shown above corresponds to the average of the utilitarian welfare (row 6 of Table 1) and the lowest expected utility of any neighborhood (row 8 of Table 1). The performance of the Home-Based plan and of the optimized plan are based on 100,000 independent simulations. See Appendix E for explanation of the theoretical upper bound.

7. Discussion

This paper suggests the following approach to design a priority-based allocation mechanism for a given application:

1. Classify agents into segments based on institutional constraints of what the priority system can depend on, and estimate the size and utility distribution of each segment.
2. Ask policy makers about their desired objectives and formulate a mathematical program for the optimal budget set probability matrix, as in Sections 3.1 and 6.1.
3. Solve the mathematical program and translate the solution into a concrete priority-based allocation mechanism of the desired form, as in the proof of Theorem 1 and in Section 6.3.
4. Estimate the performance of the mechanism in a discrete simulation model that takes into account the stochastic nature of demand and other institutional considerations, and bound the optimality gap as in Figure 4.

A potential concern is that the estimates of the segment sizes and utility distributions in Step 1 may be subject to estimation error, in which case the optimization would be based on incorrect inputs. For example, in the context of school choice in Boston, Pathak and Shi (2019) documents a substantial change in the demand estimates from 2013 to 2014, which corresponds to the first year of implementation of the Home-Based plan. Appendix F simulates the optimized plan from Section 6.3 using updated population and utility distributions, and shows that the plan's performance is relatively robust despite having sizable errors in its distributional assumptions. It would be interesting future work to explicitly account for the possibility of errors in demand estimates, perhaps using techniques from robust optimization.

Another potential concern is that the optimized priority distributions may be too complex for practical implementation. While the framework in this paper allows policy makers to designate which observable characteristics can inform priorities via suitable definition of segments, the optimal mapping between segments and priorities can be quite complex: in the optimized plan in Section 6.3, there is a priority boost b_{tj} for each neighborhood-school pair, so there are many more

degrees of freedom than in the implemented system. One approach on simplifying the priority system is to modify the mathematical program to force its solution to exhibit additional structure, such as by restricting which budget sets can be used. An alternative approach is to start with the optimal but complex priority system, and then approximate the optimal budget set probabilities using simpler priorities, or using other policy levers such as agent-specific quotas or set asides. Since simplicity and interpretability depend on the context, this may need to be done on a case by case basis using ad-hoc techniques. The benefit of being able to solve for the optimal system is that one can use it as a benchmark to quantify the loss of optimality incurred by each type of simplification, so as to rigorously navigate the trade-off between optimality and simplicity.

References

Abdulkadiroğlu A, Che Y, Yasuda Y (2015) Expanding "choice" in school choice. *American Economic Journal: Microeconomics* 7(1):1–42.

Abdulkadiroğlu A, Che Y, Yasuda Y (2011) Resolving conflicting preferences in school choice: The Boston mechanism reconsidered. *The American Economic Review* 101(1):399–410.

Abdulkadiroğlu A, Che YK, Pathak PA, Roth AE, Tercieux O (2017) Minimizing justified envy in school choice: The design of New Orleans' OneApp. Working Paper 23265, National Bureau of Economic Research.

Abdulkadiroğlu A, Pathak PA, Roth AE (2009) Strategy-proofness versus efficiency in matching with indifferences: Redesigning the NYC high school match. *American Economic Review* 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 Department of Economics.

Abdulkadiroğlu A, Sönmez T (1998) Random serial dictatorship and the core from random endowments in house allocation problems. *Econometrica* 66(3):689.

Abdulkadiroğlu A, Sönmez T (2003) School choice: A mechanism design approach. *American Economic Review* 93(3):729–747.

Abdulkadiroğlu A, Sönmez T (2013) Matching markets: Theory and practice. Acemoğlu D, 2010 Econometric Society / World Congress 10 S, Arellano M, Dekel E, eds., *Advances in Economics and Econometrics: Tenth World Congress: Economic Theory*, 3–47, Econometric Society Monographs (Cambridge University Press), ISBN 9781107016040.

Agarwal N, Somaini P (2019) Revealed preference analysis of school choice models. *Annual Reviews of Economics* Forthcoming.

Akbarpour M, Van Dijk W (2018) School choice with unequal outside options. Working paper, Stanford University.

Arnosti N (2016) Centralized clearinghouse design: A quantity-quality tradeoff. Technical report, Stanford University.

Ashlagi I, Nikzad A (2017) What matters in tie-breaking rules? how competition guides design. Technical report, Stanford University.

Ashlagi I, Nikzad A, Romm A (2019) Assigning more students to their top choices: A comparison of tie-breaking rules. *Games and Economic Behavior* 115:167 – 187.

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

Ashlagi I, Shi P (2015) Optimal allocation without money: an engineering approach. *Management Science* 64(4):1078–1097.

Azevedo EM, Leshno JD (2016) A supply and demand framework for two-sided matching markets. *Journal of Political Economy* 124(5):1235–1268.

Bertsekas D (2015) *Convex Optimization Algorithms* (Athena Scientific).

Bertsimas D, Tsitsiklis J (1997) *Introduction to Linear Optimization* (Athena Scientific), 1st edition.

Biró P, Kiselgof S (2015) College admissions with stable score-limits. *Central European Journal of Operations Research* 23(4):727–741.

Blanchet J, Gallego G, Goyal V (2016) A Markov chain approximation to choice modeling. *Operations Research* 64(4):886–905.

Bodoh-Creed A (2020) Optimizing for distributional goals in school choice problems. *Management Science* Forthcoming.

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

Bront JJM, Méndez-Díaz I, Vulcano G (2009) A column generation algorithm for choice-based network revenue management. *Operations Research* 57(3):769–784.

Calsamiglia C, Martinez-Mora F, Miralles A (2019) Cardinal assignment mechanisms: Money matters more than it should. Working paper, Barcelona GSE.

Cardell NS (1997) Variance components structures for the extreme-value and logistic distributions with application to models of heterogeneity. *Econometric Theory* 13(2):185–213.

Dubins LE, Freedman DA (1981) Machiavelli and the Gale-Shapley algorithm. *American mathematical monthly* 485–494.

Dur U, Kominers SD, Pathak PA, Sönmez T (2018) Reserve design: Unintended consequences and the demise of Bostons walk zones. *Journal of Political Economy* 126(6):2457–2479.

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

Feigenbaum I, Kanoria Y, Lo I, Sethuraman J (2020) Dynamic matching in school choice: Efficient seat reassignment after late cancellations. *Operations Research* Forthcoming.

Feldman JB, Topaloglu H (2017) Revenue management under the Markov chain choice model. *Operations Research* 65(5).

Gallego G, Iyengar G, Phillips R, Dubey A (2004) Managing flexible products on a network. Technical report, Columbia University.

Gallego G, Topaloglu H (2014) Constrained assortment optimization for the nested logit model. *Management Science* 60(10):2583–2601.

Hafalir IE, Yenmez MB, Yildirim MA (2013) Effective affirmative action in school choice. *Theoretical Economics* 8(2):325–363, ISSN 1555-7561.

Hausman JA, Ruud PA (1987) Specifying and testing econometric models for rank-ordered data. *Journal of econometrics* 34(1):83–104.

Israni AK, Salkowski N, Gustafson S, Snyder JJ, Friedewald JJ, Formica RN, Wang X, Shteyn E, Cherikh W, Stewart D, et al. (2014) New national allocation policy for deceased donor kidneys in the United States and possible effect on patient outcomes. *Journal of the American Society of Nephrology* 25(8):1842–1848.

Jeong Bh (2018) School choice in context: Can open enrollment cure segregation? Working paper, University of Melbourne.

Kojima F (2012) School choice: Impossibilities for affirmative action. *Games and Economic Behavior* 75(2):685–693.

Kominers SD, Sönmez T (2016) Matching with slot-specific priorities: theory. *Theoretical Economics* 11(2):683–710.

Leshno JD, Lo I (2018) The cutoff structure of top trading cycles in school choice. Working paper, University of Chicago.

Li G, Rusmevichientong P, Topaloglu H (2015) The d-level nested logit model: Assortment and price optimization problems. *Operations Research* 62(2):325–342.

Liu Q, Pycia M (2016) Ordinal efficiency, fairness, and incentives in large markets. Available at SSRN: `https://ssrn.com/abstract=1872713`.

Liu Q, van Ryzin G (2008) On the choice-based linear programming model for network revenue management. *MSOM* 10(2):288–310.

McFadden D (1978) Modeling the choice of residential location. *Transportation Research Record* 673.

McFadden D, Train K, et al. (2000) Mixed MNL models for discrete response. *Journal of applied Econometrics* 15(5):447–470.

Mikulecky MT (2013) Open enrollment is on the menu–but can you order it. Denver, CO: Education Commission of the States.

Pathak P, Shi P (2019) How well do structural demand models work? Counterfactual predictions in school choice. *Journal of Econometrics* Forthcoming.

Pathak PA (2017) What really matters in designing school choice mechanisms. Honor B, Pakes A, Piazzesi M, Samuelson L, eds., *Advances in Economics and Econometrics: Eleventh World Congress*, volume 1 of *Econometric Society Monographs*, 176214 (Cambridge University Press).

Pathak PA, Shi P (2013) Simulating alternative school choice options in Boston. Technical report, MIT School Effectiveness and Inequality Initiative.

Pycia M (2019) Evaluating with statistics: Which outcome measures differentiate among matching mechanisms? Working paper, University of Zurich.

Pycia M, Ünver MU (2017) Incentive compatible allocation and exchange of discrete resources. *Theoretical Economics* 12(1):287–329.

Roth AE (1982) The economics of matching: Stability and incentives. *Mathematics of operations research* 7(4):617–628.

Rusmevichientong P, Shen ZJM, Shmoys DB (2010) Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations Research* 58(6):1666–1680.

Rusmevichientong P, Shmoys D, Tong C, Topaloglu H (2014) Assortment optimization under the multinomial logit model with random choice parameters. *Production and Operations Management* 23(11):2023–2039.

Shi P (2015) Guiding school-choice reform through novel applications of operations research. *Interfaces* 45(2):117–132.

Su X, Zenios SA (2006) Recipient choice can address the efficiency-equity trade-off in kidney transplantation: A mechanism design model. *Management Science* 52(11):1647–1660.

Thompson D, Waisanen L, Wolfe R, Merion RM, McCullough K, Rodgers A (2004) Simulating the allocation of organs for transplantation. *Health Care Management Science* 7(4):331–338.

Von Hohenbalken B (1977) Simplicial decomposition in nonlinear programming algorithms. *Mathematical Programming* 13(1):49–68, ISSN 0025-5610.

Vulkan N, Roth AE, Neeman Z (2013) *The handbook of market design* (OUP Oxford).

Xie T (2016) A combinatorial algorithm for constrained assortment optimization under nested logit model. ArXiv preprint arXiv:1603.09014.

Table of Appendices

Appendix A: Precise Definitions of Priority Cutoffs

A.1. Cutoff Vector for Serial Dictatorship (SD)

Algorithm 1 below defines the cutoff vector $z^{SD(M,G,q)}$ in serial dictatorship. It is an adaptation of the simultaneous eating algorithm in Bogomolnaia and Moulin (2001). Let g_t be the density of the priority distribution for segment t, and let $P_t(j,S)$ be the choice probability of segment t agents for item j given budget set S. Given non-negative vector $y \in \mathbb{R}^n_+$, a subset $S \subseteq J$, and a scalar $\tau \in [0,1]$, define

$$z_j(y,S,\tau) := \sup\left\{ z : \int_z^\tau \sum_{t\in[m]} \lambda_t g_t(\pi) P_t(j,S)\, d\pi = y_j \right\}. \tag{A.1}$$

This is the lowest priority needed to access item j if we assume that after agents with priorities higher than τ have already picked, the set of available items is S and the remaining quota for item j is y_j. If item j is never depleted, then the set is empty and the supremum is $-\infty$.

Algorithm 1: Computing cutoff vector $z^{SD(M,G,q)}$ in serial dictatorship.

Initialize $k \leftarrow 1$, $S^1 \leftarrow [n]\backslash\{j : q_j = 0\}$, $z^0 \leftarrow 1$, $y^0 \leftarrow q$; for each $j \in [n]$, if $q_j = 0$ then $z^*_j \leftarrow 1$, otherwise $z^*_j \leftarrow 0$;
while $|S^k| > 0$ *and* $\tau^k > 0$ **do**
  $z^k \leftarrow \max\{0, \max_{j\in S^k}\{z_j(y^k, S^k, z^k)\}\}$;
  $S^{k+1} \leftarrow S^k$;
  for $j \in S^k$ **do**
    $y^{k+1}_j \leftarrow y^k_j - \int_{z^k}^{z^{k-1}} \sum_{t\in[m]} \lambda_t g_t(\pi) P_t(j, S^k)\, d\pi$;
    if $y^{k+1}_j = 0$ **then**
      $z^*_j \leftarrow z^k$; $S^{k+1} \leftarrow S^{k+1}\backslash\{j\}$;
    end
  end
  $k \leftarrow k+1$;
end
Result: Cutoff vector $z^{SD(M,G,q)} := z^*$.

A.2. Cutoff Vector for Deferred Acceptance (DA)

Following Azevedo and Leshno (2016) and Abdulkadiroğlu et al. (2015), the cutoff vector $z^{DA(M,G,q)}$ in DA can be defined as the fixed point of a certain monotone operator. Given market M, define the demand function for item j under cutoff vector z as the mass of agents for whom this item is their favorite within their budget set,

$$D_j(z) := \sum_{t\in[m]} \lambda_t \mathbb{P}_{u\sim F_t, \pi\sim G_t}(j \in \arg\max_{j'\in J}\{u_{j'} : j' = 0 \text{ or } \pi_{j'} \geq z_{j'}\}. \tag{A.2}$$

Without loss of generality, let the space of priorities be $\Pi = [0, B]^n$, where $B > 0$ is a constant. Let $\mathbb{1}_j$ be the n-dimensional unit vector with an one in component j. Define the operator $DA : \Pi \to \Pi$ according to Algorithm 2.

Algorithm 2: Single iteration of the DA operator.

```
Input: old cutoff vector z ≥ 0;
for j ∈ [n] do
    if D_j(z) > q_j then
        z'_j ← z_j + inf{δ : D_j(z + δ𝟙_j) ≤ q_j};
    else
        z'_j ← z_j;
    end
end
Result: new cutoff vector DA(z) := z'.
```

Note that $D_j(z)$ is weakly decreasing in z_j and weakly increasing in all other components $z_{j'}$ with $j' \neq j$. This implies that $DA : [0, B]^n \to [0, B]^n$ is a monotone operator that maps the complete lattice $\Pi = [0, B]^n$ to itself, so by the Knaster-Tarski fixed point theorem, the set of fixed points $\{z : DA(z) = z\}$ is a non-empty lattice. Define $z^{DA(M,G,q)}$ to be the minimum element in this set.

Note that the above definition is valid even if the priority distribution G has atoms, so that the demand function (A.2) is not continuous. In this case, the update on z'_j in Algorithm 2 will implicitly accept all agents with priority for j equal to the cutoff, and the final allocation may overshoot the quota q, so the cutoff vector corresponds to the agent-optimal L-stable matching as defined in Biró and Kiselgof (2015). However, in all examples in this paper, the demand function (A.2) is continuous, so this phenomenon does not occur.

When the market M is regular, then the serial dictatorship cutoff $z^{SD(M,G,q)} = z^{DA(M,G',q)}$, where the utility distribution G'_t is defined by taking each realization $\pi \sim G_t$ and mapping it to the n-dimensional vector with equal components $(\pi, \pi, \cdots, \pi)$.[22] However, this may not be true if M is not regular: in the market in Example H.1 of Appendix H.1.1, the SD cutoff is $(1/2, 0)$ under priority distributions $G_1 = \text{Uniform}(1/2, 1)$ and $G_2 = \text{Uniform}(0, 1/2)$, while the DA cutoff is $(0, 0)$.

A.3. Cutoff Matrix for Top Trading Cycles (TTC)

The TTC cutoffs in Leshno and Lo (2018) arise from solving a set of differential equations modeling the trading of priorities among agents. Instead of re-deriving the cutoffs, I apply their existence result as a black box, which requires bridging the differences in the technical assumptions of the two papers.

[22] The is because when M is regular, there is a unique fixed point to the DA operator, as implied by the proof of Proposition H.1 in Appendix H.1.

In Leshno and Lo (2018), a continuum economy is given by $\mathcal{E} = (C, \tilde{\Theta}, \eta, q)$, where C is a set of items, $\tilde{\Theta} = \Pi_C \times [0,1]^{|C|}$ is the space of agent types, where Π_C is the set of permutations of C. Each agent type $\theta \in \tilde{\Theta}$ is a tuple $(\succ^\theta, r^\theta)$ where $\succ^\theta$ is a preference ranking of C and r is a vector of priorities for each item. η is a measure over $\tilde{\Theta}$ and q is a $|C|$-dimensional vector of quotas for each item with strictly positive entries in all components. An outside option is not explicitly modeled and they assume for convenience that all agents and items find each other acceptable, and there is an excess of agents $\eta(\tilde{\Theta}) > \sum_{j \in C} q_j$. Moreover, they assume that the measure η has a density ν, which is piecewise Lipschitz continuous except on a finite grid. Moreover, ν is bounded above and away from zero on its support: either $\nu(\theta) = 0$ or $0 < a \leq \nu(\theta) \leq b$ for some constants a, b. Given such an economy $\mathcal{E}$, their Theorem 2 implies that the TTC mechanism is well-defined, and there exists a cutoff matrix z^* such that the mechanism matches each agent type θ to his most preferred item in the budget set $\{j : r_k^\theta \geq z_{jk}^*\}$. The main differences with this paper are that 1) I explicitly model an outside option, which is always available to every agent; 2) I do not assume an excess of agents; and 3) I allow certain quotas q_j to be zero to make the statement of Theorem 1 cleaner.

I now define the $z^{TTC(M,G,q)}$ in my model based on their z^*, assuming that the priority distributions G satisfy the regularity conditions in Assumption 1, which is adapted from the assumptions in their paper.

Assumption 1. *For the given market M and quota vector q, the priority distribution G_t satisfies the following requirements for each segment t: Let $I = \{j \in [n] : q_j > 0\}$ be the set of items with strictly positive quota and $\tilde{G}_t$ be the projection of the measure G_t (originally defined over $[0,1]^n$) unto $[0,1]^I$ by integrating over the components not in I. The measure $\tilde{G}_t$ is continuous with density $\tilde{g}_t$, which is bounded above and bounded away from zero: for each $\pi \in [0,1]^I$, either $\tilde{g}_t(\pi) = 0$ or $a \leq \tilde{g}_t(\pi) \leq b$, where a, b are positive constants. Furthermore, g_t is piecewise Lipschitz continuous except on a finite grid: there exists a finite set $0 := d_0 < d_1 < d_2 < \cdots < d_L := 1$, such that the function $\tilde{g}_t$ is Lipschitz continuous in each open hyper-rectangle $(d_{k_1-1}, d_{k_1}) \times (d_{k_2-1}, d_{k_2}) \times \cdots \times (d_{k_{|I|}-1}, d_{k_{|I|}})$, where each $k_j \in [L]$.*

Define $I := \{j \in [n] : q_j > 0\}$ to be the set of items with strictly positive quotas. Let R be the set of permutations of the set $I \cup \{0\}$. For each $\rho \in R$, construct a dummy item 0_ρ representing the outside option of agents with this preference ranking over $I \cup \{0\}$. We need a distinct dummy item for each ranking ρ to prevent the profitable trading of outside options among agents with different preference rankings over $I \cup \{0\}$. Let $O = \{0_\rho : \rho \in R\}$ be the set of these dummy items.

Define the economy $\mathcal{E} = (C, \tilde{\Theta}, \eta, \tilde{q})$ in the notation of Leshno and Lo (2018), with $C = I \cup O$, $\tilde{\Theta} = \Pi_C \times [0,1]^{|C|}$, and quota vector $\tilde{q}$ such that $\tilde{q}_j = q_j$ for each $j \in I$ and $\tilde{q}_{0_\rho} = 1 + \sum_{t \in [m]} \lambda_t h_t(\rho)$ for each $0_\rho \in O$, where $h_t(\rho)$ be the probability that a utility vector $u \in \Theta$ drawn according to the utility distribution F_t is consistent with the ranking ρ, meaning that if $\rho = (j_1, j_2, \cdots)$, then $u_{j_1} > u_{j_2} > \cdots$. The above quota for the outside option 0_ρ is designed to be sufficiently large so that all of the original agents will be able to access their own outside option if desired.

