OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0030-364X | EISSN 1526-5463 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Improving Community Cohesion in School Choice via Correlated-Lottery Implementation

Itai Ashlagi*
MIT Sloan School of Management (iashlagi@mit.edu)

Peng Shi
MIT Operations Research Center (pengshi@mit.edu)

In school choice, children submit a preference ranking over schools to a centralized assignment algorithm, which takes into account schools' priorities over children and uses randomization to break ties. One criticism of existing school choice mechanisms is that they tend to disperse communities so children do not go to school with others from their neighborhood. We suggest to improve community cohesion by implementing a *correlated lottery* in a given school choice mechanism: we find a convex combination of deterministic assignments that maintains the original assignment probabilities, thus maintaining choice, but improving community cohesion.

To analyze the gain in cohesion for a wide class of mechanisms, we first prove the following characterization which may be of independent interest: any mechanism which, in the large market limit, is non-atomic, Bayesian incentive compatible, symmetric and efficient within each priority class, is a "lottery-plus-cutoff" mechanism. This means that the large market limit can be described as follows: given the distribution of preferences, every student receives an identically distributed lottery number, every school sets a lottery cutoff for each priority class, and a student is assigned her most preferred school for which she meets the cutoff. This generalizes Liu and Pycia (2012) to allow arbitrary priorities. Using this, we derive analytic expressions for maximum cohesion under a large market approximation. We show that the benefit of lottery-correlation is greater when students' preferences are more correlated.

In practice, although the correlated-lottery implementation problem is NP-hard, we present a heuristic that does well. We apply this to real data from Boston elementary school choice 2012 and find that we can increase cohesion by 79% for K1 and 37% for K2 new families. Greater cohesion gain is possible (tripling cohesion for K1 and doubling for K2) if we reduce the choice menus on top of applying lottery-correlation.

1. Introduction

In various school choice mechanisms, students submit a ranked list of schools they would like to attend, and schools have priorities over students; a centralized assignment algorithm, which may randomize to break

* Ashlagi acknowledges the research support of the National Science Foundation grant SES-1254768.

ties, determines the assignment. Such mechanisms are used to assign children to public schools in many metropolitan areas in the US, including Boston, New York, New Orleans, San Francisco, and Chicago.

One drawback of existing school choice systems is that children from the same community end up going to many different schools, thus weakening community ties. Ebbert and Ulmanu (2011) document 19 children on one street in Boston going to 15 different schools, as an example of the community dispersion due to a choice lottery. Community dispersion also raises transportation burdens: Boston Public Schools spent $80 million in 2012 on busing students, which represents almost 10% of its total budget (Sutherland (2012)). In his January 2012 State of the City address, Boston's mayor Menino said,

> "Pick any street. A dozen children probably attend a dozen different schools. Parents might not know each other; children might not play together. They can't carpool, or study for the same tests. We won't have the schools our kids deserve until we build school communities that serve them well." Menino (2012)

One important facet of building communities is to have children go to school with others from their community, so the families get to know one another through common activities. Partly motivated by this, some have advocated abandoning school choice altogether and switching to a neighborhood-based system, in which kids predominately go to the closest school. But choice adds value by allowing families to find an option that fits their individual tastes, and many parents vocally defend their right to choose. Is it possible to improve community cohesion without sacrificing choice?

The main insight of this paper is that much of the community dispersion is artificial, caused purely by the allocation algorithm using independently drawn lottery numbers. If we had used a "correlated-lottery," then we could implement the same assignment probabilities but improve cohesion. Consider the following example: There are 2 schools with 2 seats each, and 8 students from 2 communities. Students A, B, C, D are from community I, and E, F, G, H are from community II. Students A, B, E, F prefer school 1, and C, D, G, H prefer school 2. Suppose for simplicity that the choice mechanism is Random Serial Dictatorship– students are ordered uniformly randomly and they pick schools sequentially according to this order. It is straightforward to work out the assignment probabilities (see Table 1).

Conditional on being assigned, what is a student's chance of being assigned with someone else from the same community? Because there are 2 communities and everything is symmetric, we expect this to be roughly $\frac{1}{2}$. Working out the details, we get that the answer is in fact $\frac{13}{35} \approx 37\%$.

However, the random assignment in Table 2 generates the same assignment probabilities for everyone, while always keeping communities together.

From each individual's perspective, the second random assignment is "equivalent" to the first, in the sense that the individual has the same probabilities of being assigned to each school as before. As a result, every individual's expected distance to school, expected "academic quality" of assigned school, probabilities of

			Assignment Probability	
Community	Student	Preference	School 1	School 2
I	A	$1 \succ 2$	$61/140$	$9/140$
	B	$1 \succ 2$	$61/140$	$9/140$
	C	$2 \succ 1$	$9/140$	$61/140$
	D	$2 \succ 1$	$9/140$	$61/140$
II	E	$1 \succ 2$	$61/140$	$9/140$
	F	$1 \succ 2$	$61/140$	$9/140$
	G	$2 \succ 1$	$9/140$	$61/140$
	H	$2 \succ 1$	$9/140$	$61/140$

Table 1 Assignment probabilities from example. For example, student A from community I prefers school 1 over 2, and in the randomized assignment she is assigned to school 1 with probability $61/140$ and school 2 with probability $9/140$. With remaining probability $\frac{1}{2}$, she is not assigned (or assigned to an outside option).

		Assignment			
Community	Student	i	ii	iii	iv
I	A	1	2		
	B	1	2		
	C			1	2
	D			1	2
II	E			2	1
	F			2	1
	G	2	1		
	H	2	1		
	Probability	$61/140$	$9/140$	$61/140$	$9/140$

Table 2 Correlated lottery implementation of the assignment probabilities from Table 1. This randomizes over 4 assignments, each represented by a column. For example, assignment i is chosen with probability $61/140$, and assigns A and B to school 1, E and F to 2, and leaves the rest unassigned. Note that no matter what the realized assignment is, communities stay together as much as space allows.

getting into some set of schools, are the same as before. In some sense, the second lottery implementation achieves gains in cohesion "for free."

Formally, we define the community cohesion of a lottery as the expected number of same-community school peers a student can expect to see. We propose the following approach to increase community cohesion in any randomized allocation mechanism: estimate the assignment probabilities for every student to every school in the current mechanism, and implement the same assignment probabilities in a "community-correlated" lottery. In other words, we seek a convex combination of deterministic assignments that matches the original assignment probabilities but that maximizes cohesion. We term this optimization problem *correlated-lottery implementation.*

In this paper, we address the following questions:

1. For the most prevalent mechanisms used in practice, by how much can we hope to improve community cohesion using a correlated lottery? In what settings can we expect the most improvement?

2. How to solve the correlated-lottery optimization problem in practice?

3. For Boston (where community cohesion in school choice has been the focus of much debate), how much can correlated lottery improve community cohesion? How would such a method interact with other possible reforms?

To help us address the first question, we prove a large market characterization of all mechanisms that satisfy the following four properties:(1) non-atomicity (a single individual has negligible effect on assignment probabilities of others), (2) asymptotic Bayesian incentive compatibility (given distribution of students' preferences, taking the limit as the market size goes to infinity, students reporting truthfully forms a Bayes-Nash equilibrium), (3) symmetry within each priority class (students with same priorities to every school and same submitted preferences receive the same assignment probabilities), and (4) asymptotic efficiency within each priority class (no trading cycles within each priority class). These properties are generally satisfied by commonly implemented mechanisms, such as Deferred Acceptance (DA) or Top Trading Cycles (TTC) with randomized tie-breakers, when we take a suitable limit as market size is scaled up.[1] We show that any mechanism that satisfies these properties in the large market can be interpreted as a "lottery-plus-cutoff" mechanism, which means that students are divided into priority classes and are each given an identically distributed lottery number; given the distribution of preferences, schools set a lottery cutoff for each priority class; students are assigned to their most preferred school for which they meet the lottery cutoff. This generalizes a result by Liu and Pycia (2012) to allow for priorities.

Using this characterization, we derive clean expressions for cohesion with independent lotteries and optimally correlated lotteries. In a large market framework, we show that baseline cohesion (using independent lotteries) is equal to the sum of a measure of variation in school size and between-community variation in assignment probabilities. Under an additional assumption that each community has the same priority to all schools, we show that maximum cohesion (using optimally correlated lotteries) is the sum of baseline cohesion and the average variance of a certain "demand function" of communities for schools. This improvement term can be interpreted as a measure of preference correlation, with greater improvement under higher uncertainty of lottery or under higher within-community preference correlation. Under a random utility model, we show that if there are no priorities or between-community variation, cohesion gain from lottery correlation increases when preferences are more correlated.

We address the second question by showing that the problem is NP-hard to solve optimally and introduce a heuristic that performs reasonably well in practice. The underlying optimization problem is related to the Quadratic Assignment Problem, which is in general notoriously intractable (see Burkard et al. (1998)), but there is more structure in our case, which is exploited in our heuristic. We also derive an upper-bound to test the optimality gap of our heuristic.

We address the third question by applying our heuristic to real data from Boston elementary school choice, simulating what would have happened if we implemented lottery correlation in 2012 Round 1 assignment. The main grades under consideration are kindergarten 1 (K1) and kindergarten 2 (K2). Defining

each community to be a .5 mile by .5 mile square, we show that our method improves community cohesion by 79% for grade K1 and 37% for K2. Conditional on the student traveling outside their walk-zone (1-mile radius), cohesion improves by 140% for K1 and 64% for K2.

We also compare our approach to reforms discussed by a mayor-appointed city committee during the 2012-2013 Boston school choice reform. The main reforms discussed were to increase the walk-zone percentage and and to reduce the choice menu. As of 2012, school programs were split into two halves, one half that prioritizes students living within 1 mile (walk-zone), and one half that does not have this priority. For most programs, the walk-zone half represented 50% of seats. By increasing this percentage, policy makers can induce a closer-to-home assignments, thus increasing cohesion. However, we showed that even if we had made the percentage 100%, the gain in community cohesion would not be as much as if we had kept the walk-zone percentage unchanged but used correlated lottery. Furthermore, while increasing the walk-zone percentage would not increase the number of same-community-peers for students traveling out of their walk-zone, correlated lottery would.

The other reform discussed by the city committee was to reduce the choice menus of students. During this process many plans for choice menus were proposed, some involving dividing the city into more assignment zones, and others involving a customized menu that depended on students' addresses. Using simulated choices from a discrete choice model fitted with real data, we show that the cohesion gain from lottery correlation is comparable to sizable reductions in choice menus. More interestingly, the two strategies amplify one another. For example, consider the choice menu called "Home Based A." (This was the choice menu eventually chosen by the city committee.) If we were to apply this choice menu reform alone, we would improve cohesion by 46% for K1 and 30% for K2. If we were to apply correlated-lottery alone, the cohesion gains are 88% for K1 and 39% for K2. So correlated-lottery achieves more gain. However, if we were to apply both reform at the same time, cohesion would more than triple for K1 and more than double for K2. So the number of neighbors students can expect to see at their assigned school would dramatically increase.

We also analyze the geographic distribution of cohesion gains due to correlated lottery, and show that while performing correlated lottery alone yields uneven gains for K2, this would be largely mitigated if we simultaneously reduced the choice menu. Furthermore, we show that lottery correlation has minimal impact on racial or social-economic diversity. These analyses are in the Electronic Companion.

1.1. Related work

While there is much existing literature on school choice (see Abdulkadiroğlu and Sönmez (2003b), Abdulkadiroğlu et al. (2009, 2006), Pathak (2011)), most of the literature focuses on individual students' assignments and ignore the *correlations* between different students' assignments. One reason for this is that complementarities in matching is difficult to analyze theoretically.

The idea that different random assignments can represent the same assignment probabilities has appeared before in the literature. Abdulkadiroğlu and Sönmez (2003a), note that such random assignments may differ in their ex-post efficiency. However, in their setup there is no guideline to decide between such random assignments. In our setup, there is the added performance measure of community cohesion. Piantadosi et al. (2007) and Asadpour and Saberi (2010) seek for an "optimal" convex combination of assignments that yields the highest entropy while maintaining the same assignment probabilities. In their setting, they maximizes a concave function, which is computationally tractable. However, in our setting, we seek to maximize a convex function, and the optimization is NP-hard. This difficulty cannot be avoided by an alternative definition of cohesion because cohesion is inherently convex, as it corresponds to having greater variation between assignments (either have children from a neighborhood mostly go to one school or mostly go to another).

Our approach of defining a mechanism by implementing the marginal assignment probabilities is similar to Budish et al. (2013). They study a more general framework and address the issues of group-specific quotas, ex-ante efficiency, ex-post fairness, and implementability of lotteries under general constraints. However, their techniques do not handle issues involving complementarities, such as community cohesion, and our work expands on the applicability of their framework in this domain. Another work that studies the decomposition of assignment probabilities into deterministic assignments to achieve certain properties is Pycia and Ünver (2012).

Randomization has been much studied in school choice mechanisms. The currently most adopted mechanisms break ties in school's preferences for students using independently generated lottery numbers, and apply the deferred acceptance (DA) algorithm or the top trading cycles (TTC) algorithm. (See Abdulkadiroğlu and Sönmez (2003b).) Abdulkadiroğlu et al. (2009) study whether to use a single tie-breaker for all schools or different tie-breakers for different schools, and their simulations using New York City data shows that single tie-breaking is better. Pathak and Sethuraman (2011) show that in the absence of school priorities both tie-breaking methods are equivalent. Erdil and Ergin (2008) illustrate the potential ex-ante inefficiencies from running an ex-post efficient mechanism after random tie-breaking, and propose a method to deal with such inefficiencies, but the resultant mechanism is no longer incentive compatible. Azevedo and Leshno (2010) show that this proposed improvement may yield Nash equilibria in which the outcome is Pareto-dominated by the original mechanism. Che and Kojima (2011) show that in the large market, without school priorities, the deferred acceptance algorithm with a randomized tie-breaker is equivalent to the ordinal efficient probabilistic serial mechanism, which Liu and Pycia (2012) show is equivalent in the large market to any asymptotically efficient, symmetric, and asymptotically strategyproof ordinal allocation mechanisms. These works suggest that there may be little room for improvement over the status quo in terms of individual students' welfare, given requirements of strategyproofness and fairness. Our work is different from these because we focus on community cohesion, which can be seen as "orthogonal" to students'

individual assignment probabilities to schools. We show that there is in fact much room for improvement in this direction, while also maintaining most of the good properties of the current mechanisms.

In terms of implementing other social objectives in school choice, there has been previous work in "controlled school choice," most of which focus on achieving diversity (see e.g., Ehlers (2010), Ehlers et al. (2011), Echenique and Yenmez (2012) and Kominers and Sönmez (2012)).