Construct the measure η over $\tilde{\Theta} = \Pi_C \times [0,1]^{|C|}$ as follows. The idea is to make it consistent with the original preference distributions F and priority distributions G, while adding a sufficiently large mass of dummy agents so that there is an excess of agents in the aggregate. To ensure that the dummy agents do

not interfere with the assignment of the original agents, I make their priority for every item in C to be worse than the original agents. To do this, define the hypercubes $B_1 = [0.5, 1]^{|C|}$ and $B_2 = [0, 0.5]^{|C|}$. For each $r \in B_1$, define $\pi(r) \in [0, 1]^I$, such that for each $j \in I$, $\pi_j(r) = 2r_j - 1$. For each permutation ρ over $I \cup \{0\}$, define a corresponding permutation $\succ_\rho$ over $C = I \cup O$ by replacing the entry in ρ for the item 0 with the item $0_\rho \in O$, and by appending the remaining items in O in an arbitrary but fixed order at the end. Let the set of such $\succ_\rho$'s be R', and for each $\succ \in R'$, let the reverse mapping be $\rho(\succ)$. Define a fixed permutation $\succ_0$ over C that ranks every item in O above every item in I. Define the measure η on $\tilde{\Theta}$ with density ν as follows: for each $\succ \in \Pi_C$ and $r \in [0, 1]^{|C|}$,

$$\nu(\succ, r) = \begin{cases} 2^{|C|} \sum_{t \in [m]} \lambda_t \tilde{g}_t(\pi(r)) h_t(\rho(\succ)) & \text{if } \succ \in R' \text{ and } r \in B_1, \\ 2^{|C|} \sum_{c \in C} \tilde{q}_c & \text{if } \succ = \succ_0 \text{ and } r \in B_2, \\ 0 & \text{otherwise,} \end{cases} \tag{A.3}$$

where the density $\tilde{g}_t$ is defined as in Assumption 1 and the term $2^{|C|}$ is a normalization constant since the volume of B_1 and B_2 are $2^{-|C|}$ each. The economy $\mathcal{E} = (C, \tilde{\Theta}, \eta, \tilde{q})$ satisfies all the assumptions needed for Theorem 2 of Leshno and Lo (2018). Therefore, the cutoff matrix z^* exists. Define the TTC cutoffs in my model to be

$$z_{jk}^{TTC(M,G,q)} = \begin{cases} \max(2z_{jk}^* - 1, 0) & \text{if } j, k \in I, \\ 1 & \text{otherwise.} \end{cases} \tag{A.4}$$

Note that the cutoff $z_{jk}^{TTC(M,G,q)}$ is 1 if $q_j = 0$ or $q_k = 0$, so that no one can access items with zero quota, and no one can benefit from having a high priority to an item with zero quota.

Appendix B: Structure of DA in a Finite Market Model with Discrete Agents

Consider a model with discrete agents, in which there are l_t agents of segment t, where l_t is a non-negative integer, and the capacities are integral. Define the DA mechanism in the discrete model as follows. The inputs include a non-negative priority score matrix π and an integral quota vector q, where the component π_{ij} denotes the priority score of agent i for item j, and q_j denotes the number of copies of item j that can be allocated. Assume that no two agents have the same priority score to a given item. Moreover, each agent submits a preference ranking over $J = [n] \cup \{0\}$. The mechanism follows the two-step iterative process described in Section 2.1.2: Initialize the priority cutoff $z_j = 0$ for every item $j \in [n]$ and each agent i applies to his favorite item j for which $\pi_{ij} \geq z_j$. If the number of applicants for an item j exceeds the quota, then increase the cutoff z_j by the smallest amount so that the number of applicants with $\pi_{ij} \geq z_j$ is exactly equal to q_j. This bumps out the applicants with the worst priority scores until the quota is not violated. In the next iteration, these bumped applicants apply to their next favorite item for which their priority score meets the cutoff.

In this model, define an agent's budget set as the set of items he can be assigned to if he ranks that item first and the outside option second, assuming that everyone else submits their true preference rankings and the priorities and quotas are fixed. Note that this definition is independent of the agent's own preferences. The following proposition is the basis of the argument in Section 3.3 for using the LP in Sections 3.1 to bound the performance of the DA mechanism in the discrete model.

PROPOSITION B.1. *In the agent-proposing deferred acceptance algorithm, every agent is assigned to his favorite item among his/her budget set.*

Proof of Proposition B.1 The desired result is implied by the strategyproofness of agent-proposing DA for the agents (Roth 1982, Dubins and Freedman 1981, Abdulkadiroğlu and Sönmez 2003). To see this, suppose that in the agent-proposing DA, agent i is rejected by item $j_1, \cdots, j_{k-1}$ and finally assigned to item j_k. By the strategyproofness of agent-proposing DA, if i had ranked j_k first, i would still be matched to j_k. This is because if i cannot get j_k by ranking it first, but can get it by ranking other items first, then the agent has incentives to misreport preferences if j_k happened to be his true first choice.

It suffices to show that none of the previous items, $j_1, \cdots, j_{k-1}$ are in the budget set. This again follows from strategyproofness, because if i can get any of these items by ranking it first, then i would have an incentive to deviate because that improves upon i's current assignment of j_k. □

Appendix C: Decomposition Technique for a General Objective Function

Let w be a m-dimensional vector, and p a $m \times (n+1)$ matrix. Let $\Omega(w,p)$ be a concave function that is weakly increasing in each component of w. For each agent segment $t \in [m]$, let $\Psi_t \subseteq 2^J$ be a constraint set specifying the valid budget sets. Let U_t and P_t be defined as in (7) and (8). Consider the following convex program, which generalizes the LP in Section 3.1:

$$\begin{aligned}
&\text{Maximize:}_{y} && \Omega(w,p) && && \text{(C.1)}\\
&\text{subject to:} && y_{tS} \geq 0 && \text{for each segment } t \in [m] \text{ and } S \in \Psi_t, && \text{(C.2)}\\
&\text{(Expected utility)} && \sum_{S \in \Psi_t} U_t(S) y_{tS} = w_t && \text{for each } t \in [m], && \text{(C.3)}\\
&\text{(Assignment probability)} && \sum_{S \in \Psi_t} P_t(j,S) y_{tS} = p_{tj} && \text{for each } t \in [m],\ j \in J, && \text{(C.4)}\\
&\text{(Valid probabilities)} && \sum_{S \in \Psi_t} y_{tS} = 1, && \text{for each } t \in [m]. && \text{(C.5)}
\end{aligned}$$

The above convex program has a large number of decision variables. A standard approach to solve such problems is simplicial decomposition, which is a generalization of the column generation technique from linear programming. This is an iterative algorithm that maintains for each segment t a subset $\tilde{\Psi}_t \subseteq \Psi_t$, which is initialized to be a small set of assortments that makes the above convex program feasible. For the LP in Section 3.1, one can initialize $\tilde{\Psi}_t$ to contain only the outside option assortment $\{0\}$.

In each iteration, we first solve a master problem, which is the same as the original formulation above except that we replaces Ψ_t by the smaller set $\tilde{\Psi}_t$, which greatly simplifies the optimization as the cardinality of $\tilde{\Psi}_t$ is small by construction. Given an optimal solution of the master problem, compute a super-gradient (α, r) of the concave objective $\Omega(w,p)$, such that α_t is the component of the super-gradient for w_t and r_{tj} is the component for p_{tj}. (In the special case in which the concave objective can be expressed as a linear objective subject to linear constraints, a super-gradient naturally arises as the dual LP variables.) Note that $\alpha_t \geq 0$ by the assumption that W is weakly increasing in the component of w_t. Consider the following

sub-problem for segment t, which is equivalent to the socially optimal assortment planning problem (14) except for the additive constant of r_{t0},

$$\max_{S\in\Psi_t}\left\{\alpha_t U_t(S)+r_{t0}+\sum_{j\in[n]}(r_{tj}-r_{t0})P_t(j,S)\right\}. \tag{C.6}$$

Suppose that the optimal objective value to the sub-problem (C.6) does not exceed zero for every segment t, then terminate the algorithm and return the solution y^* to the master problem. Otherwise, for each segment t such that the optimal objective to (C.6) is strictly above zero, append the optimal solution S_t^* to (C.6) to the set $\tilde{\Psi}_t$, and iterate again to resolve the master problem.

The above approach is guaranteed to find an optimal solution to the original convex program upon termination. In practice, the number of iterations needed is often small, making it a practical approach to solve large scale optimization problems. See Von Hohenbalken (1977) for the development of the theory of simplicial decomposition and proof of correctness. For a more recent exposition, see Chapter 4 of Bertsekas (2015). In the special case in which the convex program can be formulated as the LP, the above technique is called column generation, which is explained in Chapter 6 of Bertsimas and Tsitsiklis (1997).

Appendix D: Details of the Simulation Model for Boston

This section describes the distributional assumptions behind the simulation results of Table 1. The same assumptions and parameters are used in Ashlagi and Shi (2015), and the following descriptions are reproduced here for completeness.

D.1. Student Population

The BPS data partitions students based on geographic location into 868 small neighborhoods, called geocodes. I model each neighborhood as a segment t.[23] The data also groups the 868 neighborhoods into 14 larger regions, which are based on natural divisions of the city. (For example, downtown is a region by itself.)

The number of students who apply from each neighborhood t is modeled as follows: Define a normal random variable with mean 4294 and standard deviation 115. This represents the total number of applicants and is estimated from four years of data from 2010-2013. To accommodate medium-scale regional variations, generate an independent normal random variable for each of the 14 regions, which represents the proportion of students who come from this region. The means and standard deviations are shown in Table 2, and are based on the sample means and sample standard deviations from the four years. The total number of students of each region is the product of the overall normal variable with the region-specific term, rounded to the nearest integer. Having computed this regional total, sample the neighborhood t of each student based on the historic density in 2010-2013. Generated in this way, the simulated number of applicants from each neighborhood is positively correlated both across the city and within each region, with the levels of correlation matching the historic data.

[23] It is also possible to consider other differences across students, such as race, older siblings, special education needs, and language learning needs. However, for clarity of analysis, I focus on the geographic aspects following Ashlagi and Shi (2015).

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

Table 2 Means and standard deviations of the proportion of Kindergarten-2 applicants from each region, estimated using the sample means and standard deviations from four years of historical data from 2010-2013.

D.2. Utility Distributions

The preferences of students are modeled using a MNL utility distribution, where the utility of a student i from neighborhood t for school j is modeled as

$$u_{ij} = \bar{u}_{tj} + \beta\epsilon_{ij}, \tag{D.1}$$

$$\bar{u}_{tj} = Q_j - d_{tj} + \gamma \cdot \mathbb{1}(j \in S_t^{walk}). \tag{D.2}$$

The data in the above equations are d_{tj}, and S_t^{walk}, and the parameters are Q_j, γ and β. As in Section 6.1, d_{tj} is the walking distance from the centroid of neighborhood t to school j according to Google Maps, and S_t^{walk} is the set of schools within the walk-zone of neighborhood t. Q_j is an estimated school-specific fixed effect capturing overall school popularity, which is called the *inferred quality* of school j. $\beta > 0$ is the scale of the random term in the utility. γ is a coefficient for living within one-mile. The expected utilities and choice probabilities under budget set S are given by the following closed form expressions, where $\gamma_{Euler} = 0.5772...$ is Euler's constant.

$$U_t(S) = \beta\left[\log\left(\sum_{j\in S} e^{\overline{u}_{tj}/\beta}\right) + \gamma_{Euler}\right], \tag{D.3}$$

$$P_t(j,S) = \frac{e^{\overline{u}_{tj}/\beta}}{\sum_{j'\in S} e^{\overline{u}_{tj'}/\beta}}. \tag{D.4}$$

Equation (D.2) normalizes the distance coefficient to one, instead of the scale parameter of the Gumbel distribution, so that the utilities can be interpreted in terms of distance. The parameters Q_j, γ and β are estimated from submitted preference rankings from 2013, using the maximum likelihood technique of Hausman and Ruud (1987). The estimates are shown in Table 3. The inferred qualities of schools are plotted on a map in Figure D.1b, and the lowest quality Q_j is normalized to zero.

Parameter	Value	Interpretation
Q_j	0–6.29	Quality of schools. For a school of ΔQ additional quality, holding fixed other components, a student would be willing to travel ΔQ miles further. The value for each school is graphically displayed in Figure D.1b.
γ	0.86	Additional utility for going to a school within the walk-zone.
β	1.88	Scale parameter of the Gumbel term.

Table 3 Parameters of the MNL utility distribution, estimated from preference data from 2013. The values can be interpreted in units of miles (how many additional miles a student is willing to travel for one unit of this variable).

This model estimates preference intensities from data on preference rankings. The logic is as follows: assuming that the differences in how students from different neighborhoods rank schools can entirely be explained by distances to schools, then one can infer students' preference intensities by observing how quickly they trade these preferences for distance. For example, suppose that students generally prefer school A over school B. In neighborhoods equidistant from the two schools, then one would expect more students rank A before than B. However, as one moves through neighborhoods going closer to B, one may see students preferring B more. By observing the speed at which their preference rankings change, one would have a rough estimate of how strongly students on average prefer A over B without considerations of distances.

It is possible also to add non-linear terms of distance as well as interactions between students' race and income and the school's demographics and test-scores, as in Pathak and Shi (2013), Shi (2015) and Pathak and Shi (2019). Pathak and Shi (2019) also compare the MNL utility distribution to a mixed MNL utility distribution, which allows for rich correlations in the unobserved component ϵ_{ij} across schools, and they show that the models perform similarly in prediction accuracy in the Boston data.

D.3. Schools and Quotas

There are $m = 77$ schools in the dataset, each of which has a capacity constraint c_j. Figure D.1a plots the school capacities and locations. The plot also shows the location of 19 so-called capacity schools, which are schools at which BPS can expand capacities at to accommodate excess demand. To reflect the fact that all applicants must be eventually offered a seat due to mandatory schooling laws, I assign each neighborhood t a *default school*, which is the closest capacity school.[24] For simplicity, I treat the default school as the only outside option of each neighborhood, thus ignoring the possible substitution to non-BPS schools. Moreover, I do not count students assigned to the default school against the capacity of the school, which guarantees that every student can at least be assigned to his/her default school.

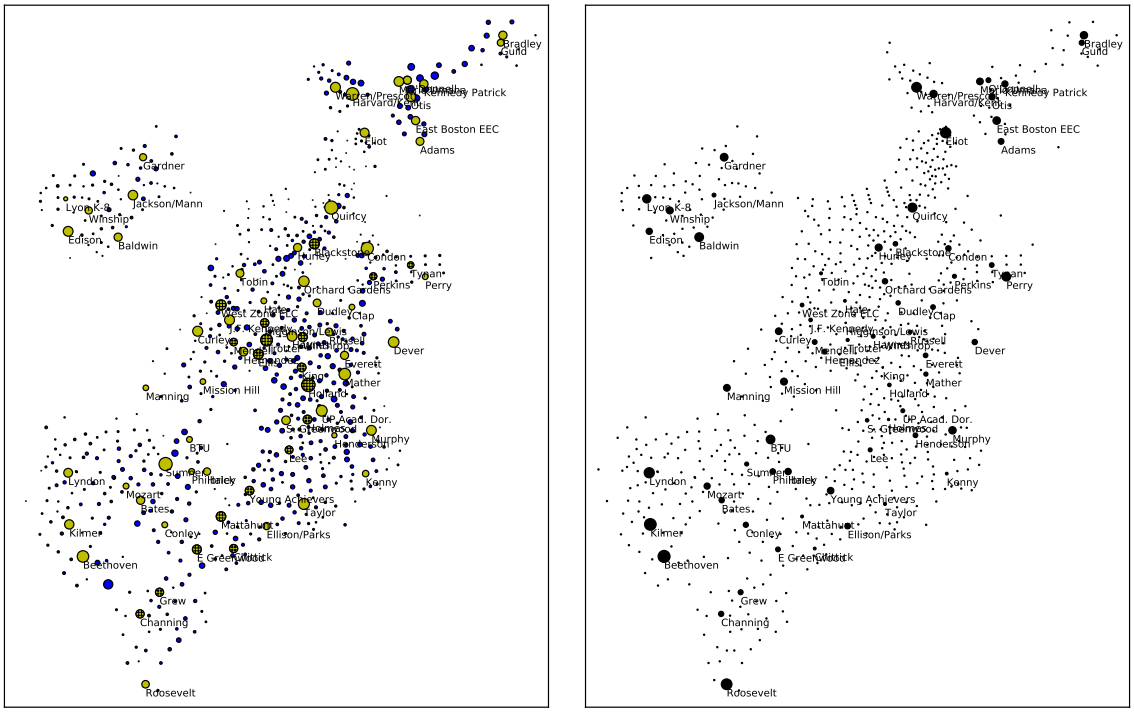
D.4. Student Assignment Plans

Each assignment plan in Table 1 of Section 6 specifies the set of school options that a student can rank, as well as the distribution of his/her priorities at the schools. Each student submits a ranking of schools in their choice set of arbitrary length, and the assignment is by the student-proposing deferred acceptance algorithm. As in Ashlagi and Shi (2015), I do not model the assignment of siblings and students with special needs, so the priorities structure being simulated is simplified.

D.4.1. The 3-Zone Plan In the 3-Zone plan in Table 1, the choice sets are as follows. The city is partitioned into three geographic zones as in Figure D.2, and every student can rank any school within the zone where they live, as well as any school within a one-mile radius of their home, called their walk-zone. From 1988 to 2013, elementary school assignment in Boston was based on these three zones.

The priorities are as follows. Except for a few citywide schools, each school is divided into two virtual halves, a walk-zone half and an open half. The capacity of the walk-zone half is rounded up and the open half rounded down. The preference ranking of each student is extended to a ranking over halves, with students in

[24] In the Home-Based and optimized plans, the default school for a neighborhood is the closest capacity school out of all such schools. In the 3-Zone plan, it is the closest capacity school within the zone-based choice menu.



(a) Schools and Students (b) School Quality

Figure D.1 **The diagram on the left shows the distribution of students and the capacities of schools. Each blue circle represents a neighborhood, with its size proportional to the expected number of students from that neighborhood. Each yellow circle represents a school, with its size proportional to the number of available seats for the grade Kindergarten-2 in 2013. The shaded schools are those at which BPS is able to expand capacity if needed, and are referred to as the capacity schools. The right shows estimates of Q_s (inferred quality) from the 2013 data. The size of the circle is proportional to the estimated Q_s, with higher quality schools having larger circles.**

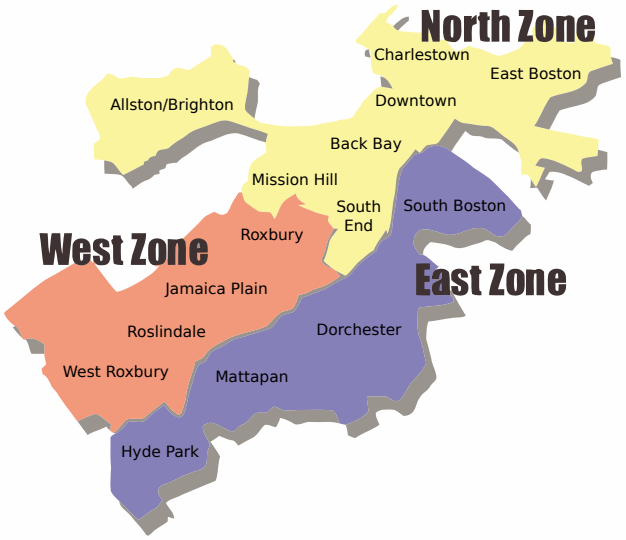


Figure D.2 **Illustration of the 3-Zone student assignment plan implemented in Boston from 1988-2013.**

a school's walk-zone applying to the walk-zone half first and the open half second, and students outside the walk-zone applying to the open-half first and the walk-zone half second. The walk-zone half prioritizes all walk-zone students over all non-walk zone students, while the open-half treats both kind of students equally. To break remaining ties, each student i is given an i.i.d. random number δ_i.

D.4.2. The Home Based Plan Since 2014, BPS has been using a Home-Based plan for elementary school assignment, which has undergone minor modifications since it was first implemented. The Home-Based plan in Table 1 corresponds to the original version in 2014, in which the choice set of each student is the union of the following sets: any school within 1 mile straight line distance; the closest 2 Tier 1 schools;[25] the closest 4 Tier 1 or 2 schools; the closest 6 Tier 1, 2 or 3 schools; the closest school with Advanced Work Class (AWC); the closest Early Learning Center (ELC); the 3 closest capacity schools;[26] the city-wide schools, which are available to everyone in the city. Furthermore, for students living in parts of Roxbury, Dorchester, and Mission Hill, their choice set also includes the Jackson/Mann school in Allston/Brighton.

The priorities are as follows: Students living in East Boston have priority for East Boston schools. Students outside of East Boston have priority for non-East Boston schools. To break remaining ties, each student i is given an i.i.d. random number δ_i.

Appendix E: Theoretical Upper Bound for the Finite Market Stochastic Model

This section explains the theoretical upper bound used in Figure 4 to bound the optimality gap of the optimized plan of Section 6.3 in the finite market stochastic model. The ideal (but intractable) optimization in the finite market model would involve evaluating all possible combinations of choice menus and priority distributions by simulation. The continuum model is only an approximation, and the optimized plan from section 6.3 may not be optimal in the finite market model due to the following discrepancies:

1. The number of agents and capacities of items may not be large enough for the continuum model to be an adequate approximation of the discrete model. This concern is explored in detail in Appendix G, where it is shown that market size is not the issue.
2. The mass of students from each neighborhood is deterministically equal to λ_t in the continuum model, whereas in the discrete model, they are randomly drawn and are positively correlated across the city, as described in Appendix D.1.
3. The ideal optimization would be using the exact constraints for the bus coverage area and the number of busing choices (28) and (29), rather than the linearized versions (26) and (27).
4. The construction in the proof of Theorem 1 requires controlling both the quotas and priority boosts in DA-STB to implement the budget set probability matrix y^{**}, whereas the assignment plan considered in the Boston reform are not allowed to set arbitrary quotas.

[25] Since 2013, BPS has been partitioning schools into four tiers based on standardized test-scores, with Tier 1 being the best.

[26] 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, the capacity schools are exactly the Tier 4 schools.

Nevertheless, as explained in Section 3.3 and formally established in Appendix B, budget sets are well defined in the DA mechanism even in the finite market model. Moreover, constraints (26) and (27) are relaxations of the exact busing constraints (28) and (29), so anything that satisfies the latter satisfies the former. This implies that if $\lambda_t y_{tS}$ is interpreted as the expected number of neighborhood t students who receive a budget set S, then y would be a feasible solution to the LP (18)-(27) in Section 6.1, if the bounds (B_1, B_2, B_3) are set to be above the simulated busing requirements of the Home-Based plan in rows 3 through 5 of Table 1. The optimal objective value of the LP is an upper bound to what can be achieved in the finite market stochastic model by any combination of choice menus and priorities, and is the bound used in Figure 4.

Appendix F: Robustness of the Optimization to Errors in Parameters

The optimization in Section 6 depends on distributional assumptions on the student population and preferences. In this section, I evaluate the robustness of the optimized plan to errors in these assumptions.

The population distribution from Appendix D.1 is based on data from 2010-2013, and the utility distribution from Appendix D.2 is estimated from students' submitted preferences from 2013 under the 3-Zone plan. In this section, I re-evaluate the various plans by using the actual application population from 2014 and by using a utility distribution estimated from 2014 preferences under the Home-Based plan. The amount of perturbation in parameters from this computational experiment represents the typical perturbation one may observe after an assignment plan reform.

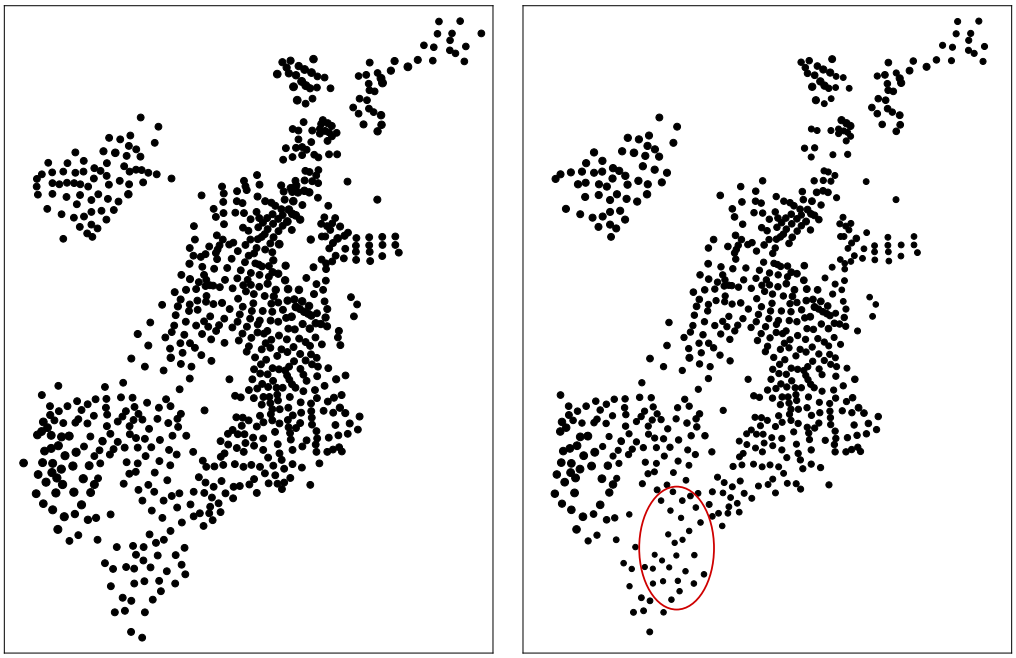
The magnitude of the perturbation in parameters is significant. For the population, instead of a forecasted total of 4294 students, only 3964 students applied in 2014. This difference is about 3 times the standard deviation of 115 students in the original population distribution. For the utility distribution, the inferred qualities of schools changed, with the average absolute change being about 0.69. (Recall that in the utility distribution, magnitudes are normalized to distance, so this is equivalent to changing students' travel distances to a school by ± 0.69 miles.) The estimated scale of the Gumbel distribution β changed from 1.88 to 1.64, and the estimated effect of coefficient for the walk-zone term γ changed from 0.86 to 0.37.

The simulation results using these updated parameters are in Table 4. Note that although the parameters for evaluation changed, the optimized plan (from Section 6) is based on parameters from before, and has not been re-optimized. As seen in Table 4, the optimized plan still dominates the 3-Zone and Home-Based plans in busing savings, expected utilities of students and predictability, despite having sizable errors in its demand estimates.

However, the magnitude of its improvements over the Home-Based plan for the lowest and the 10th percentile expected utilities are less than before. Figures F.1a and F.1b compare the expected utility for each neighborhood based on the original and the updated demand estimates. One can see that much of the decrease is in the Hyde Park region of Boston, which is shown using a red oval. To understand what happened, I compare the school qualities from the original and updated utility distributions. As can be seen in Figures F.2a and F.2b, the inferred qualities of schools (defined in Appendix D.2) in Hyde Park is much lower in 2014 than in 2013. This shows that the equity performance of the optimized plan is sensitive to systematic changes in school quality in a particular region, which is unavoidable for any location based assignment plan.

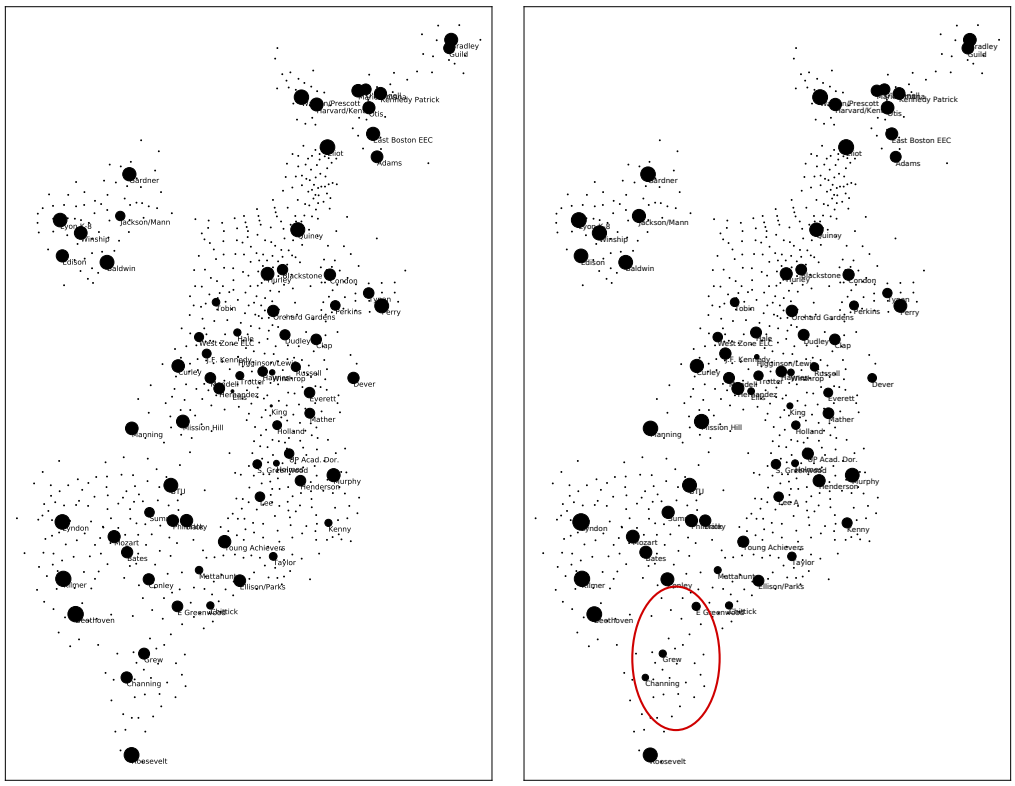
		3-Zone	Home-Based	Optimized
	Descriptive statistics			
(1)	Av. # of choices	29.21	14.78	14.46
(2)	Av. miles to assigned school	1.66	1.21	1.23
	Busing requirement			
(3)	Miles bused per student	1.15	0.60	0.63
(4)	Av. bus coverage area	-	-	-
(5)	Av. # of busing choices	22.26	8.17	8.02
	Expected utilities of neighborhoods			
(6)	Weighted average	6.68	6.30	6.73
(7)	10th percentile	5.77	5.48	6.10
(8)	Lowest	4.42	4.73	5.33
	% getting top choices in menu			
(9)	Top 1	59.8%	60.9%	76.5%
(10)	Top 3	82.0%	82.7%	93.9%

Table 4 Re-evaluation of the 3-Zone, Home-Based, and optimized plans using the actual population data from 2014 and a re-estimated utility distribution. The optimized plan is based on the old parameters and has not been re-optimized using the new data. All the results are averages from 100,000 independent simulations. The average bus coverage area (row 4) is omitted because coverage areas are not affected by the update in demand estimates.



(a) Original parameters (b) Updated parameters

Figure F.1 These plots show the expected utilities of neighborhoods under the optimized plan. Each circle represents a neighborhood and the size of the circle is proportional to the expected utility. The left plots the expected utilities under the demand estimates in Appendix D. The right plots the expected utilities under the actual 2014 population and re-estimated utility distribution. The red oval in the plot on the right shows the biggest area of utility decrease. This corresponds to the Hyde Park region of Boston.



(a) Original (b) Updated

Figure F.2 These plots show the inferred qualities of schools in the utility distribution. Each circle represents a school and the size of the circle is proportional to the inferred quality. The left plots the original quality estimates based on 2013 data. The right plots the updated quality estimates based on 2014 data. The red oval shows the decrease in quality estimates for the Hyde Park region.

Appendix G: Accuracy of the Large Market Approximation

The continuum model in this paper corresponds to the limit of a discrete model in which the number of agents of each segment goes to infinity. However, in the school choice application in Section 6, the expected number of students is 4294, and there are 868 segments, so the average number of students per segment is about 5. Nevertheless, there are reasons to expect the large-market approximation to be reasonable. First, neighborhoods that are close to one another tend to have similar choice sets, utility distributions, and priority distributions, so there are regional pooling effects. Second, the independence in preferences make it so that the number of students who prefer a school from a certain area converges quickly to its expectation.

In this section, I empirically test how well the large-market approximation performs on the Boston dataset, by comparing the outcomes of interests in the optimized plan of Section 6.3 as predicted by the continuum model with the outcomes from discrete simulations.

There are three sources of discrepancy between the two types of estimates. The first is that the market size in the Boston data is not large enough for the large-market approximation to set in. The second is that the simulations involve randomness in the student population (see Appendix D.1), while the continuum model assumes the size of each agent segment is fixed. The third is that ideally, the quotas for the optimized plan

	Continuum model	Discrete model	% Difference
Descriptive statistics			
(1) Av. # of choices	14.610	14.610	0.00%
(2) Av. miles to assigned school	1.274	1.278	0.27%
Busing requirement			
(3) Miles bused per student	0.600	0.611	1.89%
(4) Av. bus coverage area	-	-	-
(5) Av. # of busing choices	8.155	8.155	0.00%
Expected utilities of neighborhoods			
(6) Weighted average	7.555	7.488	0.89%
(7) 10th percentile	7.394	7.293	1.37%
(8) Lowest	7.394	7.193	2.72%
% getting top choices in menu			
(9) Top 1	81.83%	79.19%	3.22%
(10) Top 3	93.36%	93.15%	0.23%

Table 5 Comparison of the predictions from the continuum model with the estimates from discrete simulations when evaluating the optimized plan from Section 6.3. The percentage difference is equal to the absolute difference between columns 1 and 2, divided by the value in column 1. The numerical estimates for the discrete model are based on 100,000 independent simulations, and they are different from those of Table 1 because in order to focus on market size aspect of the approximation, I use the school quotas as prescribed in Proposition H.3, and I reduce the randomness in the applicant population. See Table 1 for explanation of the rows. The average bus coverage area (row 4) is omitted as it is identical in the two models by construction.

should be determined from the left hand side of (24) as in Proposition H.3, whereas in the simulations I use the capacities as quotas due institutional constraints.

In the exercise below, I focus on the first issue of market size. To do this, I modify the population distribution and quotas in order to remove the latter two sources of discrepancy. For the quotas, I use the left hand side of (24) at the optimal y^{**} (defined in Section 6.3), rounded to the nearest integer. For the number of students of each neighborhood, I set it to be as close to the expectation λ_t. Precisely speaking, if N_t is the number of applicants from neighborhood t, I define N_t to be a random variable that takes one of the values $\{\lfloor\lambda_t\rfloor, \lceil\lambda_t\rceil\}$ withits expectation being equal to λ_t.

Table 5 tabulates the simulation results of the optimized plan using the modified population distribution and quotas, and compare with the predictions from the continuum model. As can be seen, the estimates are all very similar between the continuum model and the discrete model, with the largest discrepancies coming from the minimum expected utility of neighborhoods, as well as the probability of getting their top choice within menu. For all metrics, the simulation results are within 3% of the large market estimates. This shows that the market size in Boston is large enough for the continuum model to be an adequate approximation.

Appendix H: Proofs

H.1. Proof of Theorem 1: Characterization of Mechanisms

Part a-i) of Theorem 1 follows from Proposition H.1, and part a-ii) from Proposition H.2 and H.3. Part b) follows from Proposition H.2, H.4 and H.5. The necessity of the assumption that M is regular for part a) is explained in Appendix H.1.1.

PROPOSITION H.1 **(Flexibility of the DA Mechanism)**. *Given a market M and an arbitrary budget set probability matrix $y \in Y^M$, define the priority distribution G_t for each segment t as follows. For an agent i, sample a set $S_i \subseteq J$ according to the probability vector y_t (with probability y_{tS_i}). Define the agent's priority score for item j as*

$$\pi_{ij} = \mathbb{1}(j \in S_i) + \delta_i, \tag{H.1}$$

where $\delta_i \sim$ Uniform$(0,1)$. Define quota q_j as the left hand side of (11), which is the mass of agents assigned to j under budget set probabilities y. If M is regular, then the DA mechanism with priority distributions G and quota vector q implements y.

Proof of Proposition H.1. Let z be the n-dimensional vector with all components equal to 1. Note that z is a fixed point of the DA operator defined in Algorithm 2 of Appendix A.2, $DA(z) = z$. This is because the demand function (A.2) evaluates to

$$D_j(z) = \sum_{t \in [m]} \lambda_t P_t(j, S) y_{tS} = q_j, \tag{H.2}$$

for every item $j \in [n]$ by the construction of the quota vector q. The space of priorities is $\Pi = [0,2]^n$. Let X be the priority-based allocation mechanism associating each priority $\pi \in \Pi$ with the budget set

$$B_\pi^X := \{j \in J : j = 0 \text{ or } \pi_j \geq z_j\}. \tag{H.3}$$

For any agent segment t and set $S \subseteq J$, note that the budget set is equal to S with probability exactly y_{tS}.

Let $z' = z^{DA(M,G,q)}$ be the DA cutoff as defined in Appendix A.2. This is the minimum element of the lattice of fixed points, $\{z' \in \Pi : DA(z') = z'\}$. Define X' to be DA mechanism with priority distribution G and quota q. This mechanism associates each priority $\pi \in \Pi$ with the budget set $B_\pi^{X'}$, which is as in (H.3) but with z replaced by z'. If the two mechanisms have the same budget set probabilities, $y^X = y^{X'}$, then we are done.

Suppose on the contrary that $y^X \neq y^{X'}$, then since $z' \leq z$ element wise, we have that the set inclusion $B_\pi^{X'} \supseteq B_\pi^X$ holds for every priority realization $\pi \in \Pi$. Moreover, the inclusion is strict with positive probability for at least one segment t. By the assumption that M is regular, we have

$$\Lambda - \sum_{j \in [n]} D_j(z') = \sum_{t \in [m]} \lambda_t \mathbb{E}_{\pi \sim G_t}[P_t(0, B_\pi^{X'})] < \sum_{t \in [m]} \lambda_t \mathbb{E}_{\pi \sim G_t}[P_t(0, B_\pi^X)] = \Lambda - \sum_{j \in [n]} q_j, \tag{H.4}$$

where $\Lambda := \sum_{t \in [m]} \lambda_t$ is the total mass of agents and D_j is the demand function defined in (A.2). The two quantities in the middle are respectively the mass of agents assigned to the outside option in X' and X. Equation (H.4) implies that $\sum_{j \in [n]} q_j < \sum_{j \in [n]} D_j(z')$, which is a contradiction since $D_j(z') \leq q_j$ for each $j \in [n]$. This is because the DA mechanism always respects the quotas if the demand function D_j defined in (A.2) is continuous, which is true for the priority distribution specified in the proposition. □

PROPOSITION H.2 **(Properties of Mechanisms)**.

a) The budget set probabilities arising from DA-STB are always nested within segment.

b) The budget set probabilities arising from TTC are always nested and non-degenerate.

c) The budget set probabilities arising from SD are always nested and non-degenerate.

Proof of Proposition H.2. For a), if priorities are parameterized as $\pi_{ij} = b_{tj} + \delta_i$ as described in Section 2.1.2, and z is the cutoff vector, then the budget set of an agent of type t with tie-breaker $\delta_i \in [0,1]$ is $\{j \in J : j = 0 \text{ or } \delta_i \geq z_j - b_{tj}\}$, which is weakly increasing in δ_i with respect to set inclusion.

For b), Corollary 1 of Leshno and Lo (2018) states that the TTC cutoffs z^*_{jk} in their model for an economy $\mathcal{E} = (C, \tilde{\Theta}, \eta, q)$ are such that there exists a relabeling of items under which $z^*_{1k} \geq z^*_{2k} \geq \cdots \geq z^*_{kk} = z^*_{(k+1)k} = \cdots = z^*_{|C|k}$ for each $k \in C$. This implies that whenever item $j \in C$ is in the budget set, every item $j' \geq j$ is also in the budget set, so the budget sets are nested. Now, the TTC cutoffs in my model inherits this property by their construction in Appendix A.3, so budget sets are nested in TTC. For non-degeneracy, observe that the definition of the cutoffs in (A.4) is that that $q_j = 0$ implies that j is removed from everyone's budget set. Moreover, if $q_j > 0$ but the left hand side of (11) is zero, then the construction in Appendix A.3 embedding my model into that of Leshno and Lo makes j present in the budget set of all agents with certainty.

For c), if the cutoff vector is z, then the budget set for an agent with priority π is $\{j \in J : j = 0 \text{ or } \pi \geq z_j\}$, which is weakly increasing in π with respect to set inclusion, so budget sets are nested. Moreover, if $q_j = 0$ then the SD cutoff $z_j = 1$ by the first line of Algorithm 1. On the other hand, if $q_j > 0$ but the left hand side of (11) is zero, then $z_j = 0$ and the item is present in all budget sets. □

PROPOSITION H.3 **(Implementation using DA-STB)**. *For a regular market M, let $y \in Y^M$ be a budget set probability matrix that is nested within segment. If priority boost b_{tj} is set to be $\sum_{S \ni j} y_{tS}$ and quota q_j is set to be the left hand side of* (11)*, then the corresponding DA-STB mechanism implements y.*

Proof of Proposition H.3 The proof is analogous to that of Proposition H.1. Let z be the n-dimensional vector with all ones, then the demand function defined in (A.2) satisfies $D_j(z) = q_j$ for every $j \in [n]$, and z is a fixed point of the DA operator defined in Appendix A.2. Let X be the mechanism that offers each agent i of segment t with tie-breaker $\delta_i \in [0,1]$ the budget set $B^X_\delta := \{j \in J : j = 0 \text{ or } \delta \geq 1 - b_{ts}\}$. Note that X implements y by construction, $y^X = y$. Since M is regular, we can apply the same argument as in the proof of Proposition H.1 and get that if X' is the DA-STB mechanism with the given priority boosts and quotas, then $y^{X'} = y^X$. □

PROPOSITION H.4 **(Implementation using Serial Dictatorship)**. *For any market M, let $y \in Y^M$ be a budget set probability matrix that is nested and non-degenerate. Let the budget sets that occur with positive probability, $\{S : y_{tS} > 0 \text{ for some } t \in [m]\}$, be parameterized as $\{S^1, S^2, \cdots, S^K\}$, where $S^1 \supsetneq S^2 \supsetneq \cdots\cdots \supsetneq S^K$ and K is the number of distinct sets. Define the priority distribution G_t for each segment t as follows: let the priority of an agent i be*

$$\pi_i = \frac{1}{K}(K - b_i + \delta_i), \tag{H.5}$$

where $\delta_i \sim$ *Uniform*$(0,1)$ *and* b_i *is a discrete random variable that equals* k *with probability* y_{tS^k} *for each* $k \in [K]$. *All random variables are independent. Let* d_j *be the total mass of item* j *assigned under* y, *which is equal to the left hand side of* (11). *Define the quota for each item* $j \in [n]$ *as*

$$q_j = \begin{cases} d_j & \text{if } d_j > 0, \\ 1 & \text{if } d_j = 0 \text{ and } j \in S^1, \\ 0 & \text{otherwise.} \end{cases} \tag{H.6}$$

Serial dictatorship (SD) with priority distributions G *and quota vector* q *implements* y.

Proof of Proposition H.4 Let $J' = S^1$. For each $j \in J'$, define $k(j) = \max\{k' : S^{k'} \ni j\}$. Let z be the SD cutoff vector defined in Appendix A.1. It suffices to show that for each $j \in J \backslash J'$, $z_j = 1$, and for each $j \in J' \backslash \{0\}$, $z_j = (K - k(j))/K$. The former claim follows from the assumption that y is non-degenerate, so $j \notin J' = S^1$ and $d_j = 0$ imply that $q_j = 0$ by (H.6), and the item is assigned a cutoff of 1 in the first line of Algorithm 1. The latter claim follows from the observation that the sets $S^1, \cdots S^K$ defined in the statement of the proposition correspond exactly to the sets defined with the same notation in Algorithm 1, and the quantity z^k in Algorithm 1 is equal to $(K-k)/K$ for each $k \in [K]$. Hence, in iteration k of Algorithm 1 ($k = 1, 2, \cdots, K$), the items in $S^k \backslash S^{k+1}$ are depleted and are assigned cutoff $(K-k)/K$. □

In a finite market, the TTC mechanism in which all items share the same priority ordering over agents is identical to SD, so the above result indicates that the TTC mechanism can also implement the desired budget set probability matrix y using the same priorities and quotas as above. Technically speaking however, the definition of the TTC mechanism based on Leshno and Lo (2018) requires priority distributions to be continuous, so items cannot share the same priorities for all agents. The following proposition resolves the issue by modifying the construction so that the tie-breakers are independently drawn across items.