A recent paper that studies neighborhood interactions in school choice is Weiwei (2013), which uses secondary school choice data from New York City to show that students tend to choose similar schools as their immediate neighbors. This provides empirical support that students value going to school with their neighbors.

Outside of school choice, there has been other studies of assignment externalities. On the theoretical side, general settings of matching with externalities usually yields negative results. (For example, in many-to-one matching of workers to firms, if preferences can be over colleagues then the game theoretic core can be empty. See Echenique and Yenmez (2007) for an example. (Klaus and Klijn 2005) show a similar impossibility result even when joint preferences are between only two agents.) One empirical study that has similar flavor to ours is Mariagiovann et al. (2012), in which they consider the assignment of faculty members to offices in a US professional school, and study how institutional and social ties between faculty affect their choices and the final assignment. They quantify the effects of these network externalities and assess the matching protocol from a welfare perspective.

2. Model

There are n students to be assigned to m schools. The set of individual students is I and the set of schools is S. Each school $s \in S$ has capacity q_s. Without loss of generality, we assume that all students must be assigned. This is because we can model unassignment if needed by including a dummy school $s_\emptyset$ with infinite capacity to denote "unassigned."

The students are partitioned into k disjoint communities:

$$I = I_1 \cup I_2 \cdots \cup I_k.$$

A community-membership function $c : I \to \{1, \cdots k\}$ maps each student to the index of the community she belongs to. (For clarity of exposition, we refer to students using the feminine gender and the social planner using the male gender.) As a slight abuse of notation, we may also index communities using c.

An *assignment* a is a mapping that takes students to schools, and we require that no school capacity is violated. Formally, $a : I \to S$, $|a^{-1}(s)| \leq q_s$, where $a^{-1}(s) = \{i : a(i) = s\}$. A random assignment x is a random variable whose realizations are assignments. Denote the set of all random assignments X.

Slightly abusing notation, we sometimes represent a as an indicator matrix, in which a_{is} is 1 if and only if student i is assigned to school s. In this notation x_{is} becomes a binary random variable for whether i is

assigned to s. And $p = E[x]$ becomes the matrix of assignment probabilities. (p_{is} is probability student i is assigned to school s.)

For each student i and school s, define the *school-specific utility* of i for s to be u_{is}. Abuse notation slightly and define $u_i(a) := u_{ia(i)}$, which is student i's school-specific utility under assignment a. Define $v_i(a)$ to be i's number of *same-community peers* under assignment a. This is the number of other students assigned to the same school and from the same community, and can be written as $v_i(a) = |(a^{-1}(a(i)) \cap I_{c(i)}) \backslash \{i\}|$.

Assume that student i's preference over random assignments is induced lexicographically by the ordered pair

$$(E[u_i(x)], E[v_i(x)]).$$

So students in our model care foremost what school they are assigned to, and within a given school they prefer to be assigned with more peers from their community. We call this preference structure *weakly community-preferring*. We assume that for each student i, u_{is} is different for different s. Hence, the u_{is}'s induce for student i a complete strict ordering $\succ_i^*$ over schools, in which her more preferred schools are ranked better. We call $\succ_i^*$ the *true preference ranking* of student i.

One intuitive measure of community cohesion under assignment a is simply the average number of same-community peers assigned to the same school. We define $f(a)$, the *cohesion* of assignment a, as follows:

$$f(a) = \frac{1}{n}\sum_i v_i(a) = -1 + \frac{1}{n}\sum_i \sum_{j \in I_{c(i)}} \mathbb{1}\{a(i) = a(j)\},$$

where $\mathbb{1}\{a(i) = a(j)\}$ is the indicator for students i and j being assigned together. Note that $nf(a)/2$ is simply the total number of pairs of students from the same community who are assigned to the same school. The cohesion of a random assignment x is defined as $E[f(x)]$. (In the example in the introduction, the expected cohesion of the first lottery is .37 and of the second is 1.)

An *assignment mechanism* $\mathcal{M}$ is a function that takes as input a strict ranking of schools $\succ_i$ from every student, and outputs a random assignment. Note that the function $\mathcal{M}$ may implicitly incorporate various rules for prioritizing one student over another, but these are given a-priori and are therefore not treated as inputs.

For any mechanism $\mathcal{M}$, we define its *max-cohesion correlated lottery implementation* as a mechanism induced by the following maximization program:

$$\mathcal{M}^{\text{max-cohesion}}(\text{input}) = \arg\max_x E[f(x)] \tag{1}$$

$$\text{s.t. } E[x] = E[\mathcal{M}(\text{input})] \tag{2}$$

$$x \in X.$$

This optimization maximizes cohesion subject to maintaining the same assignment probabilities as the original mechanism. This is equivalent to maximizing the second component of expected social welfare in our utility model, which implies that if the original mechanism is Pareto-efficient with respect to school-specific utilities, the correlated lottery implementation becomes Pareto-efficient with respect to the full lexicographic utility. By inheriting the assignment probabilities from the original mechanism, we preserve any statistic of the original mechanism that can be expressed in assignment probabilities, including expected distance to assignment, probability of attending a certain set of schools, and expected value of the first component of the preference tuple. This is attractive because the decisions of what choices students have and who gets what priority is often the result of an intensely-debated political process, so being able to maintain the exact same assignment probabilities for everyone, while improving community cohesion, makes this approach less controversial.

2.1. NP-Hardness even with 2 schools

Consider the case of 2 schools. Let the schools be labeled 1 and 2, with capacities q_1 and q_2 respectively. Suppose that both schools are acceptable to every student, and that the total number of seats matches the number of students, $q_1 + q_2 = n$.

Without loss of generality, let school 1 be the over-demanded school. We assume that students' priorities for school 1 are based on their "priority class," and a higher priority student always takes precedence over a lower priority student. For students from the same priority class, we require that if they both prefer school 1, then they get in with the same probability.

Despite the generality of the priority structure, we can characterize the structure of the assignment probabilities. There is a "cutoff priority level" for school 1 at which any student who prefers school 1 with higher than cutoff priority will get in school 1; any student with lower than cutoff priority level or prefers school 2 will get into school 2; students who who prefer school 1 with exactly the cutoff priority level will be allocated based on a fair lottery. Respectively denote these sets of students D (get in school 1 for sure), F (get in school 2 for sure), and E (allocated based on lottery). Define the number of seats to be assigned by lottery to be $q^* = q_1 - |D|$. Students in E are assigned to school 1 with probability

$$p = \frac{q^*}{|E|},$$

and school 2 otherwise. The random assignment is illustrated in figure 1.

An upper-bound on cohesion is if communities in E are always assigned together. In order to achieve this, we need to be able to partition E into subsets of size q^*, which is a hard "packing" problem. So it is computationally hard to decide whether this upper-bound can be achieved.

PROPOSITION 1. *Unless P=NP, even with* 2 *schools, there exists no polynomial time algorithm (or FPTAS) to compute the maximum achievable cohesion by correlated lottery implementation.*

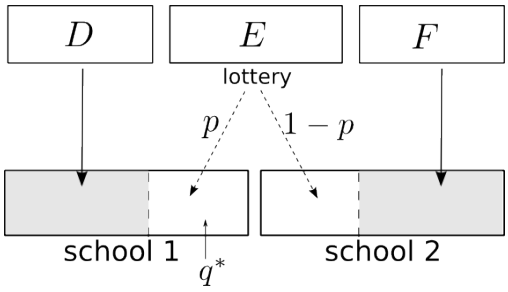

Figure 1 Random assignment in the 2-school case: students who prefer school 1 with higher than cutoff priority level (set D) get in for sure; students who prefer school 1 with exactly the cutoff priority level (set E) get in with probability $p = \frac{q^*}{|E|}$; students who prefer school 2 or whose priority level is lower than cutoff (set F) get in school 2 for sure.

However, if there are many communities and no single community dominates in size, one can show that we can achieve this upper-bound approximately. By analyzing the "large market limit" as the number of communities go to infinity, one can obtain an exact formula for max-cohesion. We defer this to the Online Appendix because the insights are similar to the insights from the large market analysis in Section 4, which uses a model that can encompass many schools.

This NP-hardness result precludes us from having nice expressions for cohesion from correlated lottery in the finite market model. To circumvent this difficulty, we "smooth away" the NP-hardness by considering the "large market" environment, in which there is a continuum of communities. Although in practice the number of communities is finite, the continuum model allows for simple analytic expressions for maximum cohesion and improvement by correlated lottery, thus illustrating the underlying insights in a clean way.

In Section 3, we setup a large market model and prove a useful characterization result, which may be of independent interest. In Section 4, we use this characterization to gain insights about how much we can gain in cohesion from lottery correlation and in what environments we can expect the most improvement.

3. Large market characterization of reasonable one-sided matching mechanisms with priorities

In this section, we define a large market model for one-sided matching markets with priorities. We show that in this setup, any mechanism that satisfies certain regularity conditions (non-atomicity, Bayesian incentive compatibility, symmetry and efficiency within each priority class) can be interpreted as "lottery-plus-cutoff": each student is given an identically distributed lottery number and each school sets a lottery-cutoff for each priority class (the cutoffs may depend on the distribution of preferences for each priority class); a student is assigned her most-preferred school for which she meets the cutoff. This sets up the basis for our analysis of cohesion in Section 4. However, this characterization itself does not have to do with communities or cohesion.

3.1. Large market model

Let the set of students, I, be represented by a subset of Euclidean space of Lebesgue measure 1. Let S be a finite set of schools, with $|S| = m$. For $s \in S$, let q_s be the school's capacity.

As before, each student i submits preferences $\succ_i$, which is a ranking of schools. We assume that every school is acceptable to every student, and that every student ranks all schools. There are $m!$ possible rankings, and we assume that for each possible ranking, the set of students that submit this ranking is measurable.

Our model is one-sided matching with priorities, which means the following. Students are partitioned into priority classes Π. We assume that for each priority class, the set of students of that priority class is measurable and of positive measure. Furthermore, the distribution of priority classes and the distribution of preferences within each priority class is common knowledge.

Given students' preferences and priorities, an assignment mechanism is represented by a random indicator function $\mathbb{1}_s(i)$, which equals 1 if student i is assigned to school s, and 0 otherwise. We assume that the assignment mechanism satisfies the following regularity conditions in the large market setting:

- Non-atomicity: Any single student changing her preferences has no effect on the assignment probabilities of others.
- Bayesian incentive compatibility in school-specific utilities: Given the distribution of students' preferences within each priority class, students reporting truthfully is a Bayes-Nash equilibrium. In other words, given the measure of students of each priority class and of each possible preference ranking, assuming that everyone else reports truthfully, a student cannot improve her expected school-specific utility $E[u_i]$ by submitting a false preference. Henceforth we denote this simply by "incentive compatibility."
- Symmetry: Students in the same priority class with the same preferences receive the same assignment probabilities.
- Efficiency within each priority class: For students in the priority class, there does not exist a Pareto improvement "trading cycle" where by $s_0 \prec_{i_0} s_1$, $s_1 \prec_{i_1} s_2 \cdots s_l \prec_{i_l} s_0$ and i_0 has positive probability of being assigned s_0, i_1 has positive probability for s_1, etc.

Note that since these conditions only need to hold in the large market, one can interpret them as only needing to be "asymptotically" true. This means that for example, the "incentive compatibility" condition in the above is actually the less-restrictive incentive compatibility "in the large" condition described in Azevedo and Budish (2012). Taking the suitable limit and ignoring"knife-edge" cases, two of the most widely used mechanisms–Deferred Acceptance with Single Tie-breaking (DA-STB) and Top Trading Cycle with Single Tie-breaking (TTC-STB), assuming that students are allowed to rank as many choices as they would like, both satisfy the above conditions in the large market, so both fall into our framework.[2] DA-STB is used in Boston, New York, Denver and San Francisco. TTC-STB is used in New Orleans. For descriptions of these mechanisms, see Abdulkadiroğlu and Sönmez (2003b) and Abdulkadiroğlu et al. (2009).

This model is similar to the continuum two-sided matching model in Azevedo and Leshno (2012). The main difference is that preference structure is one-sided in our model and two sided in their model: only students have preferences over schools in our model, while schools also have strict preferences over students in their model. In their paper, they show that a continuum two-sided matching model can be interpreted as a limit of discrete matching models. This provides a theoretical foundation for such continuum models. Other works that use continuum matching models include Abdulkadiroğlu et al. (2008), Miralles (2008), and Budish and Cantillon (2012).

3.2. A characterization theorem

We show that in the large market model, any mechanism that satisfies our four regularity conditions can be described as a "lottery-plus-cutoff" mechanism: each student gets an identically distributed lottery number and each school sets a lottery-cutoff for every priority class; a student is assigned her most preferred school for which her lottery meets the cutoff.

DEFINITION 1. A mechanism is *lottery-plus-cutoff* if it can be described as follows: given preference submissions, each student i receives an identically distributed lottery number z_i (may be jointly correlated). WLOG, $z_i \propto Uniform[0,1)$. Given the measure of students in each priority level submitting each ranking, schools have a priority-dependent lottery cutoff $z^*_{\pi,s}$. Student i is assigned her most preferred school for which she meets the cutoff ($z_i \geq z^*_{\pi(i),s}$).

If in addition the lottery numbers z_i are independently generated for different students, then we call the lottery *independently implemented*.

THEOREM 1. *In the continuum model, an assignment mechanism is non-atomic, Bayesian incentive compatible, symmetric and efficient within each priority class if and only if it is a lottery-plus-cutoff mechanism.*

The core ideas behind the proof appeared previously in Liu and Pycia (2012), which shows that without priorities and in the large market, any mechanism that is non-atomic, strategyproof, symmetric and efficient is equivalent to the so-called probabilistic serial mechanism, or equivalently lottery-plus-cutoff with one priority class. The main difference with our result is that while their setup does not have priorities, we allow arbitrary priorities. Moreover, while their analysis studies the limit as a discrete model is scaled up, our analysis is directly in the continuum setting. Our proof is given in the Online Appendix EC.3.

One way to interpret the above result is that in a large one-sided many-to-one matching market, asymptotic incentive compatibility and asymptotic efficiency within each priority class constrain the mechanism significantly, leaving policy makers with only two control levers: (1) cutoffs for each priority class and (2) lottery correlation. Assuming in addition that the mechanism does not waste desirable resources, then the first lever, determining cutoffs, is a purely distributional question: how much claim does each priority class have on each resource and whether such claims can be traded.[3] The second lever, lottery correlation, is what we study in this paper.

4. Cohesion in the large market model

The characterization in Section 3.2 allows us to study the impact of lottery-correlation in a wide class of mechanisms. Once we know that a mechanism is "lottery-plus-cutoff," we can isolate the effects of lottery-correlation by treating the cutoffs as given. By doing this, we insulate the analysis from the complexities of how the mechanism treats different priority classes, as all such subtleties are endogenized into the cutoffs.