Proposition H.5 **(Implementation using TTC)**. *For any market* M, *let* $y \in Y^M$ *be a budget set probability matrix that is nested and non-degenerate. Define the sets* $S^1, S^2, \cdots, S^K$ *and quota vector* q *as in Proposition H.4. Define the priority distribution* G_t *for each segment* t *as follows: let the priority score of an agent* i *for item* j *be*

$$\pi_{ij} = \frac{1}{K}(K - b_i + \delta_{ij}), \tag{H.7}$$

where the random variable b_i *is defined as in Proposition H.4 and* $\delta_{ij} \sim$ *Uniform*$(0,1)$ *are independently drawn. The TTC mechanism with priority distributions* G *and quota vector* q *implements* y.

Proof of Proposition H.5 Let $J' = S^1$. For each $j \in J'$, define $k(j) = \max\{k' : S^{k'} \ni j\}$. Let $z = z^{TTC(M,G,q)}$ be the TTC cutoff matrix as defined in Appendix A.3. Since every item $j \notin J'$ has quota $q_j = 0$ by (H.6), the definition of cutoffs in (A.4) sets $z_{jl} = 1$ if $j \notin J'$ or if $l \notin J'$. It suffices to show that for $j, l \in J'$, $z_{jl} = (K - k(j))/K$.

Let z^* be the cutoff matrix for the economy $\mathcal{E} = (C, \tilde{\Theta}, \eta, \tilde{q})$ as defined in Appendix A.3 using the notation of Leshno and Lo (2018), with $C = I \cup O$ where $I = \{j : q_j > 0\}$ is identically equal to J', and O is a set of dummy items representing outside options. It suffices to show that

a) $z^*_{jl} = (2K - k(j))/2K$ if $j \in S^{K-1}, l \in J'$.

b) $z^*_{jl} \leq \frac{1}{2}$ if $j \in S^K \backslash S^{K-1}, l \in J'$.

Let ζ be the projection of the measure η (which is defined on the space $\tilde{\Theta} = \Pi_C \times [0,1]^C$) onto the set $[0,1]^I$, then the support of ζ is contained in the set

$$Z := B(0, \frac{1}{2}) \cup B(\frac{1}{2}, \frac{K+1}{2K}) \cup B(\frac{K+1}{2K}, \frac{K+2}{2K}) \cdots \cup B(\frac{2K-1}{2K}, 1), \tag{H.8}$$

where $B(a,b) := [a,b]^I$, in which the components represent the priority score for each item in $I \subsetneq C$. By Definition 2 of Leshno and Lo (2018) and the definition of the marginal distribution H_a^c in their paper, the projection of their TTC path $\gamma(t)$ onto the subspace $[0,1]^I$ will always stay within the set Z. Hence, it must pass through the intermediate points $x_k := \mathbb{1}(2K-k)/2K$ for each $k \in [K]$, where $\mathbb{1}$ is the $|I|$-dimensional vector with all ones. For each $k \leq K-1$, when the projection of the TTC path passes through the point x_k, all of the items $j \in S^k \backslash S^{k+1}$ are depleted for the first time, so the cutoff $z^*_{jl} = (2K-k)/2K$ for all $l \in I$. When the curve passes through $x_K = 0.5\mathbb{1}$, all of the items $j \in S^K \backslash S^{K-1}$ are either just depleted or still with excess quota. In either case, the cutoff $z^*_{jl} \leq \frac{1}{2}$ for every $l \in I$. □

H.1.1. Limitations of the DA mechanism in non-regular markets The following example shows that when M is not regular, there are certain nested and non-degenerate budget set probabilities that cannot be implemented using the DA mechanism under any priorities or quotas.

Example H.1. There are two items of capacities 1/3 and 1 respectively, and two agent segments of unit mass. Segment 1 agents have uniformly random preferences for items 1, 2 and their outside option 0. Segment 2 agents prefer item 2 to item 0 to item 1 with 2/3 probability, and prefer outside option 0 best with remaining probability. The budget set probability matrix y with non-zero entries $y_{1\{0,1,2\}} = 1$ and $y_{2\{0,2\}} = 1$ cannot be implemented using the DA mechanism with any priority distributions and quotas. This is because if $q_1 < 1/3$, then $y_{1\{0,1,2\}} < 1$. However, if $q_1 \geq 1/3$, then the cutoff $z_1 = 0$, since the capacity is not violated when item 1 is offered to all agents. Such a cutoff is incompatible with $y_{2\{0,2\}} = 1$.

However, if we enrich the DA mechanism with additional policy levers, then it can implement arbitrary budget set probabilities in all markets. In the above example, one can implement y by eliminating the option of ranking item 1 for segment 2 agents. Such restrictions in choice sets are often implemented for school choice, as illustrated by the Boston application in Section 6.

Another policy lever that would enrich the space of outcomes that can be implemented by DA is to require agents to have a sufficiently high priority score to be eligible for an item. For example, if we set 1 to be the minimum priority score to be eligible for an item, then for any budget set probabilities y, the DA mechanism with priority distributions and quotas defined in the statement of Proposition H.1 implement y in any market, regular or non-regular. This is because the cutoff z' in the proof of Proposition H.1 would now have a lower bound of 1 in every component, so must equal to the all-1 vector. Having such a lower bound in priority score would also allow the DA-STB mechanism constructed in Proposition H.3 to implement arbitrary budget set probabilities that are nested within segment without requiring the market to be regular.

H.2. Proof of Theorem 2: Efficient Algorithms for Socially Optimal Assortment Planning

Throughout this section, I define $r_0 = 0$ for convenience, and denote the expected revenue of assortment S as $R(S) := \sum_{j \in S} r_j P(j,S)$. The socially optimal assortment planning problem is $\max_{S \in \Psi} \{\alpha U(S) + R(S)\}$. Part a) of Theorem 2 follows from Proposition H.7, part b) from Proposition H.11, part c) from Proposition H.12, and part d) from Proposition H.14.

H.2.1. MNL Utilities and Cardinality Constraint Without loss of generality, let the Gumbel distributed random term ϵ_{ij} in the MNL utility equation (15) have scale parameter 1. Define attraction weight vector $v \in (0,\infty)^J$ with component $v_j := \exp(\overline{u}_j)$ for each $j \in J = [n] \cup \{0\}$. The socially optimal assortment planning problem (14) under the MNL utility distribution can be written as

$$\max_{S\in\Psi} \left\{ \alpha \log\left(\sum_{j\in S} v_j\right) + \frac{\sum_{j\in S} r_j v_j}{\sum_{j\in S} v_j} \right\}, \tag{H.9}$$

where $\log(\cdot)$ is the natural logarithm and $\Psi \subseteq 2^J$ is the set of feasible assortments.[27] For (H.9) to be well-defined, the constraint set Ψ must not contain the empty set. This is satisfied by the constraint sets described in Section 4 as they all require the outside option 0 to be in every assortment.

The efficient algorithm for solving (H.9) under cardinality constraints is based on Proposition H.6, which says that to find an optimal assortment, it suffices to check through a certain candidate set $\mathcal{A} \subseteq \Psi$ of assortments. The proposition generalizes a result of Rusmevichientong et al. (2010), who take a similar approach to solve the revenue maximizing assortment planning problem ($\alpha = 0$). The candidate set $\mathcal{A}$ they define is identical to that in Proposition H.6, and they show that $\mathcal{A}$ can be found in $O(n^2 \log n)$ time and has cardinality at most $k(n+1-k)$. While their analysis requires additional regularity assumptions on v and r, I present a complete algorithm that removes all such assumptions at the end of this section.

Proposition H.6 **(Candidate Set for MNL)**. *For the MNL utility distribution and any constraint set Ψ that does not contain the empty set. Let $\mathcal{A} \subseteq \Psi$ be such that it contains an assortment from the set*

$$A(\lambda) = \arg\max_{S\in\Psi} \left\{ \sum_{j\in S} v_j (r_j - \lambda) \right\}, \tag{H.10}$$

for every $\lambda \in \mathbb{R}$. Then $\mathcal{A}$ contains an optimal solution to the socially optimal assortment planning problem (H.9) *for any $\alpha \geq 0$. Furthermore, for a given $\alpha \geq 0$, the set of all socially optimal assortments is equal to $\bigcup_{\lambda^*\in\Lambda^*(\alpha)} A(\lambda^*)$ for some set $\Lambda^*(\alpha) \subseteq \mathbb{R}$.*

Proof of Proposition H.6 Define $x(S) = \sum_{j\in S} v_j$, $y(S) = \sum_{j\in S} v_j r_j$, $D = \{(x(S), y(S)) : S \in \Psi\}$, and $g(x,y) = \alpha \log(x) + y/x$. The socially optimal assortment planning problem can be written as $\max_{(x,y)\in D} g(x,y)$.

Define $R = (0,\infty) \times (-\infty,\infty)$, which is an open and convex subset of $\mathbb{R}^2$. For any $\alpha \geq 0$, the function $g(x,y)$ in the domain R is quasi-convex and continuous, and it is strictly increasing in y. The desired result follows from the following lemma, which is based on the duality of convex sets in $\mathbb{R}^2$. Note that the set $A(\lambda)$ in (H.10) is equal to that in (H.14) of the lemma. □

Lemma H.1 **(Duality Lemma)**. *Let R be an open convex subset of $\mathbb{R}^2$ and D a non-empty and finite set of points from R. Let $g(x,y) : R \to \mathbb{R}$ be a continuous function that is strictly increasing in y, with the following lower level set B being a convex subset of $\mathbb{R}^2$,*

$$B = \{(x,y) \in R : g(x,y) \leq z^*\} \tag{H.11}$$

$$\text{where } z^* := \max_{(x,y)\in D} g(x,y) \tag{H.12}$$

[27] The objective function in (H.9) differs from the objective function in (14) by an additive constant of $\alpha\gamma_{Euler}$, where $\gamma_{Euler} = 0.5772...$ is Euler's constant.

Define

$$f(\lambda) = \max_{(x,y)\in D} \{y - \lambda x\}, \tag{H.13}$$

$$A(\lambda) = \arg\max_{(x,y)\in D} \{y - \lambda x\}, \tag{H.14}$$

$$h(\lambda, f) = \inf_{(x,y)\in R} \{g(x,y) : y = f + \lambda x\}. \tag{H.15}$$

Then the following supremum is attainable and equal to z^**,*

$$\sup_{\lambda\in\mathbb{R}} h(\lambda, f(\lambda)). \tag{H.16}$$

Furthermore, (x^*, y^*) *is an optimal solution to the optimization problem in* (H.12) *if and only if* $(x^*, y^*) \in A(\lambda^*)$ *for some optimal solution* λ^* *to the optimization problem in* (H.16)*.*

Proof of Lemma H.1. For convenience, denote the value of the supremum in (H.16) by z^{**}. First, I show that $z^* \geq z^{**}$. This is because for any $\lambda_0 \in \mathbb{R}$, let $(x_0, y_0) \in A(\lambda_0)$, then $(x_0, y_0) \in \{(x,y) \in R : y = f(\lambda_0) + \lambda_0 x\}$. This implies that $z^* \geq g(x_0, y_0) \geq h(\lambda_0, f(\lambda_0))$. Taking the supremum of both sides, we get $z^* \geq z^{**}$.

Conversely, I show that $z^* \leq z^{**}$. Consider the lower level set B in (H.11). Let (x^*, y^*) be an optimal solution to the optimization problem in (H.12). By the definition of z^* in (H.12), it must be that $B \supset D \ni (x^*, y^*)$. Since g is strictly increasing in y, (x^*, y^*) cannot be in the interior of B, but must lie on its boundary. By the duality of convex sets, there exists an outward pointing normal of B at (x^*, y^*) with direction $(-\lambda_0, 1)$, for some $\lambda_0 \in \mathbb{R}$. (The y-coordinate is 1 without loss of generality because g is strictly increasing in y.) Let $f_0 = y^* - \lambda_0 x^*$, then we have that both B and D are contained in the half-plane:

$$\{(x,y) : y - \lambda_0 x \leq f_0\}. \tag{H.17}$$

I now show that $h(\lambda_0, f(\lambda_0)) = g(x^*, y^*)$, from which it would follow that $z^* \leq z^{**}$ since $z^* = g(x^*, y^*)$ and $z^{**} \geq h(\lambda_0, f(\lambda_0))$. First, note that D being contained in the half-plane (H.17) implies that $f(\lambda_0) = f_0$, so $(x^*, y^*) \in A(\lambda_0)$. This in turn implies by the definition of h that $h(\lambda_0, f(\lambda_0)) \leq g(x^*, y^*)$. Now, suppose on the contrary that $h(\lambda_0, f(\lambda_0)) < g(x^*, y^*)$, then there must exist $(x_0, y_0) \in R$ such that $g(x_0, y_0) < c := g(x^*, y^*)$ and $y_0 - \lambda_0 x_0 = f(\lambda_0)$. Since R is open and g is continuous and increasing in y, there exists a sufficiently small ϵ_0, such that if $y_1 = y_0 + \epsilon$, then $(x_0, y_1) \in R$, $g(x_0, y_1) < c$ and $y_1 - \lambda_0 x_0 > f_0$. Therefore, $(x_0, y_1) \in B$ but $y_1 - \lambda_0 x > f_0$, which is a contradiction because B is contained in the half-plane specified by (H.17). Therefore, $h(\lambda_0, f(\lambda_0)) = g(x^*, y^*)$, as desired.

This shows that $z^* = z^{**}$. If (x^*, y^*) is an optimal solution to (H.12), then construct λ_0 as above from the outward pointing normal of B at the boundary point (x^*, y^*). We have that λ_0 is an optimal solution to (H.16), with $(x^*, y^*) \in A(\lambda_0)$. On the other hand, for any optimal solution λ^* to (H.16), for any $(x_0, y_0) \in A(\lambda^*)$, the argument in the first paragraph implies that (x_0, y_0) is also an optimal solution to (H.12). □

For cardinality constraint $\Psi = \{\{0\} \cup S : S \subseteq [n], |S \cap S_0| \leq k\}$, the optimization in (H.10) to construct the candidate set $\mathcal{A}$ can be represented geometrically as in Figure H.1, where each item $j \in [n]$ is represented by a line $f_j(\lambda) = v_j(r_j - \lambda)$ in $\mathbb{R}^2$. Under the assumption that the lines have distinct slopes and no three lines meet at a point, Rusmevichientong et al. (2010) derive a $O(n^2 \log n)$ algorithm for finding $\mathcal{A}$ by first sorting the pairwise intersections of the lines and the x-axis. When λ is restricted to an interval between

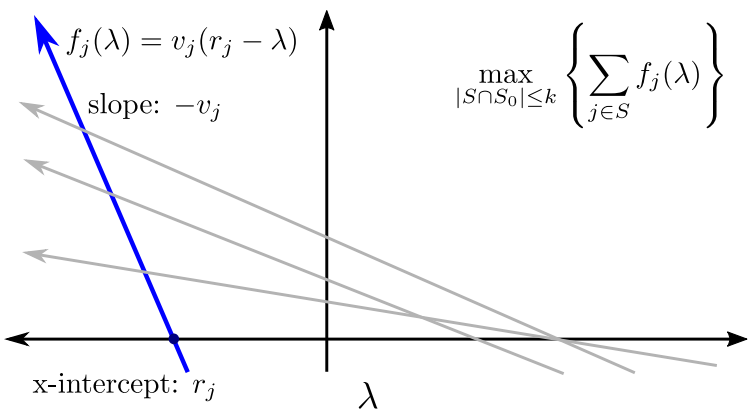


Figure H.1 Geometry of the optimization in (H.10) for computing the candidate sets $\mathcal{A}$ under cardinality constraints. Each item $j \in [n]$ is represented by a line $f_j(\lambda)$ with x-intercept r_j and slope $-v_j$. For any $\lambda \in \mathbb{R}$, an optimal assortment would be to take the top k items within S_0 with the highest and non-negative $f_j(\lambda)$, along with any item in $S\backslash S_0$ with a non-negative $f_j(\lambda)$.

two adjacent intersection points, the order of the lines is fixed, along with optimal assortments for (H.10): for each λ, an optimal assortment is to take the set of k highest non-negative lines among $j \in S_0$, and the non-negative lines among $j \in [n]\backslash S_0$.

The following algorithm uses the same idea to solve the socially optimal assortment planning problem (H.9), and includes additional details to handle generic data. In particular, it allows different items to have the same v_j, and allows arbitrarily many lines $f_j(\lambda)$'s to meet at the same point.

Algorithm 3: Socially Optimal Assortment Planning under MNL utilities and cardinality constraint

Data: attraction weight $v_j > 0$ for $j \in J = [n] \cup \{0\}$; revenue r_j for $j \in [n]$; parameter $\alpha \ge 0$; set $S_0 \subseteq [n]$ and maximum cardinality $k \in \{0, 1, \cdots, |S_0|\}$.

Step 1 (Sorting the intersection points): Define τ to be the ordered list from sorting the following set of tuples in lexicographically increasing order (comparing first component first, breaking ties using the second component, and so on):

$$\left\{ \left(\frac{v_i r_i - v_j r_j}{v_i - v_j}, -i, j \right) : i, j \in S_0, v_i > v_j \right\} \cup \{(r_i, -i, 0) : i \in [n]\}. \tag{H.18}$$

Each tuple in the first set encodes the intersection point of line $f_i(\lambda)$ and $f_j(\lambda)$ (see Figure H.1), with line j being higher to the left of the intersection and i being higher to the right. Each tuple in the second set encodes the intersection of line $f_j(\lambda)$ with the x-axis.

Step 2 (Ordering the items): Sort the items $j \in S_0$ according to the tuple $(v_j, r_j, -j)$ in lexicographic decreasing order. Let o_j denote the sorted order, with $o_j = 1$ if its tuple is the largest and $o_j = 2$ if it is the second largest, and so on. Set $o_j \leftarrow 0$ for $j \in [n]\backslash S_0$.

Step 3 (Computing the optimal objective value)

Data: τ from step 1, a copy of o from step 2, and parameters k and n.
Initialize $S \leftarrow \{j \in [n] : o_j \leq k\}$; $a \leftarrow \sum_{j \in S} v_j r_j$; $b \leftarrow \sum_{j \in S} v_j$; $z^* \leftarrow \alpha \log(a) - a/b$;
for $(\lambda, -i, j) \in \tau$ **do**
  if $j = 0$ *and* $i \in S$ `// Crossing between line` i `and the x-axis` **then**
    $a \leftarrow a - v_i r_i$; $b \leftarrow b - v_i$;
    $z \leftarrow \alpha \log(a) + a/b$;
    $S \leftarrow S \backslash \{i\}$;
  else if $j > 0$ *and* $o_i < o_j$ `// Crossing between lines` i `and` j`, with` $v_i > v_j$ `but` $o_i < o_j$`.` **then**
    Swap o_i and o_j; `// Ensure that the` o_i`'s are sorted as the` v_i`'s after a crossing`
    if $o_j = k$ *and* $i \in S$ **then**
      $a \leftarrow a - v_i r_i + v_j r_j$; $b \leftarrow b - v_i + v_j$;
      $z \leftarrow \alpha \log(a) + a/b$;
      $S \leftarrow S \cup \{j\} \backslash \{i\}$;
    end
  end
  if $z > z^*$ **then** $z^* \leftarrow z$, $\lambda^* \leftarrow \lambda$;
end
Result: Optimal objective z^* and corresponding λ^*.

Step 4 (Obtaining the optimal assortment by retracing Step 3)

Data: τ from step 1, a copy of o from step 2, λ^* from step 3, and parameters k and n.
Initialize $S \leftarrow \{j \in [n] : o_j \leq k\}$;
for $(\lambda, -i, j) \in \tau$ **do**
  if $j = 0$ *and* $i \in S$ **then**
    $S \leftarrow S \backslash \{i\}$;
  else if $j > 0$ *and* $o_i < o_j$ **then**
    Swap o_i and o_j;
    if $o_j = k$ *and* $i \in S$ **then**
      $S \leftarrow S \cup \{j\} \backslash \{i\}$;
    end
  end
  if $\lambda = \lambda^*$ **then**
    $S^* \leftarrow S$;
    break (exit the for loop);
  end
end
Result: Optimal assortment S^*.

The above algorithm has the property that the optimal assortment found is minimal according to a certain ordering over S, and the existence of such an ordering guarantees that the column generation algorithm based on the above for solving the LP in (18)-(27) will not cycle, as it is a version of the revised Simplex algorithm based on Bland's pivot rule (see Sections 3.4 and 6.1 of Bertsimas and Tsitsiklis (1997)).

The ordering is as follows: we say that set $S_1 \subseteq J$ is lexicographically less than $S_2 \subseteq J$ if either

a) it has more elements, $|S_1| > |S_2|$; or

b) it has the same number of elements ($|S_1| = |S_2|$), and the tuple from sorting the labels of the elements of S_1 in increasing order is lexicographically smaller than the analogous tuple from S_2. (If the smallest

label in S_1 is less than the smallest in S_2, then S_1 is lexicographically smaller; if there is a tie in the smallest labels, compare the second smallest, and so on.)

For example, the assortment $\{1,3,5\}$ is lexicographically smaller than $\{1,2\}$ because it has more elements, but it is bigger than $\{1,2,6\}$ as the latter has the same number of elements, and the tuple $(1,2,6)$ is lexicographically smaller than $(1,3,5)$.