We first define communities and cohesion in the large market model. Let the unit interval $C = [0, 1]$ represent a continuum of communities. (We need infinitely many communities in order to show analytic results, because the hardness result in Section 2.1 still hold even if we had infinitely many students per community but a finite number of communities.) Let the set of students within each community be also represented by the unit interval $[0, 1]$, so the set of all students can be represented by the Cartesian product of two unit intervals, which is the unit box, $I = [0, 1]^2$. Each student $i \in I$ can be represented by its coordinates (c, y), in which c is the student's community and y is the student's index in this community. (This implicitly assumes that communities have equal size, but this can be relaxed.) The community membership function $c(i)$ is simply the first coordinate of i. For simplicity, assume that the total supply of seats equals total demand, so $\sum_s q_s = 1$.

Define the cohesion for school s, $f^s = E[\mathbb{1}_s(i)\mathbb{1}_s(i')|c(i) = c(i')]$. This is the expected measure of same community pairs assigned to s. Define $f = \sum_s f^s$ as the total cohesion, which is the measure of pairs of students from the same community assigned to the same school.

Our analytical results on max-cohesion correlated lottery require an additional assumption, which we call *homogeneity of cutoffs within communities*.

ASSUMPTION 1. *For each school, everyone from the same community sees the same cutoffs to this school.*

This would hold if everyone in the same community belongs to the same priority class, which would be the case if communities and priorities were purely geographically based and communities stay together in any geographic division. Under this assumption, maximum community cohesion can be attained by giving everyone in the community the same lottery number. Without this assumption, all of the expressions for max-cohesion in this section would still hold as upper-bounds, but they might not be attainable.

Given students submitted preferences and the mechanism, define the following quantities:

$p_s(i)$:= probability of student i of being assigned to school s. For $i = (c, y)$, we also write $p_s(c, y) := p_s(i)$.

$\mathbb{1}_s(z|i)$:= Indicator for whether student i at lottery z would be assigned to school s. (It is 1 iff at lottery z, student i finds school s to be her most desirable option for which she meets the lottery cutoff.) We call this student i's demand function for school s.

$\bar{p_s}(c)$:= $\int_y p_s(c, y)dy$. The total assignment probabilities to s of students in community c.

$d_s(z|c)$:= $\int_y \mathbb{1}_s(z|(c, y))dy$. The total demand for s from community c at lottery z. We call this community c's demand function for s.

These quantities, as well as the cutoff structure, are illustrated in Figure 2. This graphs the students in a community on the horizontal axis and the lotteries on the vertical axis. It is a "slice" of a cube which would correspond to $I = [0,1]^2$ in two dimensions and lottery $z \in [0,1]$ in the third dimension.

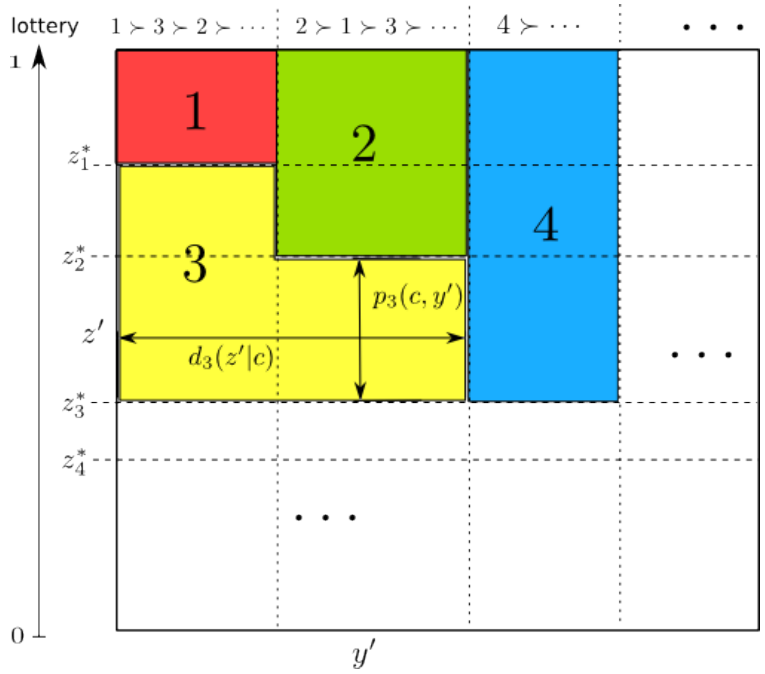

Figure 2 Assignment within community c: horizontal dimension is students and vertical dimension is lottery numbers. z_1^*, z_2^*, z_3^* and z_4^* are lottery cutoffs for community c of schools $1, 2, 3, 4$ respectively (well-defined because we assume everyone in the same community are in the same priority class). $d_3(z'|c)$ is the demand function for school 3 at lottery z'. $p_3(c, y')$ is the probability student y' is assigned to school 3. The area shaded with the same color represents the 1-dimensional measure of assignment probabilities of students from this community to a school ($\bar{p_s}(c)$).

Note from Figure 2 that $\bar{p_s}(c) = E_z[d_s(z|c)]$, because both are equal to the "area" assigned to school s. The first is integrating horizontally, and the second is integrating vertically.

The following theorem provides a simple analytical formula for cohesion to any given school in the independent and max-cohesion lottery implementations. The formula for the max-cohesion case uses the additional assumption that everyone in the same community has the same cutoff for the school.

THEOREM 2. *In the continuum model, for any school s, the cohesion from independent lottery implementation is:*

$$f^s_{independent} = E_c[\bar{p_s}^2] = q_s^2 + Var_c(\bar{p_s}).$$

Under Assumption 1,

$$f^s_{max} = E_{c,z}[d_s^2] = f^s_{independent} + E_c[Var_z(d_s|c)].$$

Thus, the total cohesion,

$$f_{independent} = \sum_s q_s^2 + \sum_s Var_c(\bar{p_s}),$$
$$f_{max} = f_{independent} + \sum_s E_c[Var_z(d_s|c)].$$

We call $f_{independent}$ the Baseline cohesion*, and $(f_{max} - f_{independent})$ the* potential gain in cohesion from correlated lottery.

The proof (see Online Appendix EC.3) follows from the cutoff structure, re-arranging orders of integration, and the law of total variance. We interpret the terms as follows.

- $\sum_s q_s^2$: Herfindahl index of school size. This is a measure of variation in school sizes. The more varied school sizes are, the higher the baseline cohesion.
- $\sum_s \text{Var}_c(\bar{p_s})$: Between community variation in assignment probabilities. The more varied assignment probabilities are between communities, the higher the baseline cohesion.
- $\sum_s E_c[\text{Var}_z(d_s|c)]$: Potential gain in cohesion from using correlated lottery. For each school, the summand is the average across communities of the variation in demand function. In other words, it is how much, on average, the lottery number affects the mass of students from a community assigned to a school. Intuitively, this has to do with how competitive schools are (high cutoffs), and how correlated preferences within a community are. This intuition is made more precise in Corollary 2.

By manipulating the expressions in Theorem 2, we immediately yield 2 corollaries. The first shows that the maximum possible improvement ratio is upper-bounded by the number of schools, which can be achieved if all schools have size $\frac{1}{m}$ and all students share the same preferences. The second corollary interprets the gain in cohesion as a weighted average of an aggregate statistic representing competition and preference correlation.

COROLLARY 1. *(Upper-bound on improvement ratio)*

$$\frac{f_{max}}{f_{independent}} \leq \frac{1}{\sum_s q_s^2} \leq m.$$

COROLLARY 2. *Under Assumption 1, the proportional improvement in cohesion for lottery correlation can be interpreted as a weighted measure of the Squared Coefficient of Variation (SCV) of the demand function. Define the Squared Coefficient of Variation $SCV_{d_s}(c) = \frac{Var_z(d_s|c)}{E_z[d_s|c]^2}$. Intuitively, the SCV of the demand function is a combination of competition (high cutoffs) and high within-community preference correlation.*

Define weights $w_c = E_z[d_s|c]^2 = \bar{p_s}(c)^2$. For any school s, the proportional improvement in cohesion is a weighted average of the SCV:

$$\frac{f^s_{max} - f^s_{independent}}{f^s_{independent}} = \frac{E_c[w_c SCV_{d_s}(c)]}{E_c[w_c]}.$$

We now show that maximum cohesion in our large market setup can also be expressed as a constant minus a measure of within-community heterogeneity. More precisely, define $\Delta_s(i,i') = |p_s(i) - p_s(i')|$. This is the absolute difference in assignment probabilities to school s for individuals i and i'. Define the mean absolute difference in assignment probability for school s and community c to be

$$\bar{\Delta}_s(c) = E_{i,i' \in c}[\Delta_s(i,i')].$$

This is the expected value of $\Delta_s(i,i')$ for two randomly drawn individuals i, i' from community c. Averaging across communities, $E_c[\bar{\Delta}_s(c)]$ is then an aggregate measure of within-community heterogeneity in assignment probabilities to school s. The following proposition shows that maximum cohesion to a school s is equal to the size of the school minus one half of this measure of within-community heterogeneity.

PROPOSITION 2. *(Cohesion is limited by within-community heterogeneity) Under Assumption 1,*

$$f^s_{max} = q_s - \frac{1}{2} E_c[\bar{\Delta}_s(c)].$$

So that summing across all schools, $f_{max} = 1 - \frac{1}{2}\sum_s E_c[\bar{\Delta}_s(c)]$, where the term being subtracted is an aggregate measure of within-community heterogeneity in assignment probabilities.

The expression for gain in cohesion in Corollary 2 is exact but difficult to think about. The SCV of demand function is hard to estimate. The following proposition bounds potential gain from lottery correlation by an easier-to-estimate measure of lottery uncertainty.

PROPOSITION 3. *(Decomposition of potential to improve) Under Assumption 1, the gain from cohesion from correlated lottery is equal to the average individual assignment variance minus a function of within-community differences in assignment probabilities.*

$$\begin{aligned} f^s_{max} - f^s_{independent} &= E_i[p_s(1-p_s)] - \frac{1}{2} E_{c(i)=c(i')}[\Delta_s(i,i')(1 - \Delta_s(i,i'))] \\ &\leq E_i[p_s(1-p_s)]. \end{aligned}$$

In the first line, the first term is the average across individuals of assignment variance to school s; this is a measure of how uncertain the lottery is for school s. The second term is minimized when $\Delta_s(i,i')$ is either close to 1 or close to 0. In other words, for correlated lottery to be most effective, we desire that for two students within the same community, they either get assigned to a school with very similar probabilities, or one gets assigned with very high probability and the other with very low probability. Because the second term is always non-negative, gain from cohesion is upper-bounded by the uncertainty of the lottery.

The following proposition shows exactly when lottery-correlation is useful. It turns out that for every school, the following three quantities add up to a constant:

1. Between-community heterogeneity in assignment probability.
2. Within-community heterogeneity in assignment probability.
3. Potential to improve cohesion by lottery-correlation.

So that we are generally in one of the following 3 cases:

1. High between-community heterogeneity: cohesion with independent lottery is already high.
2. High within-community heterogeneity: there is nothing we can do. Maximum cohesion is severely limited.
3. High potential to improve.

PROPOSITION 4. *(Structural identity) For every school s, the following three terms add to a constant:*

$$\underset{}{q_s(1-q_s)} = \underset{\substack{\textit{between-community}\\ \textit{heterogeneity}}}{Var_c(\bar{p_s})} + \underset{\substack{\textit{within-community hetero-}\\ \textit{geneity}}}{\tfrac{1}{2}E_c[\bar{\Delta}_s(c)]} + \underset{\substack{\textit{potential to improve}}}{(f^s_{max} - f^s_{independent})}$$

This relationship is illustrated in Figure 3.

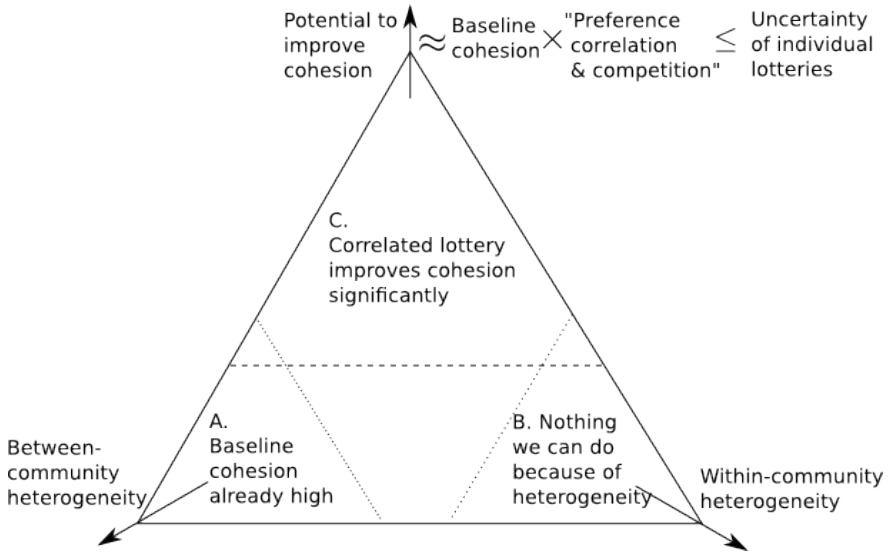

Figure 3 Diagram illustrating structural identity (Proposition 4). The triangle represents 3 quantities that add up to a constant. There are 3 possible cases: A) high between-community variation in assignment probability, so baseline cohesion is already high. B) high within-community heterogeneity in assignment probability, so there is nothing we can do by Proposition 2. C) significant potential to improve cohesion by correlated lottery. Corollary 2 shows that potential to improve can be interpreted as baseline cohesion multiplied by a weighted average of the SCV (squared coefficient of variation) of the demand function (with respect to the lottery z). The SCV can be interpreted to correspond to a mixture of preference correlation and competition. Proposition 3 shows that the improvement is upper-bounded by a measure of total uncertainty of the lottery. This relationship holds not only overall but also school by school.

4.1. Embedding a model with preferences

To more directly illustrate the relationship between preference structure and cohesion, we consider an explicit model of students' preferences. Suppose that student i's preferences for schools is driven by random utility model

$$u_{is} = \alpha\nu_s - \beta\omega_{c(i)s} + \epsilon_{is},$$

where ν_s corresponds to "quality" of school s, ω_{cs} is distance from community c to school s (more generally can capture any community-specific propensities for specific schools), and ϵ_{is} is an idiosyncratic shock drawn from a standard Gumbel distribution. This is a realistic structure of preferences that has been used to fit data in empirical studies. (See Pathak and Shi (2013).) For simplicity, we assume that for different schools s, s', $\nu_s \neq \nu_{s'}$ and $\omega_{cs} \neq \omega_{cs'}$. We assume that there are no priorities, so Assumption 1 trivially hold. We study the behavior of $f_{\text{independent}}$ and f_{max} as functions of α and β.