PROPOSITION H.7 **(Analysis of Algorithm 3)**. *Algorithm 3 (steps 1 through 4) can be implemented in $O(n^2 \log n)$ time and solves the socially optimal assortment planning problem* (H.9) *with cardinality constraint set $\Psi = \{\{0\} \cup S : S \subseteq [n], |S \cap S_0| \leq k\}$. In particular, the z^* from Step 3 is the optimal objective value of* (H.9) *and the S^* from Step 4 is an optimal assortment. Moreover, if the items in S_0 are labelled in weakly decreasing order of v_j, with $v_i \geq v_j$ if $i < j$ for $i, j \in S_0$, then the S^* from Step 4 is lexicographically the least out of all optimal assortments.*

Proof of Proposition H.7 The cardinality of the list τ from Step 1 is at most $\binom{|S_0|}{2} + n = O(n^2)$, so the sorting in Step 1 takes $O(n^2 \log n)$ time, and Steps 3 and 4 can both be completed in $O(n^2)$ time as each iteration of the for loop can be done in a constant number of operations. (Note that the set S can be maintained as a binary array, so that checking set membership, and adding or removing an element can all be done in constant time.) Step 2 is sorting a set of cardinality $O(n)$, so can be completed in $O(n \log n)$ time. The total run time is $O(n^2 \log n)$.

The optimality of z^* from Step 3 follows from Proposition H.6 as follows. Let S^0 be the initial set S in Step 3. For each $t \geq 1$, define S^t and λ^t to be the value of S and λ at the end of the tth iteration of the for loop in Step 3. For each $\lambda \in \mathbb{R}$, define

$$A(\lambda) = \underset{S:|S \cup S_0| \leq k}{\arg\max} \left\{ \sum_{j \in S} v_j (r_j - \lambda) \right\}. \tag{H.19}$$

The sorting in Step 2 implies that the initial assortment $S^0 \in A(\lambda)$ for all $\lambda \in (-\infty, \lambda^1)$: for these values of λ, the lth highest line i has order $o_i = l$. Moreover, the updates in the for loop of Step 3 ensures that the vector o always maintains a correct rank ordering of the lines above the x-axis: for each value of λ corresponding to a crossing between lines, a swap is made between o_i and o_j whenever $v_i > v_j$ but $o_i < o_j$, implying that after the crossing point λ, lines that are flatter (lower v_i's) are recognized as being higher (lower o_i's). Moreover, the set S^t always contains the k largest positive lines $j \in S_0$ at $\lambda = \lambda^t$ and any positive line $j \in [n] \backslash S_0$, and it never contains any negative lines. Hence, for each $t \leq |\tau| - 1$, $S^t \in A(\lambda)$ for all $\lambda \in [\lambda^t, \lambda^{t+1}]$, and $S^{|\tau|} = \{0\} \in A(\lambda)$ for all $\lambda \in [\lambda^{|\tau|}, \infty)$. Therefore, the set $\mathcal{A} = \{S^t : 0 \leq t \leq |\tau|\}$ satisfies the assumptions of Proposition H.6, so z^* is the optimal value of (H.9) because it achieves the highest objective among $S \in \mathcal{A}$. Moreover, Step 4 mirrors Step 3, and terminates in the same iteration of t as when the optimal objective z^* was set, so that S^* is an optimal assortment.

Finally, if the items in S_0 are initially labeled in weakly decreasing order of v_j, then S^* is lexicographically the least among all optimal assortment due to the following observations:

1. The set S^t is lexicographically increasing in t: this is because the cardinality $|S^t|$ is weakly decreasing. Moreover, whenever there is a swap, the item i being swapped out has a smaller label than the item j being swapped in, since $v_i > v_j$ in the definition of the tuples in (H.18).

2. For any $\lambda \in \mathbb{R}$, the set $\mathcal{A}$ contains the lexicographically least element in $A(\lambda)$. This is true for $\lambda \in (-\infty, \lambda^1]$ by the sorting in Step 2, which breaks ties among coincident lines to favor lines with smaller indices. (By coincident lines, I mean that they have identical slopes and x-intercepts.) Moreover, any other assortment in $A(\lambda^1)$ either has a lower cardinality than S^0 or replaces certain items $i \in S^0$ with items j with higher indices. Observe now that among a group of coincident lines, their relative order is maintained throughout the algorithm, because at each $\lambda \in \Lambda := \{\lambda^1, \lambda^2, \cdots\}$, when there are multiple lines intersecting and a swap in S is to be made, the sorting of τ in Step 1 always swaps out the element i with the largest index and replaces it with the element with the smallest index. The above arguments imply that whenever $\lambda^t < \lambda^{t+1}$, the set S^t is lexicographically the least among all sets in $A(\lambda)$ for $\lambda \in (\lambda^t, \lambda^{t+1}]$.

3. Any optimal assortment S' to the socially optimal assortment planning problem (H.9) is either equal to or lexicographically larger than an optimal assortment in $\mathcal{A}$. This is because lemma H.1 used in the proof of Proposition H.6 implies that any optimal solution S' to (H.9) belongs to the set $A(\lambda')$ for some $\lambda' \in \mathbb{R}$, and all assortments in $A(\lambda')$ are also optimal. By point 2 above, $\mathcal{A}$ contains the minimal element of $A(\lambda)$, which is an optimal assortment that is equal to or lexicographically smaller than S'.

Since Step 4 terminates as soon as it hits an optimal λ^* and S^t is lexicographically increasing, the resultant assortment S^* is lexicographically the least among all those in $\mathcal{A}$ that achieves the objective z^*. By point 3 above, it must also be lexicographically the least among all optimal assortments. □

H.2.2. d-Level Nested Logit and Trivial Constraints The d-level nested logit utility distribution is a generalization of the MNL utility distribution in (15), in which the random terms ϵ_{ij} are allowed to be correlated across items. For the 2-level nested logit, the description of the correlation structure in (16) is due to Cardell (1997). For general d, the random terms ϵ_{ij}'s in the the utility equation (15) are positive correlated with one another, with the correlation structure following a rooted tree as illustrated in Figure H.2, with d being the maximum depth of the tree. Each item in $J := [n] \cup \{0\}$ is represented by a leaf node and each internal node i represents a nest of items that are positively correlated in their utilities.

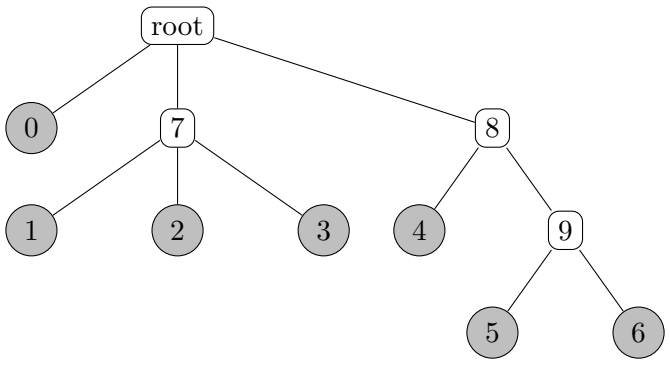


Figure H.2 Illustration of a 3-level nested logit utility distribution with 7 leaves and 4 internal nodes. Each leaf represents an item $j \in J = \{0, 1, \cdots, 6\}$ and each internal node $i \in N = \{7, 8, 9, root\}$ represents a nest, with dissimilarity parameter $\eta_i \in (0, 1]$, and $\eta_{root} = 1$. Note that the outside option 0 is always directly connected to the root.

A formal definition of a joint CDF that distribution of the ϵ_{ij}'s is deferred to (H.38) and (H.42) of Appendix H.2.4. For the analysis in this section, I do not work with the joint CDF directly but adopt the following mathematical simplifications due to Li et al. (2015).

The following notation is helpful for keeping track of the structure of the tree: Let N be the set of nests, which are represented by the internal nodes of the tree. Denote the root of the tree as *root*, and the other internal nodes by $\{n+1, n+2, \cdots, n+|N|-1\}$. Each non-root internal node $i \in N \backslash \{root\}$ has a dissimilarity parameter $\eta_i \in (0,1]$, where $\eta_i = 1$ corresponds to zero correlation within nest, and $\eta_i \to 0$ represents near perfect correlation. Define $\eta_{root} := 1$. Let $I = N \cup J$ be the set of all nodes. Each node $i \in I \backslash \{root\}$ has a unique parent node, denoted as $Parent(i)$. Each internal node $i \in N$ has a non-empty set of direct descendants, denoted as $Children(i)$. These are i's immediate neighbors when traversing away from the root. Denote the set of leaf nodes that are either direct or indirect descendants of an internal node $i \in N$ as $J_i \subseteq J$, and for each leaf node $i \in J$, define $J_i = \{i\}$. For each leaf node $i \in J$, define $Ancestors(i)$ to be the set of internal nodes when traversing from i in a direct path to the root. In the example in Figure H.2, $J_{root} = J = \{0,1,2,3,4,5,6\}$, $J_7 = \{1,2,3\}$, $J_8 = \{4,5,6\}$, and $J_9 = \{5,6\}$. Moreover, $Parent(8) = root$, $Parent(4) = 8$, $Children(8) = \{4,9\}$, $Children(9) = \{5,6\}$, $Ancestors(4) = \{root, 8\}$, and $Ancestors(5) = \{root, 8, 9\}$.

Following Li et al. (2015), assume that the outside option 0 is a direct descendant of the root, so $Parent(0) = root$. This is equivalent to assuming that the utility of the outside option is independent from the utilities of all other items. Given any assortment $S \subseteq J$ and any node $i \in I$, define the set $S_i := S \cap J_i$. For each leaf $j \in J$, define the attraction weight

$$v_j = \exp\left(\overline{u}_j \prod_{i \in Ancestors(j)} \eta_i^{-1}\right), \tag{H.20}$$

where $\overline{u}_j$ is the constant term in (15). For each node $i \in I$, define the following functions recursively:

$$V_i(S_i) = \begin{cases} \left(\sum_{j \in Children(i)} V_j(S_j)\right)^{\eta_i} & \text{if } i \in N, \\ v_i \mathbb{1}(i \in S_i) & \text{if } i \in J. \end{cases} \tag{H.21}$$

$$R_i(S_i) = \begin{cases} \frac{\sum_{j \in Children(i)} V_j(S_j) R_j(S_j)}{\sum_{j \in Children(i)} V_j(S_j)} & \text{if } i \in N \text{ and } |S_i| \geq 1, \\ 0 & \text{if } i \in N \text{ and } |S_i| = 0, \\ r_i \mathbb{1}(i \in S_i) & \text{if } i \in J. \end{cases} \tag{H.22}$$

For a d-level nested logit utility distribution, the socially optimal assortment planning problem (14) can be formulated in terms of the above functions as

$$\max_{S \in \Psi} \alpha \log(V_{root}(S)) + R_{root}(S). \tag{H.23}$$

This equivalent representation of (14) can be derived from (H.39), (H.40) and (H.42) in Appendix H.2.4.[28]

The efficient algorithm for solving (H.23) is based on the following two propositions. Proposition H.8 generalizes Proposition H.6 to the d-level nested logit utility distribution. It shows that one can solve the socially optimal assortment planning problem for any $\alpha \geq 0$ by considering a candidate set $\mathcal{A}$ of assortments, which

[28] As with the MNL utility distribution, the objective function in (H.23) differs from that in (14) by an additive constant of $\alpha \gamma_{Euler}$, where $\gamma_{Euler} = 0.5772...$ is Euler's constant.

is identical to the candidate set in Li et al. (2015) for the revenue maximizing case ($\alpha = 0$). Proposition H.9 provides a way of recursively constructing candidate sets at each internal node of the tree, which is similar to the approach in Section 5 of Li et al. (2015). However, the following analysis is based on new proof techniques arising from the Duality Lemma H.1.

PROPOSITION H.8 **(Candidate Set for Nested Logit)**. *For the d-level nested logit utility distribution and any constraint set Ψ that does not contain the empty set. Let $\mathcal{A} \subseteq \Psi$ be such that it contains an assortment from the set*

$$A(\lambda) = \arg\max_{S \in \Psi} \{V_{root}(S)(R_{root}(S) - \lambda)\}, \tag{H.24}$$

for every $\lambda \in \mathbb{R}$. Then $\mathcal{A}$ contains an optimal solution to the socially optimal assortment planning problem (H.23) *for any $\alpha \geq 0$. Furthermore, for a given $\alpha \geq 0$, the set of all socially optimal assortments is equal to $\bigcup_{\lambda^* \in \Lambda^*(\alpha)} A(\lambda^*)$ for some set $\Lambda^*(\alpha) \subseteq \mathbb{R}$.*

Proof of Proposition H.8 Define $x(S) = V_{root}(S)$ and $y(S) = V_{root}(S)R_{root}(S)$. The rest of the proof is identical to that of Proposition H.6 in Appendix H.2.1, and the desired results follow from Lemma H.1. □

PROPOSITION H.9 **(Recursive Structure of Candidate Sets)**. *For any internal node $i \in N$ of the d-level nested logit utility distribution, any constraint set $\Psi_i \subseteq 2^{J_i}$ that contains the empty set, and any given $\lambda \in \mathbb{R}$, if an assortment S is an optimal solution to the optimization problem*

$$\max_{S \in \Psi_i} \{V_i(S)[R_i(S) - \lambda]\}, \tag{H.25}$$

then there exists a $\lambda' \in \mathbb{R}$ such that S is also an optimal solution to the following optimization problem:

$$\max_{S \in \Psi_i} \{V_i^{1/\eta_i}(S)[R_i(S) - \lambda']\} \equiv \sum_{j \in Children(i)} \max_{S_j \in (\Psi_i \cap J_j)} \{V_j(S_j)[R_j(S_j) - \lambda']\}. \tag{H.26}$$

Moreover, for this value of λ', any other optimal solution S' to (H.26) *is also an optimal solution to* (H.25).

For the root node, the same statements trivially hold if the maximizations in (H.25) *and* (H.26) *are taken over an arbitrary constraint set $S \in \Psi$, because $\eta_{root} = 1$.*

Proof of Proposition H.9 Note that the equation in (H.26) is an identity by the definitions of $V_i(S)$ and $R_i(S)$ in (H.21) and (H.22). Moreover, the statement for the root node is trivially true as when $\eta_i = 1$ and $\lambda' = \lambda$, the equations (H.25) and (H.26) are identical.

Let $f(\lambda)$ and $f'(\lambda')$ be the optimal objective values to (H.25) and (H.26) respectively. Note that for any $\lambda \in \mathbb{R}$, $f(\lambda) \geq 0$ and $f'(\lambda) \geq 0$ since the empty set $S = \emptyset$ achieves the objective value of 0 in both (H.25) and (H.26). Moreover, $f(\lambda) > 0$ if and only if $f'(\lambda) > 0$ since either holds if and only if there exists an assortment $S \in \Psi_i$ with $V_i(S) > 0$ and $R_i(S) > \lambda$. As a result, if $f(\lambda) = 0$, then $f'(\lambda) = 0$, and an assortment S is an optimal solution to (H.25) if and only if $V_i(S) = 0$ or $R_i(S) = \lambda$, and the same is true for (H.26).

It remains to consider the case in which $f(\lambda) > 0$. Define $x(S) = V_i^{1/\eta_i}(S)$, $y(S) = V_i^{1/\eta_i}(S)R_i(S)$, $g(x, y) = yx^{\eta_i - 1} - \lambda x^{\eta_i}$, $R = (0, \infty) \times \mathbb{R}$, and $D = \{(x(S), y(S)) : S \in \Psi_i, x(S) > 0\} \subset R$. Since $f(\lambda) > 0$, the optimization problem (H.25) is equivalent to

$$\max_{(x,y) \in D} g(x, y). \tag{H.27}$$

In particular, the optimal objective value of (H.27) is equal to $f(\lambda)$ and an assortment S is an optimal solution to (H.25) if and only if $(x(S), y(S))$ is an optimal solution to (H.27). Moreover, the function $g(x, y)$ on the domain R is continuous and strictly increasing in y, and the lower contour set $\{(x, y) \in R : g(x, y) \leq f(\lambda)\}$ is convex. Therefore, we can apply Lemma H.1, which implies that any optimal solution $(x(S^*), y(S^*))$ to (H.27) is also an optimal solution to

$$f''(\lambda') := \max_{(x(S),y(S))\in D} \{y(S) - \lambda' x(S)\}, \tag{H.28}$$

for some $\lambda' \in \mathbb{R}$, which is guaranteed to exists and is such that

$$\inf_{(x,y)\in R} \{g(x, y) : y - \lambda' x = f''(\lambda')\} = f(\lambda). \tag{H.29}$$

Moreover, any other optimal solution $(x(S'), y(S'))$ to (H.28) at this λ' is also an optimal solution to (H.27). Note that (H.28) is identical to (H.26) except that it excludes assortments S with $V_i(S) = 0$.

To complete the proof, it suffices to show that $f''(\lambda') > 0$, as this would imply that the constraint in the definition of D that $V_i(S) > 0$ is extraneous, so the set of optimal solutions to (H.26) is the same as the set of optimal solutions to (H.28). Now, observe that for each point $(x, y) \in R$ with $y \leq \lambda' x$, we have $g(x, y) = yx^{\eta_i - 1} - \lambda x^{\eta_i} \leq (\lambda' - \lambda)x^{\eta_i}$. This implies that if $f''(\lambda') \leq 0$, then the infimum in (H.29) is less than or equal to zero, which contradicts $f(\lambda) > 0$. Therefore, $f''(\lambda') > 0$, as desired. □

Denote the trivial constraint set by $\Psi_0 := \{S \subseteq J : 0 \in S\}$. For each node $i \in I$, define $\Psi_0(i) := \{S \cap J_i : S \in \Psi_0\}$ and

$$f_i(\lambda) = \max_{S\in\Psi_0(i)} \{V_i(S)[R_i(S) - \lambda]\}. \tag{H.30}$$

The function is the upper envelope of a finite set of linear functions of λ, each of which corresponds to an assortment S, and has y-intercept $V_i(S)R_i(S)$ and slope $-V_i(S)$. Hence, $f_i(\lambda)$ is convex, piecewise linear and weakly decreasing. Note that $V_i(S) = 0$ implies that $|S \cap J_i| = 0$ by (H.21). Therefore, $f_{root}(\lambda)$ is strictly decreasing everywhere; for every other node $i \in I\backslash\{root\}$, $f_i(\lambda)$ is strictly decreasing in the range $\lambda \in (-\infty, \overline{\lambda}]$ for some $\overline{\lambda} < \infty$, and it is identically zero in the range $[\overline{\lambda}, \infty)$.

The main idea behind the efficient algorithm in this section is to compute a piecewise linear representation of $f_{root}(\cdot)$, from which one can construct a candidate set $\mathcal{A}$ satisfying the requirements of Proposition H.8 by including an assortment S corresponding to each each linear piece of $f_{root}(\cdot)$. Proposition H.8 states that it suffices to look within this candidate set $\mathcal{A}$ to identify a socially optimal assortment. Proposition H.9 suggests the following recursive procedure for computing a piecewise linear representation of the function $f_i(\cdot)$ for each node $i \in I$:

1. For each leaf node $i \in J$, the desired representation is trivial to compute: if $i \in [n]$, $f_i(\lambda) = \max\{v_i(r_i - \lambda), 0\}$; if $i = 0$, $f_i(\lambda) = -v_0\lambda$.

2. For each internal node $i \in N$, the desired representation can be computed from those of its children using Proposition H.9. Define the piecewise linear and convex function

$$f_i'(\lambda) := \sum_{j\in Children(i)} f_j(\lambda). \tag{H.31}$$

Each linear piece of f_i' also corresponds to an assortment S, which is the disjoint union of the assortments corresponding to the corresponding linear pieces from the children f_j's. Suppose the linear piece in f_i'

corresponding to assortment $S \in J_i \cap \Psi_0$ is $g'_S(\lambda) = a_S - b_S\lambda$, define the updated line $g_S(\lambda) = a_S b_S^{\eta_i - 1} - b_S^{\eta_i}\lambda$, and compute the upper envelope of the updated lines. By Proposition H.9, this new upper envelope is equal to $f_i(\lambda)$.

The above procedure for computing $f_{root}(\cdot)$ can be implemented in $O(dn \log n)$ time because a straightforward induction shows that the function $f_i(\lambda)$ has at most $|J_i|$ non-zero linear pieces. The sum in (H.31) can be computed by simply sorting all the breakpoints from the children f_j's and adding the corresponding lines between two adjacent breakpoints, so can be done in $O(|J_i| \log |J_i|)$ time. The re-computing of the upper envelope to transform f'_i into f_i can be done in $O(|J_i|)$ time using a standard algorithm from computational geometry (the dual of the monotone chain algorithm for computing the upper convex hull in $\mathbb{R}^2$). Hence, the recursive step for computing f_i from its children takes $O(|J_i| \log |J_i|)$ time, and adding this across all nodes yields the $O(dn \log n)$ bound.

Algorithm 4 is a concrete implementation of the above ideas. It represents each piecewise linear function $f_i(\lambda)$ as a doubly linked list L of $K \le |J_i|$ elements, where the kth element is a tuple $(\lambda_k, a_k, b_k, D_k)$. The list is always sorted so that $-\infty < \lambda_1 < \cdots < \lambda_K = \overline{\lambda}$. Define $\lambda_0 = -\infty$. The kth non-zero linear piece of $f_i(\lambda)$ is given by the line $a_k - b_k\lambda$ in the range $\lambda \in (\lambda_{k-1}, \lambda_k)$, and corresponds to the assortment $S_k := \bigcup_{k'=k}^{K} D_k$. See Figure H.3 for an illustration. The reason that we maintain the set difference $D_k = S_k \backslash S_{k+1}$ instead of the assortment S_k is that it makes the set operations in computing the sum H.31 more efficient. Moreover, one can show by induction that the recursive procedure above always results in nested sets. A generalization of this observation is rigorously derived in Lemma H.2 at the end of this section.