The following proposition shows the limit behavior of $f_{\text{independent}}$ and f_{max} as α (how much quality matters) or β (how much distance matters) goes to infinity. It shows that if there is no between-community heterogeneity ($\beta = 0$), cohesion with independent lotteries is fixed and is the lowest possible, regardless of how correlated students preferences are. On the other hand, with correlated lottery, perfect cohesion can be achieved without between-community heterogeneity, if preferences for quality are very correlated. With high between-community heterogeneity, perfect cohesion can be approached in both independent and correlated cases. This corroborates the triangle diagram in Figure 3.

PROPOSITION 5. *Assume that capacities are such that when everyone goes to their closest school, the closest school can accommodate. For any α_0,*

$$f_{\textit{independent}}(\alpha_0, 0) = \sum_s q_s^2, \tag{3}$$

$$\lim_{\beta\to\infty} f_{\textit{independent}}(\alpha_0, \beta) = 1. \tag{4}$$

For the correlated case, regardless of capacities, for any α_0 and β_0,

$$\lim_{\alpha\to\infty} f_{\textit{max}}(\alpha, \beta_0) = 1, \tag{5}$$

$$\lim_{\beta\to\infty} f_{\textit{max}}(\alpha_0, \beta) = 1. \tag{6}$$

While the above result is only for the limit, the following proposition shows comparative statics for the finite case, in the special case that $\beta = 0$. It shows that if capacities are not too different so that more desirable schools are also more over-demanded (more applicants per seat), then as preferences become more correlated, cohesion from independent lottery stays fixed, while cohesion from correlated lottery increases.

THEOREM 3. *(Pure vertical differentiation) Suppose that $\beta = 0$ (no between-community heterogeneity). Let $r_s = e^{\nu_s}$. Suppose that the schools are ordered so that the more desirable schools are first, so $r_1 > r_2 > r_3 \cdots$. Assume that capacities are not "too different," so that dividing by capacities do not change this relative order, so $\frac{r_1}{q_1} \geq \frac{r_2}{q_2} \geq \frac{r_3}{q_3} \cdots$. Then for every school s, while $f^s_{independent}(\alpha, 0) = q_s^2$ is constant in α, $f^s_{max}(\alpha, 0)$ is strictly increasing in α.*

5. Computing the correlated lottery implementation in practice

Fixing students' submitted rankings, let p be the assignment probabilities of the original mechanism. (p_{is} is the probability student i is assigned to school s under the mechanism.) The max-cohesion correlated-lottery implementation problem can be written as

$$\begin{aligned} \text{Max} \quad & E[f(x)] \\ \text{s.t.} \quad & E[x] = p \\ & x \in X. \end{aligned} \tag{7}$$

Define the assignment polytope

$$\mathcal{P} = \left\{ a \in \mathbb{R}^{n \times m} : \sum_s a_{is} = 1, \sum_i a_{is} \leq q_s, 0 \leq a_{is} \leq 1 \right\}.$$

Represent random assignment x explicitly as $\{(\lambda^l, a^l) : a^l \in \text{Vertices}(\mathcal{P}), \lambda^l \in [0,1], \sum_l \lambda^l = 1\}$, so that $x = a^l$ with probability λ^l, and let A be the nm by $|\text{Vertices}(\mathcal{P})|$ matrix in which the columns encode the vertices of assignment polytope $\mathcal{P}$. Define $f(A)$ naturally as the vector in which the lth component is the cohesion of the lth column of A. We can rewrite the above in the more explicit form

$$\begin{aligned} \text{Max} \quad & f(A) \cdot \lambda \\ \text{s.t.} \quad & A\lambda = p \\ & \vec{1} \cdot \lambda = 1 \\ & \lambda \geq 0. \end{aligned} \tag{8}$$

which is a standard form linear program (albeit with exponentially many variables). The theory of LP implies that there is an optimal solution with only nm positive components of λ (because $\text{rank} \begin{pmatrix} A \\ \vec{1} \end{pmatrix} \leq nm$). In other words, for any assignment probabilities $\{p_{is}\}$, there exists a random assignment x with $E[x] = p$ which randomizes over at most nm deterministic assignments and achieves maximum cohesion.

This suggests the following mechanism.

1. Estimate the individual assignment probabilities p_{is} of the original mechanism by independently simulating T times. Note that the computed estimates are unbiased and have component wise standard deviation $\sqrt{\frac{(1-p_{is})p_{is}}{T}} \leq \frac{1}{2\sqrt{T}}$.

2. Use the estimated p as inputs to program (8) and compute a convex combination of deterministic assignments $\{(\lambda^l, a^l) : \sum_l a^l = 1\}$. Output assignment a^l with probability λ^l.

For any T, the resultant randomized mechanism induces the same individual assignment probabilities as the original mechanism (using crucially the unbiasedness of the simulation in first step).

The only difficulty is what algorithm to use to solve the intractably large program (8). As shown in Proposition 1, solving the cohesion optimization is NP-hard even with 2 schools. In Online Appendix EC.1, we show that the case with many schools is related to the notoriously hard Quadratic Assignment Problem (QAP), which is NP-hard to approximate to any constant factor (Burkard et al. (1998), Sahni and Gonzalez (1976)).

We propose a simple heuristic to solve this in practice. This heuristic is related to the Birkhoff-von Neumann theorem, as it seeks to express the original assignment probabilities as a convex combination of deterministic assignments by iteratively breaking off one deterministic assignment at a time. To find a deterministic assignment at each iteration, it solves a max-weighted bipartite perfect matching problem, with the students on one side of the graph to be matched to the schools on the other side. As input to the max-weighted matching procedure at each iteration, we define an edge from a student to a school if and only if the assignment probability of that student to the school is positive, after having subtracted off the deterministic assignments from previous iterations. The weight of this edge is randomly generated, but we constrain the weights to be the same for everyone from the same community to the same school. An intuition of why this might work is that conditional on student i being assigned to school s in the max-weighted perfect matching, the edge weight of that student to the school is probably high, and so the edge weight of other students from i's community to s is probably high, so we expect many of them to be co-assigned with i to s. Another intuition is that giving the same edge weights to students of the same community reduces "local minima" in which a trading cycle can increase cohesion. Details of the heuristic and elaborated intuition is in Appendix EC.1. We implemented our correlated-lottery implementor in Java (code is available upon request). Our heuristic does not require Assumption 1 to hold. As shown in Section 6, this heuristic achieves good results even when students of the same community have different priorities. Assumption 1 was only needed to prove exact analytical results in the large market case.

5.1. An upper-bound on maximum cohesion

To evaluate the optimality gap of our heuristic, we derive a simple upper-bound to the correlated lottery implementation mathematical program (7). Consider student i and school s, conditional on the student being assigned to s, the expected number of same community peers that can be co-assigned to school s, $E[v_i(x)|x_{is}]$ is upper-bounded by

$$E[v_i(x)|x_{is}] \leq \min\left((q_s - 1), \sum_{c(i')=c(i), i' \neq i} \min(1, \frac{p_{i's}}{p_{is}})\right).$$

Where the first term follows from the capacity constraint of s, and the second term follows from $E[x_{i's}|x_{is}] = \frac{E[x_{is}x_{i's}]}{p_{is}} \leq \frac{\min(p_{is},p_{i's})}{p_{is}}$.

Summing over all students and taking the expectation over school s, we get that cohesion is upper-bounded as follows.

$$E[f(x)] = \frac{1}{n}\sum_{i,s} p_{is}E[v_i(x)|x_{is}] \tag{9}$$

$$\leq \frac{1}{n}\sum_{i,s} \min\left(p_{is}(q_s-1), \sum_{c(i')=c(i),i'\neq i} \min(p_{is},p_{i's})\right). \tag{10}$$

6. Application to Boston elementary school choice

6.1. Description of school choice in Boston

School Choice in Boston Public Schools (BPS) began with the adoption of the Controlled Choice Student Assignment Plan in 1988. The plan organized public elementary and middle schools into three zones–East, North, and West–and students were given the option to apply to any set of schools within their zone. To apply, students submitted ranked lists of their preferred schools, and a centralized lottery produced the assignment. Since then, policies regarding the assignment process have been revised numerous times, including the lottery algorithm itself, but the overall framework of the process remained the same.

Our empirical study focuses on BPS elementary school assignment in 2012, which is when Mayor Menino made his call to improve community cohesion in school assignment. We focus on elementary schools because this is arguably the time when going to school with neighbors is most important, and because this was the focus of the mayor's call for reform. The goal of this study is to analyze what would have happened if we had adopted a correlated lottery procedure to improve community cohesion in 2012, and to compare with alternative approaches to improve cohesion considered by the mayor-appointed city committee.

The vast majority of students enter BPS elementary schools via entry grades K1 and K2 (K for kindergarten). To enroll, students participate in one of 4 application rounds by submitting rankings over specific programs in schools (a school may offer several programs: regular, English Language Learner, Montessori, etc). They can rank as many schools as they would like. The first round occurs in January. and this is when the majority of families participate (about 80% of families who eventually apply first apply in Round 1.) For families that come later, there are 3 smaller subsequent rounds that take place from March to June. There is also a wait-list process in which families may get a seat at a subsequent round if more capacity becomes available or if others drop out. The wait-list favors applicants from earlier rounds so it is always the best to apply in Round 1. For simplicity, since the majority of seats are allocated in Round 1, we focus on Round 1 in this paper.

After families submit choices, the assignment is computed by Deferred Acceptance with Single Tie-Breaker (DA-STB). This was adopted in 2005 to eliminate the need for strategic manipulation. See Abdulkadiroğlu et al. (2006). More precisely, each program is internally divided into 2 halves, a walk-zone half and an open half. Students' preferences on programs are augmented into preferences on halves, such that students living in the walk-zone (within one mile of school) prefers the walk-half and students from outside the walk-zone prefer the open half. (Preferences between different programs are maintained in the augmentation.)

Each of the program halves also rank students in the following way: each student is given an i.i.d. random lottery number. The ranking over students is induced by several levels of priorities (1st level is most important, 2nd is to break 1st level ties, 3rd level is to break 2nd level ties, etc). The priorities are given in Table 3.

Hierarchy	Priority rule
1st level	continuing students $>$ others
2nd level	have sibling in this school $>$ others
3rd level (only for walk-halves)	lives in walk-zone $>$ others
4th level	by lottery number

Table 3 Order of priorities used in Boston elementary school assignment in 2012. The earlier level priorities are more important, with later levels only used to break ties. The 3rd level is applied only for walk-zone halves.

Given students' rankings on program halves and program halves' rankings on students, assignment is by the student-proposing deferred acceptance algorithm, which is as follows:

1. Find an unassigned student, have her apply to her top remaining choice.

2. If the program half is not full, accept her; otherwise, bump out the least preferred student from that program half (which may be her), and remove this program half from that student's ranking.

3. Iterate until all unassigned students have gone through all their choices.

It is well-known that the above algorithm induces a unique assignment regardless of the order of application. (See Roth and Sotomayor (1990).) This induces an assignment of students to school programs, and this assignment is mailed to families. It is well-known that if families can submit as many choices as they would like and if their submissions do not influence the priorities, then this assignment process is strategyproof for all students. (See Abdulkadiroğlu and Sönmez (2003b).)

6.2. Data

We use 2012 Round 1 choice data for grades K1-2, which have been anonymized but still contain information on students' demographics, geocode (division of Boston into 868 smaller regions), top 10 choices, and final assignment. Although we were not given capacities of programs, we were able to infer them from final assignments. As a check, we were able to replicate 98.2% of K1 assignment and 99.0% of K2 assignment

	K1	K2	K1-2
Schools	66	75	75
Programs	106	123	229
Seats	1921	3689	5610
Continuing students	167 (6%)	1904 (47%)	2071
Non-continuing siblings	690 (26%)	467 (12%)	1158
New families	1809 (68%)	1659 (41%)	3469
Total Students	2666	4030	6696

Table 4 Summary statistics of the choice data.

	K1		K2	
	% of students	% assigned top choice	% of students	% assigned top choice
Continuing	6%	92%	47%	95%
Non-continuing siblings	26%	80%	12%	79%
New families	68%	24%	41%	29%

Table 5 Percentage of students of each type getting their top choice. For example, for K2, 47% of students are continuing from K1, and 95% of them get their top choice.

(excluding students who were administratively assigned by BPS after the assignment algorithm described before has finished). Table 4 summarizes the supply and demand data.

In the data it turns out that the lottery mostly matters only for new families (non-continuing, non-siblings). Table 5 tabulates the fraction of students of each type getting their top choice.

Since continuing students and siblings are very likely to be assigned their first choice (so there is essentially no lottery for most of them), we can only hope to significantly impact via lottery correlation those who are new families. So in reporting outcomes, we focus on the new families.

Our approach also takes as input delineations of community. The city may want to do this based on natural dividing lines or other considerations. For the purpose of this study, we simply use a square grid of .5 miles in length, with each .5 mile $\times$ .5 mile square defining a community. Figure 4 plots the geographic distribution of all 6696 K1-2 students, partitioned into 205 non-empty communities.

6.3. Impact of correlated lottery implementation

We take the actual choices, simulate the current assignment algorithm 1000 times with independently drawn lottery numbers to estimate the assignment probabilities, and run the correlated lottery heuristic described in the Online Appendix EC.1.1 to produce another 1000 assignments with the same assignment counts for each student-program pair, but correlated so within the same assignment students from the same community are more likely to be co-assigned to the same school. Drawing one of the 1000 correlated assignments is then a correlated lottery that replicates the estimated assignment probabilities for all students, but has improved cohesion. In Table 6, we tabulate for various groups of students their average baseline cohesion (without correlation), improved cohesion (with correlated lottery), upper-bound to cohesion (from Section 5.1),

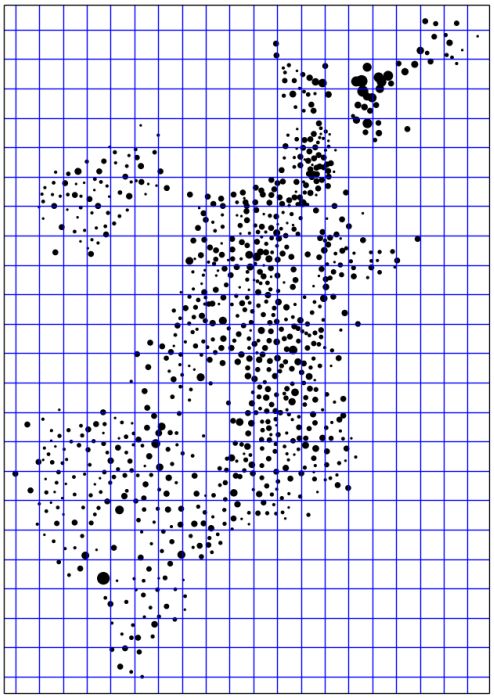

Figure 4 Partition of Boston into .5 mile×.5 mile squares. We treat each square as a community. Each circle corresponds to a geocode, and the area of the circle is proportional to the number of students residing at the geocode. Defined in this way, the average number of students per community is $2666/205 = 13$ for K1 and $4030/205 = 19$ for K2.

	Baseline	Improved	Upper-bound	Improvement	Bound on Improvement
K1					
All students	1.35	2.11	2.70	0.75	1.34
Continuing students	1.30	1.32	1.38	0.02	0.08
Non-continuing siblings	1.35	1.43	1.56	0.08	0.21
New families	1.36	2.44	3.26	1.08	1.90
K2					
All students	2.48	2.89	3.39	0.42	0.91
Continuing students	2.26	2.27	2.30	0.01	0.04
Non-continuing siblings	2.91	3.01	3.23	0.10	0.32
New families	2.61	3.58	4.69	0.97	2.08

Table 6 Impact of lottery correlation on cohesion for various groups of students. For example, for K1 new families, conditional on being assigned, students on average can expect to find 1.36 same-community peers. If we use correlated-lottery, the number of same community peers improves by 79% to 2.44. Based on the given assignment probabilities, no correlation procedure can produce cohesion greater than 3.26. Using correlated lottery, the average K1 new family gains 1.08 additional same-community peers. No correlation procedure can produce a gain greater than 1.90.

amount of improvement (improved minus baseline), and upper-bound on improvement (upper-bound minus baseline).

As can be seen, while lottery correlation does little for continuing students and siblings (who mostly get their 1st choice regardless of the lottery), it increases average cohesion for new families by about 1 in both K1 or K2. This is one additional "neighbor" these students can find at their school assignment, which is a

	Baseline	Improved	Upper-bound	Improvement	Bound on Improvement
K1					
All	0.53	1.11	1.48	0.58	0.95
Continuing students	0.55	0.58	0.60	0.03	0.04
Non-continuing siblings	0.54	0.64	0.73	0.10	0.19
New families	0.53	1.26	1.73	0.74	1.20
K2					
All	0.98	1.34	1.71	0.36	0.73
Continuing students	0.93	0.95	0.99	0.02	0.06
Non-continuing siblings	1.01	1.12	1.31	0.12	0.31
New families	1.00	1.64	2.27	0.64	1.27

Table 7 Expected number of same-community peers co-assigned conditional on being assigned outside of walk-zone. This is the number of same-community neighbors a student who is traveling outside of his/her 1-mile walk-zone can expect to find at his/her assigned program. So for a K1 new family, conditioning on the student going outside of walk-zone for school, he/she on average has only 0.53 neighborhood peers in the current lottery. But with correlated lottery this more than doubles to 1.26. Without changing assignment probabilities, we cannot expect this to be larger than 1.73.

significant increase since the baselines are 1.36 and 2.61 respectively. Moreover, even if we had solved the max-cohesion problem to optimality, we cannot expect the improvement in cohesion to be more than 1.9 for K1 and 2.08 for K2, so to achieve greater gains we would need to alter the assignment probabilities.

One motivation for increasing cohesion is so that families can car-pool and students can have neighborhood friends, but to some extent this matters only when the student is going to school not in his/her neighborhood (otherwise he/she would not need to go by car and would have neighborhood friends regardless). In Table 7, we examine the expected # of same-community peers conditional on being assigned outside of own walk-zone. The upper-bound uses a similar formula as the one in Section 5.1, except that it conditions differently. As seen, the proportional impact of correlated lottery is more pronounced in this case, more than doubling baseline cohesion for K1 new families and achieving a 64% gain over baseline cohesion for K2 new families.

6.4. Comparison and interaction with other reforms

During the 2012-2013 school choice reform, two types of reforms proposed to the city committee were increasing the walk-zone percentage and reducing the choice menu (the set of schools students from various neighborhoods could rank from). Both strategies were intended to affect assignment probabilities to result in closer to home assignment. By theorem 2, this would increase cohesion as it would increase between-community heterogeneity. We empirically estimate the increase in cohesion due to these potential reforms and compare to correlated lottery (which does not affect anyone's assignment probabilities.) We also evaluate the interaction of these reforms with lottery correlation, to see how much we can improve by applying both at the same time.

				Conditional on busing		
Walk-zone Percentage	Baseline	Improved	Improvement	Baseline	Improved	Improvement
K1						
50	1.36	2.44	1.08	0.53	1.26	0.74
60	1.44	2.56	1.12	0.48	1.13	0.65
70	1.56	2.72	1.16	0.46	1.13	0.67
80	1.69	2.86	1.18	0.44	1.07	0.64
90	1.82	3.01	1.19	0.46	0.95	0.50
100	1.91	3.11	1.20	0.54	0.86	0.33
K2						
50	2.61	3.58	0.97	1.00	1.64	0.64
60	2.74	3.73	0.99	0.95	1.59	0.65
70	2.90	3.87	0.97	0.91	1.53	0.62
80	3.09	4.05	0.96	1.00	1.62	0.61
90	3.21	4.14	0.93	1.05	1.72	0.66
100	3.32	4.21	0.89	1.08	1.70	0.62

Table 8 Baseline and Improved Cohesion with differing walk-zone percentages. The first three columns correspond to students' expected # of same-community peers co-assigned conditional on being assigned. The last three columns correspond to the same number conditional on being assigned outside of walk-zone.

6.4.1. Increasing the walk-zone percentage As described in Section 6.1, the BPS assignment algorithm in 2012 had 50% of seats of a program allocated to the walk-half (the side that respected walk-zone priority) and the rest to the open half. One approach to increase community cohesion is to induce closer-to-home assignment by increasing the percentage of seats allocated to the walk-half. For K1 and K2 new families, we show in Table 8 the result of this on their baseline cohesion, on their improved cohesion, as well as on their cohesion conditional on traveling outside of walk-zone (1 mile radius).

As can be seen, while increasing the walk-zone percentage increases cohesion, the maximum magnitude of increase (to 1.91 for K1 and 3.32 for K2) is less than if we kept the same walk-zone percentage but switched to correlated lottery (to 2.44 for K1 and 3.58 for K2). For the students who are traveling outside of their walk-zone, Table 8 shows that altering walk-zone percentage has almost no effect on their expected # of same-community peers, while lottery correlation significantly increases it. From the perspective of increasing community cohesion, both in terms of overall increase and in terms of helping those who need it the most, lottery correlation is more effective than increasing the walk-zone percentage.

6.4.2. Reducing the choice menus Another approach to increase cohesion is to decrease the choice menu, so as to focus choices from the same community to similar schools. To evaluate such a reform, we need a model for how students would choose given a new menu. We use the same demand model as in the study commissioned by the city committee during the 2012-2013 school choice reform to evaluate a range of potential outcomes. This is a multinomial logit model fitted using the same data as in this study, and includes a fixed effect for each school, and linear controls for distance to choices, racial/socio-economic interactions and school-specific affinities (whether a student is continuing student, has sibling at a school, or lives in the walk-zone of a school). The demand model is documented in detail in Pathak and Shi (2013).

				Cohesion conditional on busing		
Choice Menu	Baseline	Improved	Improvement	Baseline	Improved	Improvement
K1						
3-Zone	1.11	2.09	0.98	0.42	1.16	0.74
6-Zone	1.32	2.64	1.32	0.53	1.81	1.28
9-Zone	1.52	3.16	1.64	0.69	2.42	1.73
Home Based A	1.62	3.34	1.73	0.66	2.48	1.82
11-Zone	1.66	3.37	1.72	0.79	2.74	1.95
23-Zone	1.93	3.85	1.92	0.59	1.76	1.17
K2						
3-Zone	2.10	2.91	0.81	0.74	1.38	0.63
6-Zone	2.62	3.97	1.35	0.92	2.19	1.27
Home Based A	2.84	4.49	1.65	1.10	2.93	1.83
9-Zone	2.91	4.47	1.56	1.01	2.56	1.55
11-Zone	3.18	4.84	1.66	1.08	2.63	1.55
23-Zone	3.55	5.37	1.82	0.91	2.27	1.36

Table 9 Cohesion with differing choice menus. The first three columns correspond to students' expected # of same-community peers co-assigned conditional on being assigned. The last three correspond to the expected # of same-community peers conditional on being assigned outside of walk-zone. The rows are sorted in increasing baseline cohesion.

During the 2012 school choice reform, many different plans for choice menus were proposed. Some involved subdividing the city into smaller zones (such as the 6-zone, 9-zone, 11-zone, or 23-zone plans) and some involved giving the closest schools of certain types to each family (such as the Home Based A plan). For the purpose of this study we do not go into the details of each plan. These choice menus are documented in Pathak and Shi (2013), BPS (2013), and Shi (2013).[4] Table 9 shows the impact of various menus on baseline and correlated cohesion, as well as the impact on cohesion conditional on traveling outside of walk-zone. The statistics are averages of 25 draws of simulated choice data. (For each draw, the different plans share the same underlying random utility shocks so as to minimize sampling variance in the comparison. This is the simulation approach used in Pathak and Shi (2013).)

As can be seen in Table 9, for K1, correlated lottery beats the cohesion improvement from any menu change. For K2, correlated lottery accomplishes the same effect as tripling the number of zones to 9. For either K1 or K2, for students who are traveling outside of their walk-zone, correlated lottery alone delivers greater increase in cohesion than any choice menu change.

What is most interesting is that the impact of menu change and correlated lottery magnify one another. As can be seen, as the first column (baseline cohesion) increases, the third column (improvement due to correlated lottery) also tends to increase. For example, for K2, while changing to Home Based A (the plan adopted by the city committee) increases cohesion by $2.84 - 2.10 = 0.74$, and correlated lottery alone increases cohesion by $2.91 - 2.10 = 0.81$, doing both at the same time more than doubles cohesion, increasing it by $4.49 - 2.10 = 2.39 > 0.74 + 0.81$.

This phenomenon can be understood in terms of proposition 2. In the large market, the gain in cohesion is the product of the baseline cohesion and a weighted average of SCV (Squared Coefficient of Variation) of

VARIABLES	Improvement in cohesion
Grade	-0.741***
	(0.144)
Uncertainty	6.183***
	(1.609)
Pref. Correlation	2.503***
	(0.354)
Constant	-3.307***
	(1.055)
Observations	18
R-squared	0.931

Table 10 Descriptive regression of improvement in cohesion for new families in various choice menus. This seeks to explain variation in the 18 rows of data in Table 9 in terms of grade, uncertainty of lottery, and preference correlation. "Grade" is dummy for K2. "Uncertainty" is the average individual assignment variance, $E_i[\sum_s p_{is}(1-p_{is})]$. "Preference Correlation" is the average number of common programs in the top-3 lists of two randomly chosen students from the same neighborhood. Robust standard errors are shown in parenthesis. Consistent with Proposition 3, the magnitude of improvement is increasing in uncertainty and preference correlation.

community demand function. The SCV is positively related with within-community preference correlation. When we focus the preferences by reducing the choice menus, both the baseline cohesion and the SCV increase, thus yielding magnified gains. By Proposition 3, this should be additively decomposable into a term for individual lottery variance and a term related to within-community heterogeneity. To check this intuition, we run a regression of the improvement in cohesion from Table 9 on a dummy for grade, a measure of lottery uncertainty, and a measure of within-community preference correlation. The results are in Table 10. As seen, all of the terms are individually significant, and the R- squared is very high (93.1%), supporting the intuition from the theory.

7. Discussion

In this paper, we examine the potential of correlating the school choice lottery to increase community cohesion without affecting individual assignment probabilities. This is desirable as altering individual assignment probabilities inevitably raises questions of equity and access, which are hard to settle without a difficult political process. In contrast, a method that provides cohesion gains without changing assignment probabilities may face less opposition.

In our analysis, we show that while maximizing cohesion is an NP-hard, messy optimization problem in a discrete setting, studying the large-market limit "smooths away" the hardness and yields very clean analysis. Under such a setting, all "reasonable" one-sided many-to-one matching mechanisms can be described as lottery-plus-cutoff. This allows us to decouple the allocation from the lottery correlation, and obtain fairly general results for the potential to improve cohesion by lottery correlation. We can also build a random utility model on top and prove comparative statics, which would be unthinkably difficult if we had remained in the discrete setting.

An empirical finding is that while lottery correlation and reducing choice menus both improve community cohesion, they work best together. For Boston, in the main entry grade K2, correlated lottery alone can improve cohesion for new families by 39%, and choice menu reduction alone can improve it by 30%. But doing both at the same time more than doubles cohesion.

A question for future work is whether lottery correlation can be used to achieve other desirable social outcome. A first attempt might be to increase racial or socio-economic diversity at schools. Intuitively, one may try to "negatively correlate" race or socio-economic status in the lottery. Unfortunately, we show in the Electronic Companion that in the large market, using a correlated lottery produces no gain in diversity, because independent lotteries already achieves the optimum. Significant additional gain in diversity will require altering assignment probabilities.

From an implementation perspective, a potential criticism of our approach is that the lottery becomes less intuitive and transparent. Drawing random lottery numbers and using it to determine priorities is very intuitive, and in the current Boston system, a family may ask the school board for their lottery number. One may imagine a family dissatisfied about their assigned choice may theoretically "audit" the system by checking everyone's lottery number. However, in our approach, there is no longer a simple mapping between lottery numbers and assignment outcomes. Nevertheless, some of the intuitiveness and transparency may be restored if the school board explains the mechanism as a lottery on joint assignments instead of on individual priorities: the school board may print out a large table in which each row denotes a student, and each column denotes a joint assignment. The entry in each cell is the student's assigned school in an assignment. The row counts of each student being assigned to each school would be consistent with the simulated assignment probabilities under the original mechanism. The lottery is then to draw a random column from this table, and a family dissatisfied about the outcome may theoretically "audit" the system by checking this table.

While in this paper we assumed that increasing cohesion is desirable from a policy perspective, this claim warrants further investigation. Although families may in general prefer having more neighboring kids going to the same school so that the children can travel together, share homework help, car-pool on certain occasions, or play together after school, how much they value this and how much this contributes to the students' development are yet to be determined. Moreover, it may be that not everyone wants to be with their neighbors, and for families living in high crime, high-poverty neighborhoods, it is conceivable that some may want the opposite. To accommodate this, one may want to use an opt-in or opt-out system in practice.