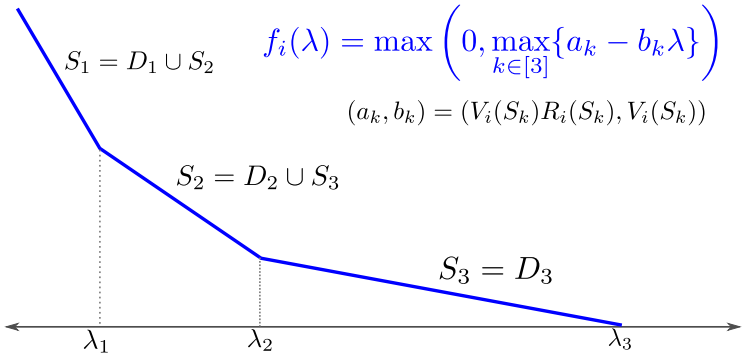


Figure H.3 Illustration of the meaning of the doubly linked list data structure $L = (\lambda_k, a_k, b_k, D_k)$ in Algorithm 4 for encoding a piecewise linear convex function $f_i(\lambda)$. In this example, i is a non-root internal node and has 3 non-zero linear pieces, so L has $K = 3$ elements.

The algorithm makes use of several standard data structures. It manipulates a doubly linked list L using the following notation:

- (): create an empty list.
- $L.insertEnd(x)$: insert the element x at the end of the list.
- $L.insertBeginning(x)$: insert the element x at the beginning of the list.
- $L.removeEnd()$: remove the last element of the list.

- $L[k]$: access the kth entry of the list.

The algorithm manipulates a priority queue Q using the following notation:

- $Q.top()$: obtain the smallest element in the queue.
- $Q.pop()$: obtain the smallest element and remove it from the queue.
- $Q.push(x)$: insert a new element x into the queue.
- $Q.pushAll(L)$: insert each element x of the list L into the queue.

The algorithm makes use of several functions, as summarized below:

- $generateCandidates(i)$: recursively compute a piecewise linear representation of the function $f_i(\lambda)$ for each node $i \in I \backslash \{0, root\}$. For the root node, the result is a piecewise linear representation of $f_{root}(\lambda) - f_0(\lambda) = f_{root}(\lambda) + v_0\lambda$.
- $firstDifference(L)$: replace each a_k and b_k in the doubly linked list $L = (\lambda_k, a_k, b_k, D_k)$ by $a_k - a_{k+1}$ and $b_k - b_{k+1}$. The purpose is to make the function sum in (H.31) easier to compute.
- $cumulativeSum(L')$: the inverse of $firstDifference(L)$.
- $reCompute(\eta_i, L)$: given $\eta_i \in (0, 1]$ and a representation of $f'(\lambda)$ as in (H.31), compute a representation of $f(\lambda)$ by updating each line and computing the upper envelope.
- $upperEnvelope(A, B)$: given two linked lists $A = (a_k)$ and $B = (b_k)$ encoding a set of lines $a_k - b_k\lambda$, with B sorted in strictly decreasing order, compute the upper envelope of the lines. By the duality of points and lines in $\mathbb{R}^2$, the correctness of this algorithm follows from that of the Monotone Chain algorithm (a.k.a. Andrew's algorithm) for computing the convex hull of the points (a_k, b_k).

The priority queue Q used below sorts tuples of the form $(\lambda_k, a_k, b_k, D_k)$ by the first component λ_k in weakly increasing order, and break ties arbitrarily.

Function generateCandidates(i)

Data: Node $i \in I \backslash \{0\}$, and all the parameters of the nested logit utility distribution: v, r, η, and $Children(\cdot)$.

Initialize $L \leftarrow ()$;
if $i \in [n]$ **then**
  $L.insertEnd((r_i, r_i v_i, v_i, \{i\}))$;
else if $i \in N$ **then**
  $Q \leftarrow$empty priority queue that orders tuples by the first component;
  for $j \in Children(i) \backslash \{0\}$ **do** $Q.pushAll(firstDifference(generateCandidates(j)))$;
  while $|Q| > 0$ **do**
    $(\lambda, a, b, D) \leftarrow Q.pop()$; $(\lambda', a', b', D') \leftarrow Q.top()$;
    while $\lambda = \lambda'$ **do**
      $a \leftarrow a + a'$; $b \leftarrow b + b'$; $D \leftarrow D \cup D'$;
      $Q.pop()$; $(\lambda', a', b', D') \leftarrow Q.top()$;
    end
    $L.insertEnd((\lambda, a, b, D))$;
  end
  if $i = root$ **then** $L.insertEnd((\infty, 0, v_0, \{0\}))$;
  $L \leftarrow cumulativeSum(L)$;
  $L \leftarrow reCompute(\eta_i, L)$;
end

Result: Doubly linked list L of cardinality $|L| \leq |J_i|$.

Function firstDifference(L)

Data: A doubly linked list L of the aforementioned format.
Initialize $L' \leftarrow ()$; $a' \leftarrow 0$; $b' \leftarrow 0$;
for $(\lambda, a, b, D) \in L$ *in reverse order of* λ **do**
  $L'.insertBeginning((\lambda, a - a', b - b', D))$;
  $a' \leftarrow a$, $b' \leftarrow b$;
end
Result: Doubly linked list L' of cardinality $|L'| = |L|$.

Function cumulativeSum(L')

Data: A doubly linked list L' of the aforementioned format.
Initialize $L \leftarrow ()$; $a' \leftarrow 0$; $b' \leftarrow 0$;
for $(\lambda, a, b, D) \in L$ *in reverse order of* λ **do**
  $a' \leftarrow a' + a$, $b' \leftarrow b' + b$;
  $L.insertBeginning((\lambda, a', b', D))$;
end
Result: Doubly linked list L of cardinality $|L| = |L'|$.

Function reCompute(η,L)

Data: Dissimilarity parameter $\eta \in (0, 1]$, and a doubly linked list L of the aforementioned format.
if $\eta_i = 1$ **then** $L^* \leftarrow L$;
else
  Initialize $A \leftarrow ()$; $B \leftarrow ()$; $\mathcal{D} \leftarrow ()$, $L^* \leftarrow ()$;
  for $(\lambda, a, b, D) \in L$ **do**
    $A.insertEnd(ab^{\eta_i})$; $B.insertEnd(b^{\eta})$; $\mathcal{D}.insertEnd(D)$;
  end
  $T \leftarrow upperEnvelope(A, B)$;
  $A.insertEnd(0)$; $B.insertEnd(0)$; $T.insertEnd(|L| + 1)$;
  for $k = 1, 2, \cdots, |T| - 1$ **do**
    $L^*.insertEnd(\frac{A[T[k]] - A[T[k+1]]}{B[T[k]] - B[T[k+1]]}, A[t], B[t], \bigcup_{t=T[k]}^{T[k+1]-1} \mathcal{D}[t])$;
  end
end
Result: Doubly linked list L^* of cardinality $|L^*| \leq |L|$.

Function upperEnvelope(A,B)

Data: Two lists of real numbers of equal cardinality $|A| = |B|$. The list B is sorted in strictly decreasing order.
Initialize $T \leftarrow ()$;
for $i = 1, 2, \cdots, |A|$ **do**
  while $|T| \geq 2$ *and* $\frac{A[T[-2]] - A[T[-1]]}{B[T[-2]] - B[T[-1]]} \geq \frac{A[T[-2]] - A[i]}{B[T[-2]] - B[i]}$ **do**
    $T.removeEnd()$;
  end
  $T.insertEnd(i)$;
end
Result: Doubly linked list T of indices, with cardinality $|T| \leq |A|$.

Algorithm 4: Socially optimal assortment planning under d-level nested logit utilities and trivial constraint

Data: Parameter $\alpha \geq 0$, and attraction weight $v_0 = \exp(\overline{u}_0)$.
Initialize $z^* \leftarrow -\infty$; $L \leftarrow generateCandidates(root)$;
for $(\lambda, a, b, D) \in L$ **do**
  $z \leftarrow \alpha \log(b) + b/a$;
  if $z > z^*$ **then** $z^* \leftarrow z$, $\lambda^* \leftarrow \lambda$;
end
$S^* \leftarrow \bigcup\{D : (\lambda, a, b, D) \in L \text{ and } \lambda \geq \lambda^*\}$;
Result: Optimal objective value z^* and optimal assortment S^*.

The following structural result is used in Proposition H.11 to state an additional property of the assortment S^* obtained by Algorithm 4. Moreover, it is the basis of the proof of Proposition 1 in Appendix H.3.

PROPOSITION H.10 **(Lattice Structure of Optimal Assortments)**. *For the d-level nested logit utility distribution and the trivial constraint set $\Psi_0 = \{S \subseteq J : 0 \in S\}$, the set of optimal solutions to the socially optimal assortment planning problem* (H.23) *forms a complete lattice: if S and S' are optimal assortments, then so are their union $S \cup S'$ and their intersection $S \cap S'$.*

Proof of Proposition H.10 The desired result follows from Proposition H.8 and the following lemma, since the set of optimal assortments can be written as $\bigcup_{\lambda^* \in \Lambda^*} A_{root}(\lambda^*)$ for some set $\Lambda^* \subseteq \mathbb{R}$, where $A_{root}(\cdot)$ is defined in (H.32) in the lemma. □

LEMMA H.2. *Let $\Psi_0 = \{S \subseteq J : 0 \in S\}$ be the trivial constraint set. For any node $i \in I$ of a d-level nested logit utility distribution, define the following set, which is parameterized by $\lambda \in \mathbb{R}$,*

$$A_i(\lambda) = \underset{S \in \Psi_0(i)}{\arg\max}\{V_i(S)[R_i(S) - \lambda]\}. \tag{H.32}$$

Suppose that $S \in A_i(\lambda)$ and $S' \in A_i(\lambda')$. If $\lambda' < \lambda$, then $S \subseteq S'$. If $\lambda' = \lambda$, then $S \cup S' \in A(\lambda)$ and $S \cap S' \in A(\lambda)$.

Proof of Lemma H.2 The proof is by induction. The above property is trivially true for the outside option node as $\Psi_0(0) = \{\{0\}\}$ has cardinality one. It is true for every leaf node $i \in [n]$ since the assortment $\{i\} \in A_i(\lambda)$ if and only if $\lambda \in (-\infty, r_i]$ and the empty assortment $\emptyset \in A_i(\lambda)$ if and only if $\lambda \in [r_i, \infty)$.

For an internal node $i \in N$, suppose that for all its children nodes $j \in Children(j)$, the above property is true for the parameterized set $A_j(\cdot)$, I show that it is also true for $A_i(\cdot)$. Suppose that $S \in A_i(\lambda)$ and $S' \in A_i(\lambda')$ with $\lambda' \leq \lambda$. Define the parameterized set

$$\tilde{A}_i(\lambda) := \left\{ \bigcup_{j \in Children(i)} T_j : T_j \in A_j(\lambda) \right\}. \tag{H.33}$$

By Proposition H.9, there exists $\tilde{\lambda}, \tilde{\lambda}' \in \mathbb{R}$ such that $S \in \tilde{A}_i(\tilde{\lambda}) \subseteq A_i(\lambda)$ and $S' \in \tilde{A}_i(\tilde{\lambda}') \subseteq A_i(\lambda')$. Define the function $f_i(\lambda)$ as in (H.30) and the function $f_i'(\lambda)$ as in (H.31). Both functions are convex and weakly decreasing. Moreover, $-V_i(S)$ is a subderivative of f_i at λ and $-V_i(S')$ is a subderivative of f_i at $\lambda' \leq \lambda$, so $V_i(S') \geq V_i(S)$. By the identity in (H.26), $-V_i^{1/\eta_i}(S)$ is a subderivative of f_i' at $\tilde{\lambda}$ and $-V_i^{1/\eta_i}(S')$ is a subderivative of f_i' at $\tilde{\lambda}'$, so $\tilde{\lambda}' \leq \tilde{\lambda}$ since $V_i(S')^{1/\eta_i} \geq V_i(S)^{1/\eta_i}$.

Suppose that $\tilde{\lambda}' < \tilde{\lambda}$, then $S \subseteq S'$ by (H.33) and the induction hypothesis. This makes S and S' satisfy the desired property regardless of whether $\lambda' < \lambda$ or $\lambda' = \lambda$.

Suppose that $\tilde{\lambda}' = \tilde{\lambda}$, then we have by the induction hypothesis and (H.33) that $S' \cup S, S' \cap S \in \tilde{A}_i(\tilde{\lambda})$, which is a subset of both $A_i(\lambda)$ and $A_i(\lambda')$. This already shows the desired result if $\lambda' = \lambda$. If $\lambda' < \lambda$, then observe that for any assortment $\tilde{S} \in \tilde{A}_i(\tilde{\lambda})$, $-V_i(\tilde{S})$ must be a subderivative to f_i at both λ and λ', so its value is pinned down:

$$-V_i(\tilde{S}) = \frac{f_i(\lambda) - f_i(\lambda')}{\lambda - \lambda'}. \tag{H.34}$$

The above argument implies that $V_i(S \cap S') = V_i(S \cup S')$. But $S \cap S' \subseteq S \cup S'$, so the definition of the function V_i in (H.21) implies that these sets must equal, so $S = S'$, which satisfies $S \subseteq S'$, as desired. □

PROPOSITION H.11. *For any d-level nested logit utility distribution and any $\alpha \geq 0$, Algorithm 4 solves the socially optimal assortment planning problem* (H.23) *under the trivial constraint set $\Psi = \{S \subseteq J : 0 \in S\}$, and is guaranteed to run in $O(dn \log n)$ time. In particular, it returns the correct objective value z^*, as well as the optimal assortment S^* with the largest cardinality, which is unique by Proposition H.10.*

Proof of Proposition H.11 The run time guarantee follows from the observation that the linked list L in $generateCandidates(i)$ has a maximum cardinality of $|J_i|$, so each priority queue push or pop takes $O(\log |J_i|)$ time. Moreover, the sets D are disjoint and the only operations needed on them is union, so they can be implemented efficiently using linked lists with each union $D \cup D'$ taking constant time. Furthermore, observe that the functions $firstDifference(L)$, $cumulativeSum(L)$, $reCompute(\eta, L)$ can all be implemented in $O(|L|)$ time and $upperEnvelope(A, B)$ can be implemented in $O(|A|)$ time. Hence, the total run time of algorithm 4 is upper bounded by

$$\sum_{i \in I} O(|J_i| \log |J_i|) \leq \sum_{i \in I} O(|J_i| \log n) = O(dn \log n). \tag{H.35}$$

The rest of the proposition follows from Proposition H.10 and the following property of the function $generateCandidates(i)$ for each node $i \in I \backslash \{0\}$: if the list returned is $L = (\lambda_k, a_k, b_k, D_k)$ and $S_k := \bigcup_{k' \geq k} D_k$, then L is a piecewise linear representation of $f_i(\lambda)$ (see Figure H.2) and S_k is the assortment in $A_i(\lambda)$ of largest cardinality for each $\lambda \in (\lambda_{k-1}, \lambda_k]$. (See (H.30) and (H.32) for definitions of f_i and A_i.) This property can be shown by induction: it is true for each leaf node $i \in [n]$ by construction. Suppose that it is true for all nodes in $Children(i)$, then it is true for node i because by the end of the line $L \leftarrow cumultativeSum(L)$ in $generateCandidates(i)$, the linked list L is a piecewise linear representation of $f_i'(\lambda)$ as defined in (H.31). By the induction hypothesis, the associated assortment S_k is the member of $\tilde{A}_i(\lambda)$ of highest cardinality for each $\lambda \in (\lambda_{k-1}, \lambda_k]$, where the set $\tilde{A}_i$ is defined in (H.33). The desired property for node i follows from Proposition H.9 and the correctness of the $upperEnvelope((a_k), (b_k))$ function for computing the upper envelope of the lines $(a_k - b_k\lambda)$, thus completing the induction. Moreover, the result of the $reCompute$ function at the end of $generateCandidates(i)$ is a representation of the upper envelope $f_i(\lambda)$ using the minimum possible number of lines, so that no three lines pass through the same point.

Finally, note that after obtaining the candidate assortments (S_k) from $generateCandidate(root)$, Algorithm 4 identifies the assortment $S^* = S_{k^*}$ that achieves the highest objective value (H.23), and in case of a tie, it returns the one with the smallest k^*, which is the one with the highest cardinality since $S_1 \supset S_2 \supset \cdots$. The optimality of S^* and its maximality among all optimal assortments follow from Proposition H.8. □

H.2.3. 2-Level Nested Logit Utilities and Cardinality Constraint within Nest When $d = 2$, define $N' = N \backslash \{root\}$, so that $[n] = \bigcup_{s \in N'} J_s$. For any set $\Psi \subseteq 2^J$ and any nest $s \in N'$, define $\Psi(s) = \{S \cap J_s : S \in \Psi\}$. Suppose that the constraint set Ψ is such that $\Psi \subseteq \Psi_0$ and the empty set $\emptyset \in \Psi(s)$ for each nest $s \in N'$, Propositions H.8 and H.9 imply that the socially optimal assortment planning problem (H.23) can be solved as follows:

1. For each nest $s \in N'$, construct a candidate set $\mathcal{A}_s$ that for each $\lambda \in \mathbb{R}$, contains an optimal solution to
$$\max_{S \subseteq \Psi(s)} \{\sum_{j \in S} v_j(r_j - \lambda)\}. \tag{H.36}$$
2. Compute a piecewise linear representation of the convex function
$$f_{root}(\lambda) = -v_0\lambda + \sum_{s \in N'} \max_{S \in \mathcal{A}_s} \{V_j(S)[R_j(S) - \lambda]\}. \tag{H.37}$$
Each linear piece corresponds to an assortment S which is the union of $\{0\}$ and an element of $\mathcal{A}_s$ for each nest $s \in N'$. By Proposition H.8, one of these assortments is guaranteed to be an optimal solution to (H.23) regardless of the parameter α.

For the cardinality within nest constraint set $\Psi = \{S \subseteq J : 0 \in S, |S \cap J_s| \leq k_s \text{ for each nest } s \in N'\}$, the candidate set $\mathcal{A}_s$ can be found using a modification of Algorithm 3, and a piecewise linear representation of (H.37) can be found using a modification of the $generateCandidates(root)$ function. A full implementation is given in Algorithm 5 below.

One difference from Algorithm 4 is that the sets S_k are no longer nested due to the cardinality constraints, so the algorithm encodes a piecewise linear representation by a list $L = (\lambda_k, a_k, b_k, D_k, E_k)$, where the associated assortment S_k for the kth piece satisfies $S_k = S_{k+1} \cup D_k \backslash E_k$. Moreover, define the functions $firstDifference'(L)$, $cumultativeSum'(L')$ and $reCompute'(\eta, L)$ to be the trivial modifications of the corresponding functions from Algorithm 4 in which whatever is done to the fourth component D of each tuple is also done to the fifth component E.[29]

Function solveNest(s)

Data: Nest $s \in N'$, along with the parameters $k_s \leq |J_s|$, $\eta_s \in (0, 1]$, and (v_j, r_j) for each $j \in J_s$.
$\tau \leftarrow$ the result of step 1 of Algorithm 3 with the set S_0 replaced by J_s and k replaced by k_s;
$o \leftarrow$ the result of step 2 of Algorithm 3 with the sets S_0 and $[n]$ replaced by J_s;
Initialize $L \leftarrow ()$; $S \leftarrow \{j \in [n] : o_j \leq k\}$; $a \leftarrow \sum_{j \in S} v_j r_j$; $b \leftarrow \sum_{j \in S} v_j$;
for $(\lambda, -i, j) \in \tau$ **do**
  if $j = 0$ *and* $i \in S$ **then**
    $L.insertEnd((\lambda, a, b, \{i\}, \{\}))$;
    $a \leftarrow a - v_i r_i$, $b \leftarrow b - v_i$;
  else if $j > 0$ *and* $o_i < o_j$ **then**
    Swap o_i and o_j;
    if $o_j = k$ *and* $i \in S$ **then**
      $L.insertEnd((\lambda, a, b, \{i\}, \{j\}))$;
      $a \leftarrow a - v_i r_i + v_j r_j$, $b \leftarrow b - v_i + v_j$;
    end
  end
end
$L \leftarrow reCompute'(\eta_s, L)$;
Result: Doubly linked list L of cardinality $|L| \leq \binom{|J_s|+1}{2}$.

[29] For example, in $reCompute'(\eta, L)$, one would create a list $\mathcal{E}$ analogous to $\mathcal{D}$, and the tuple inserted to L^* would have its last component being $\bigcup_{t=T[k]}^{T[k+1]-1} \mathcal{E}[t]$.

Algorithm 5: Socially optimal assortment planning under 2-level nested logit utilities and cardinality within nest constraint

Data: All the parameters of the 2-level nested logit utility distribution: v, r, η, J_s for each nest $s \in N'$. Parameter $\alpha \geq 0$.