Another potential criticism of our approach is that it is unclear how one ought to define "community." In the empirical study, we used a .5 mile by .5 mile grid, and a natural extension would be to optimize the grid size. But doing this optimization requires more in-depth understanding of what community means: too find a grid and the community definition may be too restrictive, too course and we may no longer be capturing "neighborhoods." More broadly, it is unclear if any geographic delineation can adequately capture

the notion of community, since this may be more determined by family background, work proximity, and other factors than home proximity. One way to micro-found our setting is to assume students preferences are lexicographic in the sense that each student first cares to which school she is assigned to, and for each assignment, she prefers to be co-assigned with more students from her own community. To relax these issues one potential is to allow families to define communities for themselves, or even let students report with which peers they would prefer to be assigned with. But this makes preferences for peers private knowledge and raises difficult incentive problems due to the complementarities such preferences introduce. For example, in the closely related matching markets, when couples exist in the market (who have joint preferences) a "stable" assignment may not exist (Klaus and Klijn 2005). One can similarly show that an "envy-free" allocation (with respect to the priorities) in the school choice problem need not exist even when just two students wish to be co-assigned. One direction is to limit the complementarities in the market. Ashlagi et al. (2010) show that as long as the number of couples in the market does not grow too fast, a stable matching exist. Students however may have asymmetric preferences over who they wish to be co-assigned with. Dur and Wiseman (2014) adopts a mechanism design approach that allows students to submit preferences over peers but relaxing the stability notion.[5]

Endnotes

1. Following Azevedo and Leshno (2012), a suitable limit would be any sequence of finite economies with market size going to infinity, but with the proportion of students of each preference and priority, as well as the relative capacities of programs, converging to the large market limit. Both DA-STB and TTC-STB satisfy incentive compatibility and symmetry regardless of market size. Azevedo and Leshno (2012) Theorem 1 and 2 imply that except when capacities fall into some measure-zero set, DA-STB satisfies asymptotic non-atomicity. On the other hand, TTC-STB satisfies asymptotic non-atomicity always, because another student changing her preferences only affects me if her change is affecting the allocation of the "last seat" at a school, which happens with negligible probability. Both mechanisms satisfy efficiency within each priority class because in the large market limit, the distribution of preferences of students of the same priority class is fixed regardless of lottery number, and this precludes trading cycles within each priority class.
2. See endnote 1.
3. In DA-STB, the claims cannot be traded, while in DA-TTC, the claims can.
4. In Pathak and Shi (2013) and Shi (2013), Home Based A is called "ClosestTypes2463."
5. See also Lavy (2012) for empirical findings from a natural experiment in Tel Aviv which allowed students to report also preferences over peers and not just students.

References

Abdulkadiroğlu, A., Y. Che, Y. Yasuda. 2008. Expanding "choice" in school choice. Available at `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1308730`.

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

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

Abdulkadiroğlu, A., T. Sönmez. 2003a. Ordinal efficiency and dominated sets of assignments. *Journal of Economic Theory* **112** 157–172.

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

Asadpour, A., A. Saberi. 2010. An approximation algorithm for max-min fair allocation of indivisible goods. *SIAM Journal on Computing* **39**(7) 2970–2989.

Ashlagi, I., M. Braverman, A. Hassidim. 2010. Stability in large matching markets with complementarities. Forthcoming in Operations Research.

Azevedo, E., E. Budish. 2012. Strategyproofness in the large. Tech. rep., Working paper.

Azevedo, E., J. Leshno. 2010. Can we make school choice more efficient? an example. Tech. rep., mimeo, Harvard University.

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

BPS. 2013. Improving school choice. http://bostonschoolchoice.org/.

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

Budish, E., Y. Che, F. Kojima, P. Milgrom. 2013. Designing random allocation mechanisms: Theory and applications. *The American Economic Review* **103**(2) 585–623.

Burkard, R. E., E. Cela, P. M. Pardalos, L. S. Pitsoulis. 1998. The quadratic assignment problem. Available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.52.3116.

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

Dur, U., T. Wiseman. 2014. School choice with neighbors. Working paper.

Ebbert, S., M. Ulmanu. 2011. A neighborhood scatters for school. *Boston Globe* (12 Jun. 2011).

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

Echenique, Federico, M Bumin Yenmez. 2007. A solution to matching with preferences over colleagues. *Games and Economic Behavior* **59**(1) 46–71.

Ehlers, L. 2010. School choice with control. Cahiers de recherche, Universite de Montreal, Departement de sciences economiques.

Ehlers, L., I. Hafalir, B. Yenmez, M. Yildirim. 2011. School choice with controlled choice constraints: Hard bounds versus soft bounds. GSIA Working Papers 2012-E20, Carnegie Mellon University, Tepper School of Business.

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

Garey, M. R., D. S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness (Series of Books in the Mathematical Sciences)*. First edition ed. W. H. Freeman.

Klaus, B., F. Klijn. 2005. Stable Matchings and Preferences of Couples. *Journal of Economic Theory* **121** 75–106.

Kominers, S. D., T. Sönmez. 2012. Designing for diversity: Matching with slot-specific priorities. Boston College Working Papers in Economics 806, Boston College Department of Economics.

Lavy, V. 2012. How students social networks affect their academic achievement and well-being? Working paper.

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

Mariagiovann, B., A. Ímrohoroğlu, A. J. Wilson, L. Yariv. 2012. A field study on matching with network externalities. *American Economic Review* **102**(5) 1773–1804.

Menino, T. M. 2012. The honorable mayor thomas m. menino state of the city address: January 17, 2012. Working paper.

Miralles, A. 2008. School choice: The case for the boston mechanism.

Pathak, P. A. 2011. The mechanism design approach to student assignment. *Annual Review of Economics* **3**.

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

Pathak, P. A., P. Shi. 2013. Simulating alternative school choice options in boston. Tech. rep., MIT School Effectiveness and Inequality Initiative.

Piantadosi, J., P. Howlett, J. Boland. 2007. Matching the grade correlation coefficient using a copula with maximum disorder. *Journal of Industrial and Management Optimization* **3**(2) 305.

Pycia, M., U. Ünver. 2012. Decomposing random mechanisms. Working paper.

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

Sahni, S., T. Gonzalez. 1976. P-complete approximation problems. *Journal of the ACM* **23**(3) 555–565.

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

Sutherland, A. 2012. To bus or not: Bostons school-choice program. *BU Today* (2012).

Weiwei, H. 2013. Neighborhood interactions and school choices: Evidence from the new york city. Working paper.

Proofs and Additional Results

EC.1. Solving the Correlated-Lottery Implementation Problem

The natural approach to tackling (8) in Section 5 is column generation. Using the language of the simplex algorithm for LP, suppose we are currently at basis B, let $\beta = A_B^{-1} f(A_B)$, the column generation subproblem of finding the deterministic assignment with the highest rate of improvement when we pivot to it (in the sense of the Simplex algorithm) is

$$\begin{aligned} \text{Max} \quad & f(a) - \beta \cdot a \\ \text{s.t.} \quad & a \in \mathcal{P}. \end{aligned} \tag{EC.1}$$

Note that in the above we've relaxed constraint $a \in \text{Vertices}(\mathcal{P})$ to $a \in \mathcal{P}$. We can do this because the objective is convex. In fact it is quadratic as cohesion can be written as

$$f(a) = \frac{1}{n}\left(-n + \sum_{s,c}(\sum_{i \in I_c} a_{is})^2\right).$$

Either the optimum objective of (EC.1) is positive in which case we pivot to the corresponding column a^l representing the optimal solution found, or the optimum is non-positive and we yield a certificate of optimality for (8).

However, this subproblem (EC.1) seeks to maximize a quadratic over the assignment polytope, which in its general case is exactly the notoriously difficult quadratic assignment problem (QAP) (see Burkard et al. (1998)), which is NP-hard to approximate to any constant factor (Sahni and Gonzalez (1976)). One hope is that since our quadratic is of a simpler form, we may still find a polynomial time algorithm. Unfortunately, one can show that the decision problem for both the original (8) and subproblem (EC.1) are NP-complete in the strong sense, even when community sizes are constrained to be 2 or 3. The proof can be done by reduction from Not-All-Equal-3-SAT.

EC.1.1. A practical heuristic

We present an efficient heuristic for correlated-lottery implementation.

The idea of this heuristic is similar to in the Birkhoff-von Neumann theorem. We express the original assignment probabilities as a convex combination of deterministic assignments, by iteratively breaking off deterministic assignments with high cohesion. In each iteration, we find a deterministic assignment with high cohesion that limits to assigning students to schools for which they have positive assignment probability. This ensures that we can break off a positive multiple of this deterministic assignment. To find such a deterministic assignment, we solve a max-weight perfect matching problem on a bipartite graph, in

which one side of the graph are students and the other is schools. We define an edge in this bipartite graph between a student and a school if and only if the student still has positive assignment probability to that school, after subtracting off deterministic assignments found in earlier iterations. The weight of each edge is random, but we constrain the weights to be equal for everyone from the same community to the same school. The intuition of why this may work is that conditional on student i being assigned to school s in the max-weighted matching, the weight between i and s is probably high, so the weight of everyone else from same community as i to s is probably high, so many of them are probably assigned to school s.

To solve the max-weight bipartite perfect matching problem in each iteration, we use a specialized primal-dual implementation with worst case running time $O(n^2m)$, but which is closer to $O(nm^2)$ in practice. (This helps because in our case $n >> m$.) The total number of iterations is at most $\min(mn, T)$, because each iteration reduces the number of non-zero assignment probabilities by at least 1. The number if iterations is also upper-bounded by T because if all assignment probabilities are multiples of $\frac{1}{T}$, then the amount subtracted each time is at least $\frac{1}{T}$. The total running time guarantee is $O(n^2m\min(nm, T))$.

A precise pseudo-code of the algorithm is given below.

Algorithm 1 Heuristic for program 8

Require: $p \in \mathcal{P}$ (assignment probabilities to implement)

$x \leftarrow \emptyset$

while $p \neq \vec{0}$ **do**

$u_{cs} \leftarrow$ random weight, $\forall c \in C, s \in S$

$w_{is} \leftarrow u_{c(i)s}$, $\forall i \in I, s \in S$

$e_{is} = \begin{cases} 1 & \text{if } p_{is} > 0 \\ 0 & \text{otherwise} \end{cases}, \forall i \in I, s \in S$

$a \leftarrow \arg\max_{a'}\{w \cdot a' : a' \in \mathcal{P}, a' \leq e\}$

$\lambda \leftarrow \max\{\lambda' : \frac{p-\lambda a}{1-\lambda} \in \mathcal{P}\}$

$x \leftarrow x \cup (\lambda, a)$

$p \leftarrow p - \lambda a$

end while

return random assignment $x = \{(\lambda_j, a_j)\}$

EC.1.1.1. Explanation for the inner step Another intuition behind our method of finding a deterministic assignment in each iteration with high cohesion is as follows. Let e_{is} be the indicator for whether the assignment probability of i to s is positive. We want to find deterministic assignment $a \leq e$ each time that maximizes cohesion $f(a)$, but since this is NP-hard, we settle for "not-too-bad" solutions. It turns out that

our method of generating community-specific random weights and solving max-weight assignment always avoids a kind of "locally-suboptimal" solutions, in which cohesion can be improved by trading cycles.

To describe our notion of local sub-optimality, we first define trading cycles: Given an assignment a, we say that there is a trading cycle (between communities and schools) $c_0 \to s_0 \to c_1 \to s_1 \cdots \to s_{l-1} \to c_0$ if for each community c_j there is some student $i_j \in c_j$ such that $a_{i_j s_{j-1}} = 1$ and $e_{i_j s_j} = 1$ (where arithmetic on j is modulo l). In other words, by undoing the assignments $i_j \to s_{j-1}$ and re-assigning $i_j \to s_j$, we arrive at another feasible assignment a'. We say that this trading cycle leads a to a'.

Our notion of local sub-optimality occurs in assignment a when both trading cycles $c_0 \to s_0 \to c_1 \to s_1 \cdots \to s_{l-1} \to c_0$ and the reverse $c_0 \leftarrow s_0 \leftarrow c_1 \leftarrow s_1 \cdots \leftarrow s_{l-1} \leftarrow c_0$ exist in a. Suppose the first leads to a' and the second to a'', then we have that the demographic counts, encoding for each community c and school s the number of assignments from c to s, satisfies $d^a = \frac{1}{2}(d^{a'} + d^{a''})$. However, cohesion $f(\cdot)$ can be written as a strictly convex function on d^a, so by convexity $f(a) < \frac{1}{2}(f(a') + f(a''))$ where both a' and a'' are feasible assignments $\leq e$. So a is suboptimal.

However, this situation cannot occur by our method because for any cycle $c_0 \to s_0 \to c_1 \to s_1 \cdots \to s_{l-1} \to c_0$ and any $\{i_j \in c_j\}$, either $\sum_j w_{i_j s_j} - \sum_j w_{i_j s_{j-1}} > 0$ for all such $\{i_j \in c_j\}$ or the reverse for all (since weights from the same community to the same school is defined to be the same, and the difference is non-zero with probability 1 as weights are randomly generated); since a max-weight matching pushes as much as possible along positive weight trading cycles and reverses as much as possible any negative weight trading cycle, the above kind of local non-optimality can never occur in our inner step.

EC.2. Additional empirical results

EC.2.1. Distribution of gains

We examine how the gains in cohesion from correlated lottery are distributed among various neighborhoods. Figure EC.1 shows baseline cohesion and improvement (correlated-baseline) for various geocodes for K1. Figure EC.2 does the same for K2. The plots show that the gains from correlated lottery for K1 is reasonably evenly distributed geographically, with some amounts of "green" throughout the city. On the other hand, the gains for K2 is very uneven, being concentrated in north east corner, south west corner, and a small wedge on the eastern part of the city. This is reflected also in the box plots, as the median improvement for K1 is about 0.6, while the median improvement for K2 is only 0.25. (The mean improvement is roughly 1 in both cases.) This is a disappointing finding as K2 is where the majority of students enter BPS.

How do we make sense of the uneven distribution of improvement in cohesion? The theory in Section 3 suggests that cohesion improvement should benefit those whose lottery is most uncertain (high variance) and whose communities have lowest within-community heterogeneity. Moreover, we expect that within a community, lottery correlation should help the most those whose preferences are most similar to the rest of the community. For ease of interpretation, we define the following proxy: for student i, comparing her

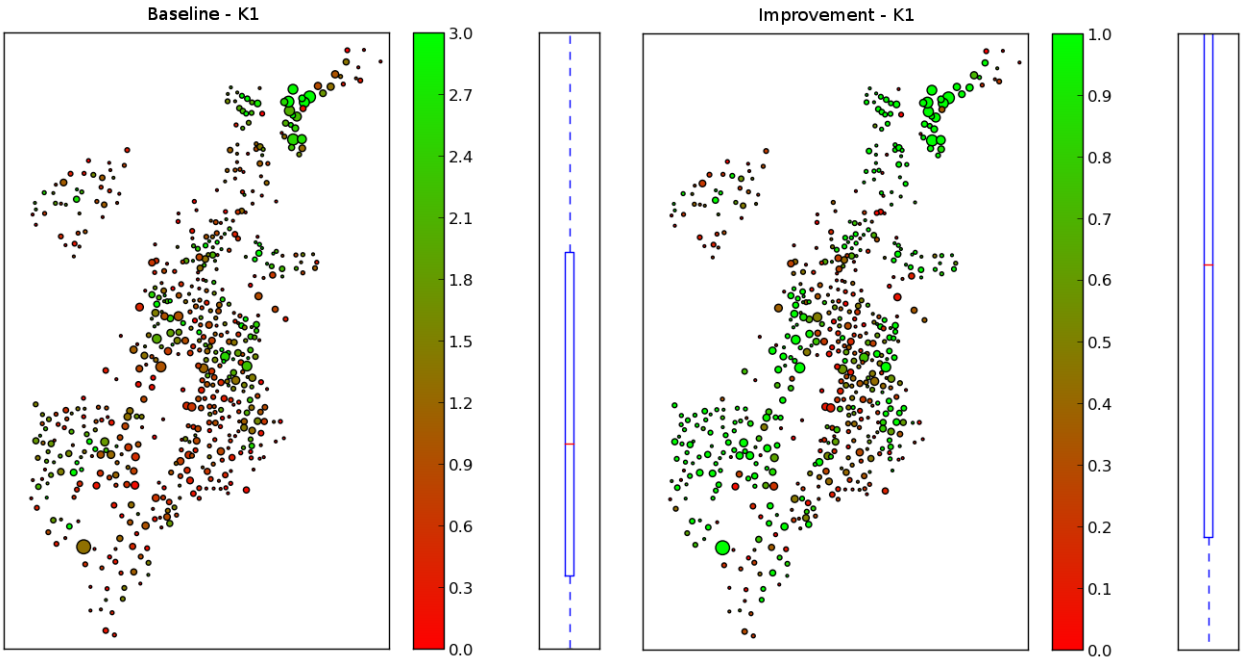

Figure EC.1 Geographic distribution of baseline cohesion and cohesion improvement for K1 new families. Each circle corresponds to a geocode. Its color corresponds to the average value of students living at that location, and its size is proportional to the number of students. The bar on the right of each map shows a box-plot based on the same color scale of the student level distribution, with the ends of the box showing the 25 and 75 percentiles, and the red line shows the median (parts of box plot maybe outside of range to have colors show reasonable contrast). From the left plot, we see that cohesion is highest in East Boston (upper-right island), and varies in different pockets in the city. Improvement in cohesion from correlated-lottery is unevenly distributed, being highest in north east, south west, and various pockets of the city.

with a randomly drawn peer from her community, what is the expected # of programs they have in common among their top 3 choices? We label this proxy choiceAgreement(i).

We plot this for K1 and K2 new families in Figure EC.3. As shown, this metric of within-community preference correlation is about twice higher for K1 than for K2, and the geographic distribution seems to match the distribution of cohesion gains in Figures EC.1 and EC.2. To illustrate this relationship more precisely, we regress individual-level improvement in cohesion on 4 variables: individual lottery variance ($\sum_s p_{is}(1-p_{is})$), scaled choice agreement ($|c(i)|$choiceAgreement(i)), and dummies for whether student i is continuing student or non-continuing sibling. The scaling in the choice agreement is to control for community size. The regression results are tabulated in Table EC.1. As can be seen, much of the variation in gains from cohesion can be explained by uncertainty of lottery and preference correlation and incoming status, with highest gains for new families who face uncertain lottery and whose preferences are highly correlated with their community.

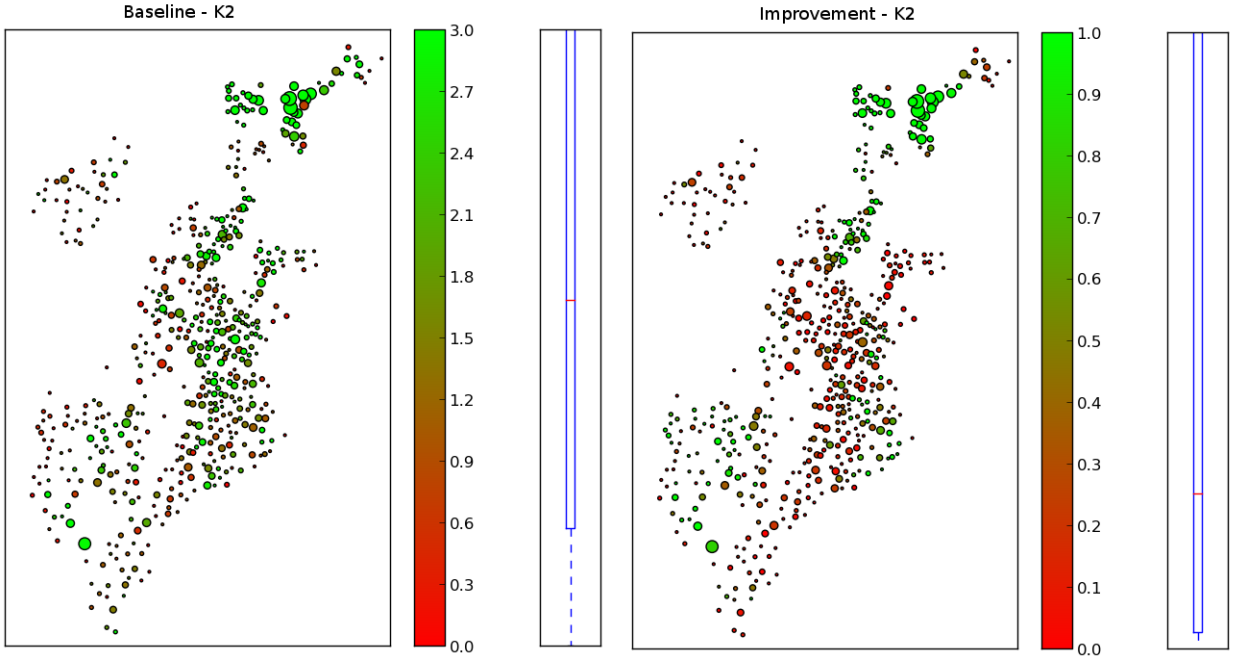

Figure EC.2 Geographic distribution of baseline cohesion and cohesion improvement for K2 new families. From the left plot, we see that cohesion is highest in East Boston (upper-right island), and various pockets in the city. Improvement in cohesion from correlated-lottery is very uneven, concentrated in north east, extreme south west, and eastern part of the city, with very small gains in the center of the city. The median cohesion gain is only about 0.25.

	(1)	(2)
VARIABLES	K1	K2
Uncertainty	0.396***	0.381**
	(0.139)	(0.158)
Pref. Agreement	0.0386***	0.0434***
	(0.00390)	(0.00493)
Non-continuing sibling	-0.823***	-0.726***
	(0.106)	(0.188)
Continuing	-0.394***	-0.161
	(0.0958)	(0.111)
Constant	0.089	-0.153*
	(0.0877)	(0.0863)
Observations	2,654	4,020
R-squared	0.434	0.533

Table EC.1 Descriptive regression of individual improvement in expected # of same-community peers from correlated lottery. Uncertainty of lottery is measured by variance of individual lottery ($\sum_s p_{is}(1-p_{is})$). Preference agreement is proxied by product of size of community and average # of top 3 choice agreements with community ($|c(i)|\text{choiceAgreement}(i)$). Non-continuing sibling and continuing are dummies. The standard errors are from clustering on the 205 communities. As can be seen, students for whom the lottery is more uncertain, whose preferences are highly correlated with their communities, and who are not siblings nor continuing students can expect the greatest cohesion gains from correlated lottery.

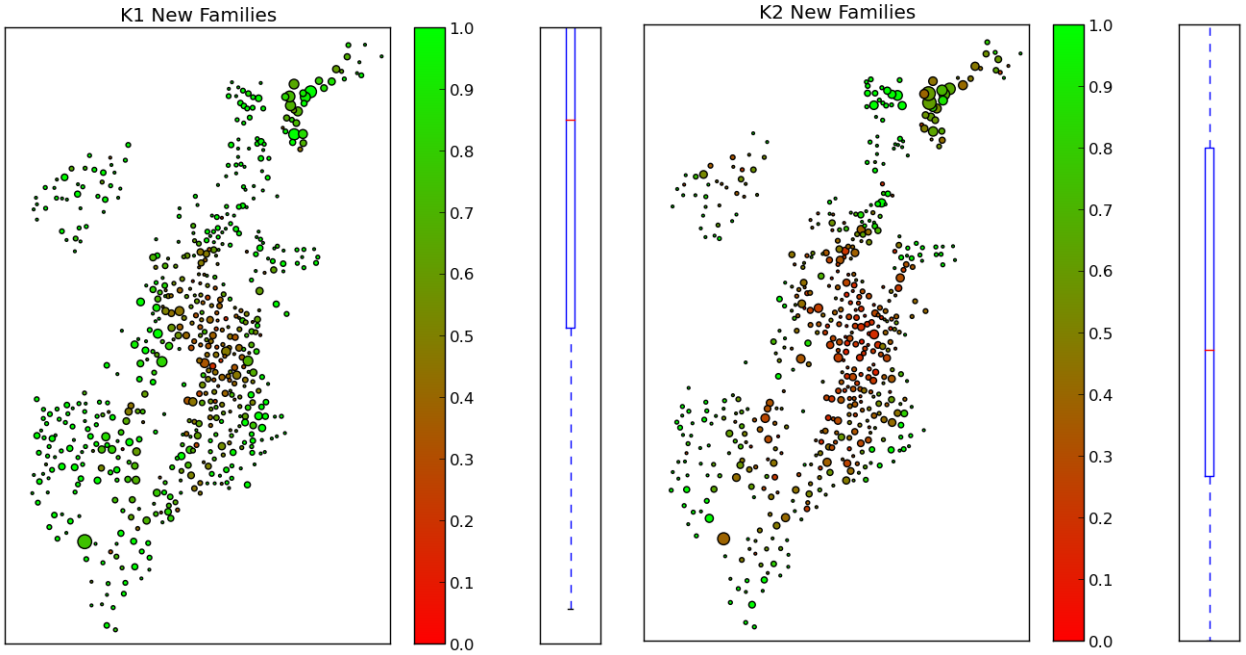

Figure EC.3 Plot of choiceAgreement(i) for K1 and K2 new families. For each student i, this is the average # of programs she has in common in her top 3 choices with an uniformly random peer i' from the same community. As seen, the median is almost twice larger for K1 new families than for K2. Moreover, for K2, this is low in the middle portions of the city, which matches the pattern of low gain from correlated lottery from Figure EC.2.

EC.2.2. Equitable distribution when applying both menu change and correlated lottery

While the cohesion gain from correlated lottery alone is not equitably distributed across the city, it turns out that doing correlated lottery along with a reduced choice menu remedies this inequity. Recall that the choice menu reduction chosen by the city in the 2012-2013 school assignment reform was Home Based A. By applying correlated lottery to Home Based A, we can yield cohesion improvements across all neighborhoods, with 75% of students experiencing a cohesion gain of about 1 or more. This is shown in Figure EC.4, which plots the geographic distribution of baseline cohesion, improvement from correlation alone, improvement from choice menu reduction alone, and improvement from applying both at the same time.

EC.2.3. Impacts on racial and socio-economic diversity

One concern of using a correlated lottery is that it may cause racial or socio-economic segregation, because race and socio-economic status are correlated with geography, so higher chances to go to school with neighbor may entail higher chances to go to school with others of same race or socio-economic status. We check whether this is a cause for concern in Boston.

One practical concern with our approach is that it might harm racial or socio-economic diversity, which historically is a key reason why many school choice systems were created in the first place. However, in the

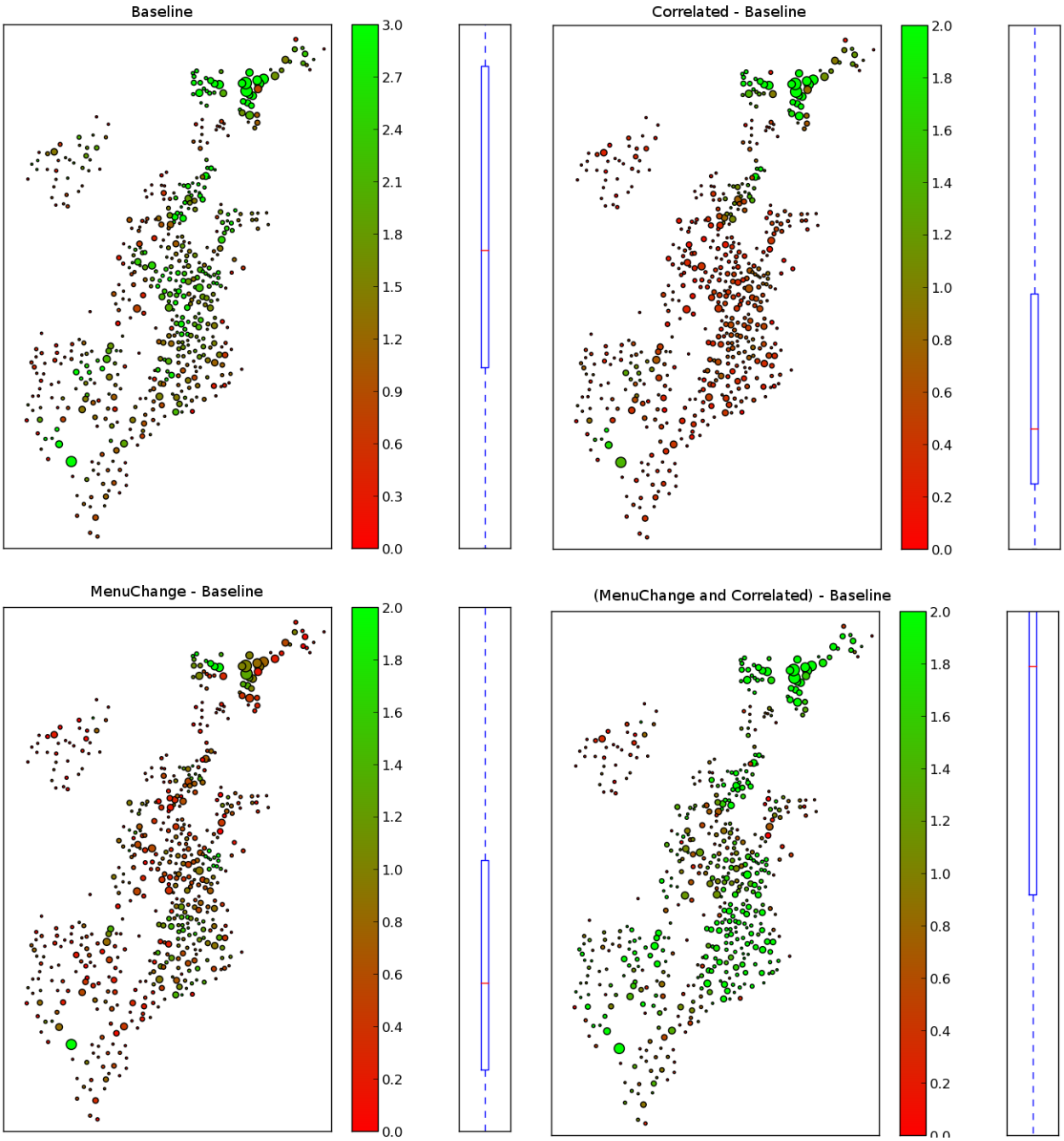

Figure EC.4 Interaction of menu change and correlated lottery for K2 new families. The menu we change to is Home Based A (the one adopted by the city). As can be seen, while the median increase from correlated lottery alone is about 0.4, and the median increase from menu change alone is about 0.6, the median increase from both together is about 1.8. The red region in the middle of the city in the second and third graph effectively disappears in the last graph, showing that large cohesion gain is achieved throughout the city when we apply these reforms together.

data, we find that lottery correlation has minimal impact on diversity. As a measure of diversity, we compute the probability that two random students assigned to the same school are of the same race or socio-economic status.