Initialize $L \leftarrow 0$; $Q \leftarrow$ empty priority queue that orders tuples by the first component; $z^* \leftarrow -\infty$;
for $s \in N'$ **do** $Q.pushAll(firstDifference'(solveNest(s)))$;
while $|Q| > 0$ **do**
  $(\lambda, a, b, D, E) \leftarrow Q.pop()$; $(\lambda', a', b', D', E') \leftarrow Q.top()$;
  while $\lambda = \lambda'$ **do**
    $a \leftarrow a + a'$; $b \leftarrow b + b'$; $D \leftarrow D \cup D'$; $E \leftarrow E \cup E'$;
    $Q.pop()$; $(\lambda', a', b', D', E') \leftarrow Q.top()$;
  end
  $L.insertEnd((\lambda, a, b, D, E))$;
end
$L.insertEnd((\infty, 0, v_0, \{0\}, \{\}))$;
$L \leftarrow cumulativeSum'(L)$;
for $(\lambda, a, b, D, E) \in L$ **do**
  $z \leftarrow \alpha \log(b) + a/b$;
  if $z > z^*$ **then** $z^* \leftarrow z$, $\lambda^* \leftarrow \lambda$;
end
$D_{sum} \leftarrow \{\}$; $E_{sum} \leftarrow \{\}$;
for $(\lambda, a, b, D, E) \in L$ *if* $\lambda \geq \lambda^*$ **do** $D_{sum} \leftarrow D_{sum} \cup D$, $E_{sum} \leftarrow E_{sum} \cup E$;
$S^* \leftarrow D_{sum} \backslash E_{sum}$;

Result: Optimal objective value z^* and assortment S^*

PROPOSITION H.12. *Algorithm 5 can be implemented in $O(\sum_{j \in N'} |J_s|^2 \log n) = O(n^2 \log n)$ time and solves the socially optimal assortment planning problem* (H.23)*, with z^* being the optimal objective value and S^* an optimal assortment. Moreover, if within each nest $s \in N'$, the items are labelled in weakly decreasing order of v_j, with $v_j \geq v_{j'}$ if $j' < j$ for $j, j' \in J_s$, then the assortment S^* is lexicographically the least out of all optimal assortments. (The lexicographic ordering for assortments is defined immediately before Proposition H.7.)*

Proof of Proposition H.12 The time guarantee follows from the observation that the $solveNest(s)$ function can be implemented in $O(|J_s|^2 \log |J_s|)$ time as with Algorithm 3, and the list returned has at most as many elements as there are intersection points between the lines $v_j(r_j - \lambda)$ and the zero function, which is bounded by $\binom{|J_s|+1}{2}$. Hence, the total number of linear segments to f_{root} is $O(\sum_{s \in N'} |J_s|^2)$, which implies that each push or pop of the priority queue in Algorithm 5 can be implemented in $O(\log n)$ time, and the whole algorithm takes $O(\sum_{s \in N'} |J_s|^2 \log n)$ time.

The rest of the proposition follows from Propositions H.8 and H.9 and the following observations:

1. By the correctness of Algorithm 3, the function $solveNest(s)$ for each nest $s \in N'$ returns a piecewise linear representation of the function $f_s(\lambda) = \max_{S \subseteq J_s, |S| \leq k_s} \{V_i(S)[R_i(S) - \lambda]\}$. Moreover, when the items in each nest are labeled in weakly decreasing order of v_j, then as in the proof of Proposition H.7, the assortment S_k corresponding to the kth piece is the lexicographically smallest assortment in $A_s(\lambda) = \arg\max_{S \subseteq J_s, |S| \leq k_s} \{V_i(S)[R_i(S) - \lambda]\}$. Moreover, the sequence (S_k) is lexicographically increasing.
2. The list L in Algorithm 5 after the line $L \leftarrow cumulativeSum'(L)$ is a piecewise linear representation of the function $f_{root}(\lambda)$ from (H.37). Moreover, if (S_k) is the sequence of assortments corresponding to this

piecewise linear representation, the assortments are lexicographically increasing in k by construction, and the S^* returned by Algorithm 5 is the first S_k that achieves the optimal objective value.

□

H.2.4. Generalization for GEV Utility Distributions The MNL and d-level Nested logit utility distribution are both special cases of Generalized Extreme Value (GEV) utility distributions, which are first studied in McFadden (1978). The utility of an agent i for item $j \in J$ is parameterized as $u_{ij} = \overline{u}_j + \epsilon_{ij}$ as in the MNL utility distribution, except that the random term ϵ_{ij}'s are allowed to be richly correlated across items, so that the $(n+1)$-dimensional vector ϵ_i is distributed according to a joint CDF $F : \mathbb{R}^{n+1} \to R$ with the following form:

$$F(x) = \mathbb{P}(\{\epsilon_{ij} \le x_j \text{ for all } j \in J) = \exp(-G(e^{-x_0}, e^{-x_1}, \cdots, e^{-x_n})), \tag{H.38}$$

where $G : \mathbb{R}^{n+1} \to \mathbb{R}$ is called a GEV generating function and satisfies the following properties:

1. Non-negativity on the positive orthant: $G(x) \ge 0$ for each non-negative $|J|$-dimensional vector $x \ge 0$. Moreover, $G(\mathbb{1}_j) > 0$ for each unit vector $\mathbb{1}_j$, which has 1 in component j and 0 elsewhere.
2. Homogeneous of degree one: for all $\alpha \ge 0$, $G(\alpha x) = \alpha G(x)$.
3. Differentiable, with non-positive even and non-negative odd mixed partial derivatives: for any $k \ge 1$ and any distinct indices $j_1, j_2, \cdots, j_k \in J$, the partial derivative $(-1)^k \frac{\partial G}{\partial_{j_1} \cdots \partial_{j_k}}(x) \le 0$ for all $x \in \mathbb{R}^{n+1}$.

McFadden (1978) shows that (H.38) is a valid CDF, and the expected utilities and choice probabilities for the corresponding utility distribution are given by,

$$U(S) = \log(G(\nu(S)) + \gamma_{\text{Euler}}, \tag{H.39}$$

$$P(j, S) = \frac{\nu_j(S) \partial_j G(\nu(S))}{G(\nu(S))}, \tag{H.40}$$

$$\text{where } \nu_j(S) = \begin{cases} \exp(\overline{u}_j) & \text{if } j \in S, \\ 0 & \text{otherwise,} \end{cases} \tag{H.41}$$

and $\nu(S)$ is a $(n+1)$-dimensional vector with the jth component equal to $\nu_j(S)$. $\gamma_{Euler} = 0.5772...$ is Euler's constant.

For the MNL utility distribution, the generating function is $G(x) = \sum_{j=0}^{n} x_j$. For the 2-level nested logit utility distribution with nests $\{J_s\}$, it is $G(x) = x_0 + \sum_s \left(\sum_{j \in J_s} x_j^{1/\eta_s} \right)^{\eta_s}$. For the d-level nested logit utility distribution,[30] $G(x) = G_{root}(x)$, where for each node of the tree $i \in I := N \cup J$, the function $G_i(x) : \mathbb{R}^{n+1} \to \mathbb{R}$ is defined recursively as follows

$$G_i(x) = \begin{cases} \left(\sum_{j \in Children(i)} G_j(x) \right)^{\eta_i} & \text{if } i \in N, \\ x_i^{1/\prod_{j \in Ancestors(i)} \eta_j} & \text{if } i \in J. \end{cases} \tag{H.42}$$

The following result generalizes Propositions H.6 and H.8 to arbitrary GEV utility distributions, and characterizes the optimality conditions for the parameter λ.

[30] See Appendix H.2.2 for explanation of the tree notation in the d-level nested logit utility distribution.

PROPOSITION H.13. *For any GEV utility distribution, any constraint set Ψ that does not contain the empty set, and any $\alpha \geq 0$, an assortment S is an optimal solution to the socially optimal assortment planning problem*

$$\max_{S \in \Psi} \alpha U(S) + R(S), \tag{H.43}$$

if and only if S is an optimal solution to the optimization problem

$$f(\lambda) := \max_{S \in \Psi} \{G(v(S))(R(S) - \lambda)\}, \tag{H.44}$$

for some λ belonging to the set

$$\Lambda^*_\alpha := \begin{cases} \arg\max_{\lambda' \in \mathbb{R}} \{f(\lambda)\exp(\lambda/\alpha)\} & \text{if } \alpha > 0, \\ \{\lambda' : f(\lambda') = 0\} & \text{if } \alpha = 0. \end{cases} \tag{H.45}$$

Moreover, the set Λ^* can be rewritten as $\{R(S) - \alpha : S$ is an optimal solution to (H.43)$\}$.

Proof of Proposition H.13 As in the proof of Proposition H.6 and H.8, define $x(S) = G(v(S))$, $y(S) = G(v(S))R(S)$, $D = \{(x(S), y(S)), S \in \Psi\}$, $g(x, y) = \alpha \log(x) + y/x$, and $R = (0, \infty) \times \mathbb{R}$. Note that the function $g(x, y)$ on the open convex domain R is quasi-convex, continuous, strictly increasing in y. The socially optimal assortment planning problem (H.43) can be written as $\max_{(x,y) \in D} g(x, y)$.

Define the function

$$h(\lambda, f) = \inf_{x \in (0,\infty)} \left\{ \alpha \log(x) + \frac{f + \lambda x}{x} \right\}. \tag{H.46}$$

Note that when $\alpha > 0$, $h(\lambda, f) = \alpha \log(f/\alpha) + \alpha + \lambda$. When $\alpha = 0$,

$$h(\lambda, f) = \begin{cases} \lambda & \text{if } f \geq 0, \\ -\infty & \text{if } f < 0. \end{cases} \tag{H.47}$$

The desired result follows from Lemma H.1 and the observation that the function $f(\lambda)$ in (H.44) is equal to $\max_{(x,y) \in D}\{y - \lambda x\}$ and is strictly decreasing in λ. The last statement on the equivalent representation of Λ^* when $\alpha > 0$ follows from the first order condition for the optimization in λ for the function $\log(f(\lambda)) + \lambda/\alpha$, which is the logarithm of (H.45). When $\alpha = 0$, the desired result follows from the observation that $f(\lambda) = 0$ is equivalent to $R(S^*) = \lambda$, where S^* is an optimal solution to (H.43). □

H.2.5. Markov Chain Based Choice Model and Trivial Constraint The Markov chain based choice model is proposed by Blanchet et al. (2016) as a tractable approximation to the mixed MNL utility distribution, which McFadden et al. (2000) show can approximate any random utility model to any degree accuracy. (The mixed MNL utility distribution itself is intractable for assortment optimization even with $\alpha = 0$, as shown by Bront et al. (2009) and Rusmevichientong et al. (2014).) In this section, I modify the Markov chain based choice model to add a measure of preference intensity, and adapt the algorithm of Feldman and Topaloglu (2017) for solving the revenue-maximizing assortment planning problem ($\alpha = 0$) to the socially optimal case ($\alpha \geq 0$).

The modified Markov chain based utility model is as follows: each agent has an initial utility $\overline{v}$, which is an arbitrary constant. Let $J = [n] \cup \{0\}$. Define a Markov chain with $n + 1$ states, in which each state corresponds to an item $j \in J$. There are three sets of parameters in the utility distribution:

1. an *arrival rate* $a_j \geq 0$ for each state $j \in J$, with their sum being equal to one;

2. a *transition probability* $\rho_{kj} \geq 0$ from each state $k \in [n]$ to each state $j \in J$, with $\sum_{j \in J} \rho_{kj} = 1$;

3. a *disappointment cost* $c_{kj} \geq 0$ for each transition from $k \in [n]$ to $j \in J$.

Note that there are no transitions out of the outside option 0.

Given any assortment $S \subseteq J$ containing the outside option 0, consider the following stochastic process: agents arrive at each state according to the arrival rates. Whenever they arrive at a state $j \in S$, they leave the system. Otherwise, at each time step, they follow the transition probabilities to their next state, and continue in the system until they arrive at one of the states in S. The expected utilities and choice probabilities are defined as follows:

$$V(S) = \overline{v} - \mathbb{E}[\text{total disappointment cost incurred before leaving}] \tag{H.48}$$

$$P(j,S) = \begin{cases} \mathbb{P}(\text{The state when they leave the system is } j) & \text{if } j \in S, \\ 0 & \text{otherwise.} \end{cases} \tag{H.49}$$

Proposition H.14. *Consider a Markov chain based utility model with arrival probabilities a, transition probabilities ρ, and disappointment costs c. For any $\alpha \geq 0$ and any revenue vector r, let (x^*, z^*) be an optimal basic solution to the linear program:*

$$\underset{x,z}{\textit{Maximize:}} \qquad \alpha \left(\overline{v} - \sum_{k \in [n], j \in J} c_{kj} \rho_{kj} z_k \right) + \sum_{j \in [n]} r_j x_j \tag{H.50}$$

$$\textit{subject to:} \qquad x_j + z_j = a_j + \sum_{k \in [n]} \rho_{kj} z_k \qquad \textit{for each } j \in [n]. \tag{H.51}$$

$$x_j, z_j \geq 0 \qquad \textit{for each } j \in [n]. \tag{H.52}$$

Then the assortment $S^* = \{j \in [n] : x_k^* > 0\} \cup \{0\}$ is an optimal solution to the socially optimal assortment planning problem (14) *under the trivial constraint set $\Psi_0 = \{S \subseteq J : 0 \in S\}$, and the optimal LP objective value* (H.50) *is the optimal objective value of* (14).

Proof of Proposition H.14. As in Section 1 of Feldman and Topaloglu (2017), for any assortment $S \subseteq J$, let $R(j,S)$ be the steady state rate of agents leaving state j. We have that

$$P(j,S) = \begin{cases} a_j + \sum_{k \in [n]} \rho_{kj} R(k,S) & \text{if } j \in S, \\ 0 & \text{otherwise.} \end{cases}$$

$$R(j,S) = \begin{cases} a_j + \sum_{k \in [n]} \rho_{kj} R(k,S) & \text{if } j \in [n] \backslash S \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, for any assortment S, setting $x_k = P(k,S)$ and $z_k = R(k,S)$ yields a feasible solution to the LP. The social welfare of this assortment according to the objective function (14) is exactly the LP objective (H.50), so the maximum social welfare of any assortment $S \in \Psi_0$ is upper bounded by the optimal LP objective value. Moreover, by Lemma 1 of Feldman and Topaloglu (2017), the polyhedron described by the constraints (H.51) and (H.52) is such that for any vertex and any $j \in J$, either $x_j = 0$ or $z_j = 0$. This implies that if S^* are as defined in the theorem, then $P(j,S^*)$ and $R(j,S^*)$ are exactly given by x_j and z_j, so we can attain the optimal LP objective value with assortment S^*. □

H.3. Proof of Proposition 1: Setting in which RSD is Optimal

Let $\Psi_0 = \{S \subseteq J : 0 \in S\}$ be the trivial constraint set, and Λ be the total mass of agents. Since there is only one agent segment, the LP in Section 3.1 can be simplified as follows, with the subscript t omitted everywhere, and the vector $p^S \in [0,1]^n$ defined to be equal to the choice probability $P(j,S)$ in its jth component.

$$\text{Maximize}_y \quad \sum_{S\in\Psi_0} U(S) y_S \tag{H.53}$$

$$\text{s.t.} \quad \sum_{S\in\Psi_0} y_S = \Lambda \tag{H.54}$$

$$\sum_{S\in\Psi_0} p_j^S y_S \le c_j \qquad \text{for each item } j \in [n] \tag{H.55}$$

$$y_S \ge 0 \tag{H.56}$$

The remainder of the proof shows that if $c_j > 0$ for every $j \in [n]$, then there exists an optimal solution y^* to the above LP with

$$y_J^* = \min(1, \min_{j\in[n]}\{c_j/p_j^J\}). \tag{H.57}$$

Once this is established, then the optimality of RSD follows from induction on the number of items with positive capacity, as by the induction hypothesis, RSD would be optimal when (Λ, c) is replaced by $(\Lambda - y_J^*, c - p^J y_J^*)$, in which case the number of items with positive capacity strictly decreases, as $\{j : c_j - p_j^J y_J^* > 0\} \subsetneq \{j : c_j > 0\}$.

Let ϕ be the dual variable for constraint (H.54) and γ_j for constraint (H.55). The dual LP is

$$\text{Minimize}_{\phi,\gamma} \quad \Lambda\phi + c\cdot\gamma \tag{H.58}$$

$$\text{s.t.} \quad \phi + p^S\cdot\gamma \ge U(S) \qquad \text{for every } S \in \Psi_0. \tag{H.59}$$

$$\gamma \ge 0 \tag{H.60}$$

For a given optimal dual solution (ϕ,γ), let $\mathcal{A}$ be the set of budget sets S for which the constraint (H.59) is tight. I show that $J \in \mathcal{A}$. This is because the optimality of (ϕ,γ) implies that $\phi = \max_{S\in\Psi_0}\{U(S) - p^S\cdot\gamma\}$, which means that ϕ is the optimal objective value of the socially optimal assortment planning problem with constraint set Ψ_0, parameter $\alpha = 1$ and revenue $r_j = -\gamma_j^*$. Moreover, $\mathcal{A}$ is the set of optimal assortments. Now, for any j such that $\gamma_j = 0$, it must be that $j \in S$ for every optimal assortment $S \in \mathcal{A}$, as including the item in any assortment would increase the expected utility without incurring any penalty. If $\gamma_j > 0$, then by complementary slackness, (H.55) is tight at the optimal primal solution y^*, so there exists a $S \ni j$ with $y_S^* > 0$, which implies that $S \in \mathcal{A}$. Thus, $J = \bigcup_{S\in\mathcal{A}} S$. By Proposition H.10 in Appendix H.2.2, $\mathcal{A}$ is a complete lattice, which implies that $J \in \mathcal{A}$.

Define $f(\Lambda, c)$ to be the optimal objective of the dual LP with coefficients (Λ, c) in the objective function. Let x^* be the right hand side (RHS) of (H.57). For any $x < x^*$, we have

$$f(\Lambda, c) = U(J)x + f(\Lambda - x, c - p^J x). \tag{H.61}$$

This is because if (ϕ,γ) is an optimal dual solution for parameters (Λ, c) and (ϕ',γ') is an optimal dual solution for parameters $(\Lambda - x, c - p^J x)$ with $x < x^*$, then the argument in the above paragraph shows that

$$\phi + p^J\cdot\gamma = U(J) = \phi' + p^J\cdot\gamma', \tag{H.62}$$

so the left hand side (LHS) of (H.61) is less than or equal to the RHS by the optimality of (ϕ, γ), and the RHS is less than or equal to the LHS by the optimality of (ϕ', γ'). Since $f(\Lambda, c)$ is a finite and concave function for any non-negative inputs, we have by continuity that (H.61) holds also for $x = x^*$, which is what we needed to prove, as the desired y^* can be constructed from any optimal solution of the primal LP with constraint bounds $(\Lambda - x^*, c - p^J x)$ and setting $y^*_J = x^*$. □

H.4. Proof of Proposition 2: Analysis of Example 1

I analytically solve the LP in (9)-(13) for the market M in Example 1. Define the sets $S_0 = \{0\}, S_1 = \{0,1\}, S_2 = \{0,2\}, S_3 = \{0,1,2\}$. Without loss of generality, it suffices to consider feasible solutions $y \in Y^M$ satisfying the following symmetry condition: $y_{1S_3} = y_{2S_3}$, $y_{1S_0} = y_{2S_0}$, and $y_{1S_1} = y_{2S_2}$, $y_{1S_2} = y_{2S_1}$. This is because if y is not symmetric, then define y' so that the two neighborhoods and two schools have labels switched, then $y'' = (y + y')/2$ is symmetric. Moreover, y'' satisfies all the constraints of the LP, and yields the same objective value as y.

By symmetry, we can rewrite the LP in the following simpler way, in which the decision variable z_k corresponds to y_{1S_k} for $k \in \{1,2,3\}$.

$$\text{Maximize} \qquad \sum_{k=1}^{3} [U(S_k) - U(S_0)] z_k \tag{H.63}$$

$$\text{s.t.} \qquad [P(1,S_3) + P(2,S_3)] z_3 + P(2,S_2) z_2 + P(1,S_1) z_1 \leq c \tag{H.64}$$

$$z_1 + z_2 + z_3 \leq 1 \tag{H.65}$$

$$z_1, z_2, z_3 \geq 0 \tag{H.66}$$

In the above, $c := c_1 = c_2$ is the common capacity of the two schools, and U and P are defined as in (7) and (8) for neighborhood 1 except that the subscript $t = 1$ is omitted here for simplicity. The original objective in (9) is equal to (H.63) plus $U(S_0)$ then multiplied by 2. By the assumptions in Example 1, $P(1,S_1) > 0$, $P(2,S_2) > 0$, and $U(S_k) - U(S_0) > 0$ for each $k \in \{1,2,3\}$. Moreover, $c \leq P(1,S_1)$, and $[U(S_1) - U(S_0)]/P(1,S_1) > [U(S_2) - U(S_0)]/P(1,S_2)$.

Observe that in any optimal solution z, it must be that $z_2 = 0$: if it were not so, then $z' := ([P(2,S_2) z_2 + P(1,S_1) z_1]/P(1,S_1), 0, z_3)$ would be a strictly a better feasible solution. This is because z' satisfies (H.64) by construction, and satisfies (H.65) since

$$z_1 + z_3 \leq z_1 + \frac{P(1,S_3) + P(2,S_3)}{P(1,S_1)} z_3 \leq \frac{c}{P(1,S_1)} \leq 1, \tag{H.67}$$

where the first inequality follows from $P(1,S_3) + P(2,S_3) = 1 - P(0,S_3) \geq 1 - P(0,S_1) = P(1,S_1)$, the second inequality from (H.64), and the third inequality from the assumption that capacities are scarce. The above argument also shows that under the assumptions of Proposition 2, $z_2 = 0$ implies (H.65).

After setting $z_2 = 0$ and observing that constraint (H.65) is always fulfilled, the feasible region is now a triangle with three vertices. The vertex $(z_1, z_2, z_3) = (0,0,0)$ is never optimal since its objective value is zero while the other two are strictly positive. The remaining two vertices are $(z_1, z_2, z_3) = (c/P(1,S_1), 0, 0)$ and $(0, 0, c/[P(1,S_3) + P(2,S_3)])$. The first corresponds to the neighborhood assignment plan, with objective value equal to c times the left hand side of (17). The second corresponds to the open enrollment plan, with objective value equal to c times the right hand side of (17). □

H.5. Proof of Proposition 3: Analysis of Example 2

The proof is based on analytically solving the LP in (9)-(13), and making use of the following lemma, whose proof is given at the end of this section.

LEMMA H.3. *Let $X_1, X_2, \cdots$ be i.i.d. random variables with CDF F and $Y_k := \max_{1\le i\le k}\{X_i\}$. Let Z be a random variable with continuous CDF H, such that $\mathbb{P}(X_1 \ge Z) > 0$. For each $k \in \{1, 2, \cdots\}$, define*

$$\phi_k := \mathbb{E}[Y_k - Z | Y_k \ge Z]. \tag{H.68}$$

a) Suppose that H has a light left-tail, then ϕ_k is weakly increasing in k.

b) Suppose that H has a heavy left-tail and its upper support is weakly larger than that of F: for any $x \in \mathbb{R}$, $H(x) = 1$ implies that $F(x) = 1$. Then ϕ_k is weakly decreasing in k.

Without loss of generality, normalize the outside option distribution H so that it has mean zero, and label the schools so that capacities are weakly decreasing, $c_1 \ge c_2 \ge \cdots \ge c_n$. Define the function ϕ_k as in the statement of Lemma H.3, based on the utility distribution F and the outside option distribution H. Define $p_k = \mathbb{P}(Y_k \ge Z)$, where Y_k and Z are as in the statement of Lemma H.3, with $0 < p_1 \le p_2 \le \cdots \le p_n$. Define $\Psi = \{S \subseteq [n] : |S| > 0\}$. For any budget set probability matrix $y \in Y^M$ for the market M in Example 2, define a corresponding $(2^n - 1)$-dimensional vector x such that for every $S \in \Psi$, $x_S := \sum_{t\in[m]} y_{t(S\cup\{0\})}$. Define $\Lambda = \sum_{t\in[m]} \lambda_t$. By symmetry, the LP in (9)-(13) can be equivalently formulated in terms of x as follows:

$$\text{Maximize} \quad \sum_{S\in\Psi} \phi_{|S|} p_{|S|} x_S \tag{H.69}$$

$$\text{s.t.} \quad \sum_{S\in\Psi} \frac{1}{|S|} p_{|S|} \mathbb{1}(j \in S) x_S \le c_j \qquad \text{for each } j \in [n]. \tag{H.70}$$

$$\sum_{S\in\Psi} x_S \le \Lambda \tag{H.71}$$

$$x \ge 0 \tag{H.72}$$

The neighborhood assignment plan corresponds to

$$x_S = \begin{cases} c_j/p_1 & \text{if } S = \{j\}, \\ 0 & \text{otherwise,} \end{cases} \tag{H.73}$$

with objective value $W^{neighbor} := C\phi_1$, where $C := \sum_{j=1}^n c_j$. For convenience, define $c_{n+1} = 0$. The open enrollment plan (RSD) corresponds to

$$x_S = \begin{cases} k(c_k - c_{k+1})/p_k & \text{if } S = \{1, 2, \cdots, k\} \text{ for } 1 \le k \le n, \\ 0 & \text{otherwise,} \end{cases} \tag{H.74}$$

with objective value $W^{open} := \sum_{k=1}^n k(c_k - c_{k+1})\phi_k$. It suffices to show that under the assumptions of Example 2, $W^{neighbor}$ is the optimal objective value when H has a heavy left-tail and W^{open} is optimal when it has a light left-tail.

First, observe that the constraint (H.71) is extraneous given the assumption in Example 2 that capacities are scarce: $C/\Lambda \le p_1$. This is because if we sum (H.70) for all $j \in [n]$ and use the fact that $p_k \ge p_1$ for all $k \ge 1$, we get

$$p_1 \sum_{S\in\Psi} x_S \le \sum_{S\in\Psi} p_{|S|} x_S \le C, \tag{H.75}$$

so $\sum_{S\in\Psi} x_S \le C/p_1 \le \Lambda$ is implied.

Let γ_j be the shadow price of the constraint (H.70). The dual to the above LP without the extraneous constraint (H.71) is as follows.

$$\text{Minimize} \qquad \sum_{j=1}^{n} c_j\gamma_j \tag{H.76}$$

$$\text{s.t.} \qquad \phi_{|S|} \le \frac{1}{|S|}\sum_{j\in S}\gamma_j \qquad \text{for each } S\in\Psi. \tag{H.77}$$

$$\gamma \ge 0 \tag{H.78}$$

Observe that if γ is a feasible solution to the above, and if the components are permuted, then the resultant vector γ' is also a feasible solution. Therefore, the optimal γ must be in reverse order from the c_j's, so $\gamma_1 \le \gamma_2 \le \cdots \le \gamma_n$. Given this ordering, it suffices to consider the constraints (H.77) in which $S = \{1, 2, \cdots, k\}$ for some $k \in [n]$. Hence, the above dual LP is equivalent to the following:

$$\text{Minimize} \qquad \sum_{j=1}^{n} c_j\gamma_j \tag{H.79}$$

$$\text{s.t.} \qquad \phi_k \le \frac{1}{k}\sum_{j=1}^{k}\gamma_j \qquad \text{for each } k\in[n]. \tag{H.80}$$

$$\gamma_j \le \gamma_{j+1} \qquad \text{for each } j\in[n-1]. \tag{H.81}$$

$$\gamma \ge 0 \tag{H.82}$$

By Lemma H.3, when H has a heavy left-tail, we have $\phi_1 \ge \phi_2 \ge \cdots \ge \phi_n$, in which case an optimal solution to the above LP will have all constraints (H.81) tight and γ_1 as small as possible, so $\gamma_j = \phi_1$ for all $j \in [n]$. The objective value is equal to $W^{neighbor}$. When H has a light left-tail, we have $\phi_1 \le \phi_2 \le \cdots \le \phi_n$, in which case an optimal solution will have every constraint (H.80) tight, and the constraints (H.81) will be extraneous. Solving, we get $\gamma_j = j\phi_j - (j-1)\phi_{j-1}$, where we define $\phi_0 = 0$ for convenience. The objective value is equal to W^{open}. □

The proof of Lemma H.3 makes use of the following basic property of light and heavy tailed distributions.

LEMMA H.4. *Let H be the CDF of a continuous distribution and let $\underline{x} := \sup\{x : H(x) = 0\}$ and $\overline{x} := \sup\{x : H(x) = 1\}$ be the lower and upper bounds to its support. (Note that $\underline{x}$ may be $-\infty$ and $\overline{x}$ may be ∞.) Define the function $\zeta : (\underline{x}, \infty) \to \mathbb{R}$,*

$$\zeta(x) := \mathbb{E}_{Z\sim H}[x - Z | x \ge Z]. \tag{H.83}$$

a) If H has a light left-tail, then $\zeta(x)$ is weakly increasing on $(\underline{x}, \infty)$.

b) If H has a heavy left-tail, then $\zeta(x)$ is weakly decreasing on $(\underline{x}, \overline{x})$ and strictly increasing on $(\overline{x}, \infty)$.

Proof of Lemma H.4 If H has a light left-tail, then $\log(H(x))$ is weakly increasing and concave for $x \in (\underline{x}, \infty)$. Therefore, for any $x > \underline{x}$, $y \in [\underline{x}, x]$ and $\delta > 0$,

$$\frac{H(y)}{H(y+\delta)} \le \frac{H(x)}{H(x+\delta)}, \tag{H.84}$$

Thus,

$$\int_{\underline{x}}^{x} H(y)\,dy = \int_{\underline{x}-\delta}^{x} H(y)\,dy = \int_{\underline{x}-\delta}^{x} \frac{H(y)}{H(y+\delta)} H(y+\delta)\,dy \le \frac{H(x)}{H(x+\delta)} \int_{\underline{x}}^{x+\delta} H(y)\,dy, \tag{H.85}$$

which rearranges to

$$\zeta(x) = \frac{\int_{\underline{x}}^{x} H(y)\,dy}{H(x)} \le \frac{\int_{\underline{x}+\delta}^{x} H(y)\,dy}{H(x+\delta)} = \zeta(x+\delta). \tag{H.86}$$

On the other hand, if H has a heavy left-tail, then $\log(H(x))$ is weakly increasing and convex for $x \in (\underline{x}, \overline{x})$ and identically zero for $x \in (\overline{x}, \infty)$. For $\delta > 0$, $x \in (\underline{x}, \overline{x} - \delta]$, $y \in [\underline{x}, x]$, (H.84) holds with the inequality reversed, and similarly (H.85) and (H.86) hold with $\le$ changed to $\ge$. So $\zeta(x)$ is weakly decreasing for $x \in (\underline{x}, \overline{x})$. When $x \ge \overline{x}$, $\zeta(x) = (x - \overline{x}) + \zeta(\overline{x})$, so $\zeta(x)$ is strictly increasing in x. □

Proof of Lemma H.3 Consider the conditional distribution of Y_k given $Y_k \ge Z$. It has CDF,

$$\Gamma_k(y) := \frac{\int_{-\infty}^{y} H(x)\,dF^k(x)}{\int_{-\infty}^{\infty} H(x)\,dF^k(x)}. \tag{H.87}$$

The denominator is positive since $\mathbb{P}(Y_k \ge Z) \ge P(X_1 \ge Z) > 0$. Let $\tilde{Y}_k$ be a random variable with the above CDF, then $\phi_k = \mathbb{E}[\zeta(\tilde{Y}_k)]$, where the function ζ is defined as in Lemma H.4. Moreover, note that the upper support of $\tilde{Y}_k$ is no more than that of F: $F(x) = 1$ implies that $\Gamma_k(x) = 1$. Therefore, it suffices to show that $\tilde{Y}_{k'}$ first order stochastically dominates $\tilde{Y}_k$ if $k' > k \ge 1$: $\Gamma_{k'}(y) \le \Gamma_k(y)$ for all $y \in \mathbb{R}$.

If $\Gamma_k(y) = 1$, then there's nothing to show. If $\Gamma_k(y) = 0$, then $\Gamma_{k'}(y) = 0$ because the numerator

$$\int_{-\infty}^{y} H(x)\,dF^{k'}(x) = \frac{k'}{k} \int_{-\infty}^{y} H(x) F^{k'-k}(x)\,dF^k(x) \le \frac{k'}{k} \int_{-\infty}^{y} H(x)\,dF^k(x) = 0. \tag{H.88}$$

If $0 < \Gamma_k(y) < 1$, then we have

$$\frac{\int_{-\infty}^{y} H(x)\,dF^{k'}(x)}{\int_{-\infty}^{y} H(x)\,dF^k(x)} = \frac{\frac{k'}{k}\int_{-\infty}^{y} H(x) F^{k'-k}(x)\,dF^k(x)}{\int_{-\infty}^{y} H(x)\,dF^k(x)} \le \frac{\int_{y}^{\infty} H(x)\,dF^{k'}(x)}{\int_{y}^{\infty} H(x)\,dF^k(x)}. \tag{H.89}$$

The above inequality holds because the fraction to left of the "$\le$" sign is equal to $\mathbb{E}[s(\tilde{Y}_k)|\tilde{Y}_k \le y]$ and the fraction to the right is equal to $\mathbb{E}[s(\tilde{Y}_k)|\tilde{Y}_k \ge y]$, where the function $s(x) = k' F^{k'-k}(x)/k$ is weakly increasing in x. Note that the denominator in the left most term of(H.89) is strictly positive, as well the denominator of the right most term. Moreover, the sum of the numerator of the left most term and that of the right most term is $\mathbb{P}(\tilde{Y}_{k'} \ge Z) > 0$, so the numerator of the right most term is also strictly positive. Therefore, (H.89) can be rearranged to $\Gamma_{k'}(y) \le \Gamma_k(y)$, as desired. □

H.6. Negative Externality of Choice

While the examples in Section 5 illustrate the tradeoff between the benefit and cost of allowing more choices, the following one-item example more clearly illustrates the negative externality of choice.

PROPOSITION H.15 **(Optimal Mechanism with Homogeneous Items)**. *Suppose that there are m segments but only one type of item ($n = 1$), then a priority-based allocation mechanism that maximizes utilitarian welfare is as follows. Prioritize agents based on their segment t, in decreasing order of the segment's average marginal value ϕ_t for obtaining the item conditioning on desiring it over the outside option: if $(\alpha, v) \sim F_t$ represents the agent's utility for the outside option and the item, then*

$$\phi_t := \begin{cases} 0 & \text{if } \mathbb{P}(v > \alpha) = 0, \\ \mathbb{E}[v - \alpha | v > \alpha] & \text{otherwise.} \end{cases} \tag{H.90}$$

In the above result, ϕ_t is the expected welfare loss for withholding one unit of supply to interested agents from segment t, and is a precise measure of the negative externality incurred when someone's decision to choose the item displaces an agent from segment t.

Proof of Proposition H.15 Since there is only one item, we can simplify the LP in (9)-(13) by defining the change of variables: $z_t = \lambda_t y_{tS_1}/P_t(1,S_1)$, where $S_1 = \{0,1\}$. Define $S_0 = \{0\}$. Note that $\phi_t = (U_t(S_1) - U_t(S_0))/P_t(1,S_1)$. Let c be the capacity of the one item, and let $K = \sum_{t\in[m]} U(S_0)$ be the utilitarian welfare when everyone is assigned the outside option. The simplified LP is as follows:

$$\text{Maximize:} \qquad K + \sum_{t\in[m]} \phi_t z_t \tag{H.91}$$

$$\text{s.t.} \qquad \sum_{t\in[m]} z_t \le c \tag{H.92}$$

$$0 \le z_t \le \lambda_t \qquad \text{for each segment } t \in [m]. \tag{H.93}$$

This is exactly the LP for the fractional knapsack problem, and the solution is to initialize $z_t = 0$ for all segments, then update z_t segment by segment in decreasing order of ϕ_t: for each segment t, set $z_t = \lambda_t$ if capacity allows, and otherwise set z_t to be the maximum value that doesn't violate the capacity constraint $\sum_t z_t \le c$. This solution corresponds exactly to the mechanism described in Proposition H.15. □

H.7. Intuitive Interpretation of Optimized Budget Sets

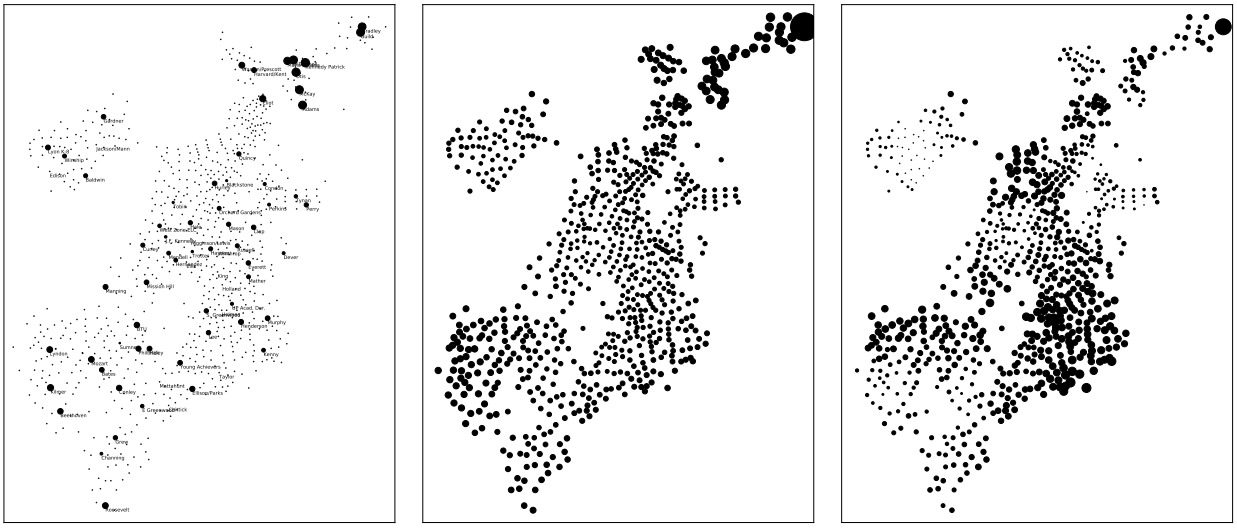
The optimal budget sets from the LP (18)-(27) have the following intuitive structure: each school j is associated with a certain $cost_j \ge 0$, which is proportional to the shadow price of the capacity constraint (24), and is higher for schools that are more popular but have lower capacities. Each neighborhood t is given a certain endowment $e_t \ge 0$ of points, as well as an allowance $k_t \ge 0$ of schools outside of its walk-zone. The parameter $cost_j$ is deterministic, whereas e_t and k_t may be random for each student from the neighborhood. The following proposition summarizes the dependence of the budget sets on the parameters $cost_j$, e_t and k_t.

PROPOSITION H.16 **(Structure of Optimal Budget Sets)**. *Let y^* be an optimal solution to the LP (18)-(27) for the MNL utility distribution, and ξ_1, ξ_2 and ξ_3 be the corresponding shadow prices for the constraints (25)-(27). Suppose ξ_1 is strictly positive, and at least one of ξ_2 or ξ_3 is strictly positive. For any neighborhood t and budget set S such that $y^*_{tS} > 0$, there exists parameters $e_t, k_t \ge 0$ such that S can be expressed as the union of*

- *the default school j_t;*
- *all schools j within the one-mile walk-zone with $cost_j < e_t$;*
- *The top k_t other schools with the highest score σ_{tj}, defined as*

$$\sigma_{tj} = \bar{u}_{tj} + \beta \log(e_t - cost_j - d_{tj}). \tag{H.94}$$

where $\bar{u}_{tj}$ is the average utility of students from neighborhood t for school j, β is the scale parameter of the MNL utility distribution, and d_{tj} is the distance in miles from neighborhood t to school j. If fewer than k_t schools have a positive sum within the logarithm, then include only the ones that do.



(a) School costs (b) Expected endowments (c) Expected # of busing choices

Figure H.4 These plots illustrate the parameters that encode the optimal budget sets for Boston. Subfigure (a) plots the schools as circles, with the size of the circle for school j being proportional to $cost_j$, which is proportional to the shadow cost of its capacity constraint (24). Subfigure (b) plots the distribution of expected endowments across neighborhoods, with each circle representing a neighborhood and the size of the circle proportional to $\mathbb{E}[e_t]$. A larger endowment gives a neighborhood higher access to over-demanded and faraway schools. Subfigure (c) plots the distribution of the expected number of additional school options in the budget set outside of the walk-zone. Each circle represents a neighborhood and the size of the circle is proportional to $\mathbb{E}[k_t]$, with the exception of the largest circle at the top right corner, which should be larger than what is shown but is capped for visibility of nearby neighborhoods.

- *In the knife-edge case in which $cost_j = e_t$ for some school j within the walk-zone, then S may or may not include the school j.*[31]

A high endowment e_t allows a neighborhood to access over-demanded schools that are faraway. For schools within the one-mile walk-zone, busing is not required, so such schools are included as long as they are not too over-demanded, $cost_j < e_t$. For schools outside of the walk-zone, the parameters e_t and k_t limit the size of the coverage area as well as the number of busing options. The score σ_{tj} defined in (H.94) favors schools that the neighborhood likes on average, and penalizes schools that are faraway or highly desired by other neighborhoods, thus optimally balancing the expected utility of students from neighborhood t with the negative externalities they impose on others when they occupy seats at a school. Figure H.4 plots the geographic distribution of the values $cost_j$, $\mathbb{E}[e_t]$ and $\mathbb{E}[k_t]$ for the Boston dataset.

Proof of Proposition H.16 Consider the equivalent formulation of the LP in (18)-(27) in which constraints (19)-(21) are substituted into the rest, and all summations of S are taken over Ψ_t as defined in (32). By writing down the dual of this LP and applying complementary slackness, we get that if y^* is an optimal solution, then $y^*_{tS} > 0$ implies that S is an optimal solution to the assortment planning problem (31) with ν_t,

[31] This knife-edge case is theoretically possible but does not occur in the Boston dataset.

γ_t, ξ_1, ξ_2, and ξ_3 being the shadow prices of (23)-(27) at y^*. By Proposition H.13, S is an optimal solution to (31) only if it is an optimal solution to the following optimization for some $k^* \in \mathbb{Z}$ and $x \in \mathbb{R}$,

$$\max_{S: j_t \in S, |S \setminus S_t^{walk} \setminus \{j_t\}| \leq k^*} \left\{ \sum_{j \in S} (r_j - x) e^{\overline{u}_{tj}/\beta} \right\}, \tag{H.95}$$

where r_j is defined in (34). Due to the linear structure of (H.95), any optimal assortment S contains j_t, as well as every school $j \in S_t^{walk}$ for which $r_j > x$, and may or may not contain schools for which $r_j < x$. Moreover, it contains no school $j \in [n] \setminus \{j_t\}$ for which $r_j < x$. Out of schools within the set $L_t = [n] \setminus S_t^{walk} \setminus \{j_t\}$ for which $r_j - x > 0$, it contains the k^* schools with the highest and positive $(r_j - x)e^{\overline{u}_{tj}/\beta}$. Now, when $|\{j : r_j > x, j \in L_t\}| < k^*$, it is possible that an optimal solution to (H.95) includes a school $j \in L_t$ for which $r_j = x$. However, this will not arise as an optimal solution to (31) as Proposition H.13 implies that removing the schools with $r_j = x$ yields an assortment with a smaller cardinality $|S \setminus S_t^{walk}|$ yielding the same value for the first two components of (31) but incurring a strictly smaller penalty in the third component, since the coefficient ζ is strictly positive if one of ξ_2 and ξ_3 is strictly positive by (35). The rest of Proposition H.16 follows from the formula for r_j in (34) and by defining $cost_j = n\gamma_j/\xi_1$, $e_t = -nx/(\lambda_t \xi_1)$ and $k_t = k^*$. □