To do this, we compute the probability for a random pair of students who are assigned to the same school to be of the same race or socio-economic status. The races in our data are black (21%), white (16%), asian (7%), hispanic (43%), other (3%), or missing (10%). The percentages in the parenthesis denote the percentage of applicants of that race counting both K1 and K2. For socio-economic status, we use eligibility to receive free or reduced lunch as proxy. The options are free lunch (49%), reduced lunch (4%), non-free/reduced lunch (12%), and missing (36%). To computing the diversity measure, we go through the 1000 deterministic assignments that form the lottery, cycle through all pairs of distinct students who are assigned to the same program, and count the percentage of times they are of the same race or lunch status. We call these the "peer same race %" and "peer same lunch status %." Table EC.2 shows the results. This is based on actual Round 1 choice data.

	Baseline	Improved
K1		
Peer Same Race %	36.7%	36.9%
Peer Same Lunch Status %	41.6%	41.8%
K2		
Peer Same Race %	42.6%	42.7%
Peer Same Lunch Status %	45.2%	45.2%

Table EC.2 Impact of correlated lottery on racial or socio-economic diversity. For example, when we use independent lottery implementation for K1, on average 36.7% of pairs of students assigned to the same school are of the same race. This percentage increases by only 0.2% to 36.9% when we use correlated lottery.

As can be seen, the impact of correlated lottery on racial or socio-economic diversity is minimal. For example, for K1, the same-race probability increases from 36.7% with independent lottery to 36.9% with correlated lottery; for K2, this increases from 42.6% to 42.7%. The impact is similarly small for same-lunch-status probability. One reason why the impact is so small is that geographic patterns of racial and socio-economic concentration in Boston are at a larger scale than the neighborhood sizes we define (.5 mile by .5 mile squares). Moreover, because choices are heterogeneous to begin with, the impact of lottery correlation is on the order of magnitude of about 1 additional neighbor, which does not significantly impact diversity because class sizes are usually 22 or more.

EC.3. Proofs

Proof of Proposition 1 (NP-hardness of 2 school case): Since there are only 2 schools and since only students in E have non-deterministic assignments, the max-cohesion problem becomes selecting a random q^*-subset of E so that each element of E is selected with probability $p = \frac{q^*}{|E|}$, and the E's stay in their

communities as much as possible. Partition E into communities, defining $E_c = E \cap I_c$. (And similarly $D_c = D \cap I_c$, $F_c = F \cap I_c$.) Let x_i be the binary random variable that equals 1 iff i is assigned to school 1. To simplify notation, we restrict attention to components of x that corresponds to E (because D is assigned to school 1 with probability 1, and E to school 2 with probability 1). Define A to be all elements of $\{0,1\}^{|E|}$ that sum to q^*. The max-cohesion problem becomes:

$$\begin{aligned} f_{\max} = \text{Max} \quad & \Gamma + \left[\sum_{c=1}^{k} \sum_{i \neq j \in E_c} (x_i x_j + (1-x_i)(1-x_j))\right] \\ \text{s.t.} \quad & E[x_i] = p \quad \forall i \in E \\ & x \in \Delta(A), \end{aligned} \tag{EC.2}$$

where the constant $\Gamma = 2\sum_{c=1}^{k}\left[\binom{|D_c|}{2} + \binom{|F_c|}{2} + p|D_c||E_c| + (1-p)|E_c||F_c|\right]$.

An upper-bound on the maximum cohesion is $\Gamma + 2\sum_c \binom{|E_c|}{2}$, which corresponds to $|E|$'s always being assigned with their community members.

It suffices to prove that it is strongly NP-hard to decide whether this upper-bound is achievable. To do this, we reduce from the following version of 3-PARTITION: Given a multi-set of $k = 3h$ integers $\{a_c\}$, for which $\sum_{c=1}^{k} a_c = hB$, and $\frac{B}{4} < a_c < \frac{B}{2}$. Can the a_c's be partitioned so the sum of each partition is B? This is well-known to be strongly NP-hard. A proof is given in Garey and Johnson (1979).

Reduction: given an instance of 3-PARTITION, construct a case of the 2-school max-cohesion problem for which $|E_c| = a_k$, $q^* = B$, then we can achieve perfect cohesion of U iff the $|E_c|$'s can be partitioned so the sum of each partition is exactly q^*, which is what we needed. □

Proof of Theorem 1 (Large Market Characterization): We show that a mechanism is non-atomic, incentive compatible, symmetric and efficient within each priority class, if and only if it is lottery-plus-cutoff.

If a mechanism is lottery-plus-cutoff, then it is non-atomic because any single agent changing preferences does not affect the measure of students in each priority class submitting each ranking, which is what cutoffs depend on. The mechanism is symmetric because within the same priority class, the cutoffs are the same for everyone, so since the lottery numbers are identically distributed, the assignment probabilities are the same for those submitting the same preferences. The mechanism is incentive compatible in school-specific utilities because no individual student can change her cutoff, so for each lottery realization her set of feasible schools is fixed. Submitting a false preference, if it changes her assignment at some lottery number, can only place her in an suboptimal school for those lottery numbers. (This is because with a truthful submission, she would already be assigned her most preferred schools among the feasible set by definition.) The mechanism is efficient within each priority class because if student i_0 prefers school s_1 over s_0 ($s_0 \prec_{i_0} s_1$) and i_0 is assigned to the less-preferred school s_0 with positive probability, then the cutoff for s_1 must be higher:

$z^*_{\pi(i_0),s_0} < z^*_{\pi(i_0),s_1}$, because otherwise there would be no lottery number at which i_0 would select s_0 as her most preferred school. So trading cycles are prohibited by the arithmetic impossibility:

$$z^*_{\pi,s_0} < z^*_{\pi,s_1} < z^*_{\pi,s_2} < \cdots < z^*_{\pi,s_l} < z^*_{\pi,s_0}.$$

Conversely, suppose a mechanism is non-atomic, incentive compatible, and symmetric and efficient within each priority class, we show that it is lottery-plus-cutoff. To do this, we first use non-atomicity to produce a finite set of "representative" students for each priority class, such that every one of the $m!$ possible preference orderings are submitted by someone from the representative set. By symmetry within each priority class, this pins down the assignment probabilities of everyone in that priority class. We then use incentive compatibility and efficiency within each priority class to produce a set of cutoffs, and show that the assignment probabilities of everyone in the representative set is consistent with these cutoffs. This implies that the assignment probabilities from the original mechanism is consistent with a lottery-plus-cutoff description. To show that the mechanism itself is lottery-cutoff, we use the flexibility of arbitrary correlation of lottery numbers to define the lottery numbers so that they produce the exact random assignment of the original mechanism. The fact that the assignment probabilities are consistent with the cutoffs will mean that our lottery numbers can be distributed as Uniform$[0,1)$.

As outlined above, the first step is to produce a finite set "representative" students for each priority class using non-atomicity. For any priority class $\pi \in \Pi$, let the set of students in this class be I_π. Its measure is $\mu(I_\pi) > 0$. Because there are finitely many rankings, there exists some ranking $\succ_0$ that is picked by a positive measure of students in I_π. We choose $m!$ of these students who submitted $\succ_0$ and have them alter their preference submission to each of the $m!$ different possible rankings. By iteratively applying non-atomicity, this does not alter the assignment probabilities of anyone else. By symmetry, the assignment probabilities of each of these $m!$ students is representative of the assignment probabilities of anyone with their preference within the whole priority class. Denote the set of representative students I_{rep}. For any ranking $\succ$, define $p(\succ, s)$ as the assignment probability to school s of the student in I_{rep} corresponding to ranking $\succ$.

The second step is to use incentive compatibility and efficiency within each priority class to produce a set of valid cutoffs. First, note that incentive compatibility in school-specific utilities implies that if any student submits preference list $s_1 \succ s_2 \succ \cdots s_l \succ s_{l+1} \succ \cdots$, reshuffling the relative order of the top l choices among themselves, or reshuffling the later choices among themselves, cannot affect the total assignment probability to $\{s_1, \cdots s_l\}$ (any of the top l choices). To see this, suppose her utilities for top l schools are all $M + \epsilon_s$ and for later schools ϵ_s with $M >> \epsilon_s$ for all s, then the above condition is necessary for her not to have incentive to deviate from truthful preferences at some preference ranking (because for sufficiently large M, the potential to gain a greater probability of getting M trumps any other considerations).

Fix a priority class, focusing on the representative student of this class, we define a complete ordering $>$ over schools, such that $s > s'$ if there exists a student in I_{rep} who prefers the first school s but receives a positive probability of being assigned to the second school s'. The lack of trading cycles among I_{rep} makes this a well-defined ordering. In the case two schools are not comparable directly, we use transitivity to define their relative order. If after taking into account all ordering relations two schools are still not comparable, we can order them arbitrarily. Re-label the schools so that $s_1 > s_2 > s_3 \cdots > s_m$. Call this the over-demand ordering for this priority class. Intuitively, the earlier schools in this ordering are more highly demanded relative to the supply allocated to this priority class. For each j, define "run-out time" t^*_{π,s_j} as the assignment probability to s_j of students in I_{rep} who rank s_j first. (The nomenclature is motivated from the probabilistic serial mechanism.) This is well-defined because of the previous argument from incentive compatibility.

Having defined the "run-out times," the key step is to note that if school s is more over-demanded than school s' according to the cutoff-ordering, then the run-out time of school s must be weakly smaller, that is $t^*_{\pi,s} \leq t^*_{\pi,s'}$. To see this, note that $t^*_{\pi,s}$ is a student's chances of getting into s if she ranks it first, and other schools in arbitrary order. Similarly, $t^*_{\pi,s'}$ is a student's chances of getting into s' if she ranks s' first and s second, and other schools in arbitrary order. Now, if this student reorders her top two rankings to rank s first and s' second, then her total probability of getting into one of these schools must still be $t^*_{\pi,s'}$ by incentive compatibility, but her chance of getting into school s is $t^*_{\pi,s}$, which has to be smaller than the total probability. So $t^*_{\pi,s} \leq t^*_{\pi,s'}$. This means that the run-out times follow a reverse ordering to that of the over-demand ordering of schools, $t^*_{\pi,s_1} \leq t^*_{\pi,s_2} \leq \cdots t^*_{\pi,s_m}$.

We now can define valid cutoffs using these run-out times. Define cutoffs $z^*_{\pi,s} = 1 - t^*_{\pi,s}$. We need to show that everyone's assignment probabilities are consistent with applying lottery-plus-cutoff on these cutoffs. For any subset of schools $U \subseteq S$, let $\min U$ denote the least over-demanded school in U, which by the above paragraph is also the school with the largest run-out time. We observe that if a student in priority class π ranks schools U first in arbitrary order, followed by other schools in arbitrary order, her total assignment probability to some school in U must be equal to $t^*_{\pi,\min U} = \max_{s \in U} t^*_{\pi,s}$. This is because this probability must not depend on her relative ranking within U by incentive compatibility, and so must equal to the probability when she ranks school $\min U$ first, which equals to $t^*_{\pi,\min U}$ because she gets into school $\min U$ with this probability and other schools in U with zero probability since those are more over-demanded.

The above observation allows us to pin down everyone's assignment probabilities to every school, since if a student ranks schools U first, followed by s, followed by other schools, her assignment probability to s must be

$$\max_{s' \in U \cup \{s\}} t^*_{\pi,s'} - \max_{s' \in U} t^*_{\pi,s'}.$$

Therefore, if a student in priority class π submits ranking $\succ$, her assignment probability to school s is equal to

$$p(\succ, s) = \max(0, t^*_{\pi,s} - \max_{s' \succ s} t^*_{\pi,s'}).$$

This can be re-written as

$$p(\succ, s) = \max(0, \min_{s' \succ s} z^*_{\pi,s'} - z^*_{\pi,s}),$$

which is exactly the assignment probability induced by cutoffs $z^*_{\pi,s}$ and lottery numbers $\propto \text{Uniform}[0,1)$.

Having shown that the assignment probabilities are consistent with some set of cutoffs, we now show that by correlating the lottery numbers suitably, we can recover the original random assignment exactly. We do this by reverse construction. For student i in priority class π who submits ranking $\succ_i$, suppose an instantiation of the random assignment x of the original mechanism assigns i to s. Among the schools which i prefers over s, let s' be the one with the lowest cutoff. It must be that the cutoff of s' is higher than s because otherwise the student would have never been assigned to s. We generate

$$(z_i \text{ conditional on } x_i = s) \sim \text{Uniform}[z^*_{\pi,s}, z^*_{\pi,s'}).$$

Since i is assigned to s with probability exactly $z^*_{\pi,s'} - z^*_{\pi,s}$, generated in this way, the unconditional z_i would be distributed Uniform$[0,1)$. The lottery-plus-cutoff mechanism using these lottery numbers z_i, along with the cutoffs $z^*_{\pi,s}$, generates the exact same random assignment as the original mechanism, which is what we needed to show. □

Proof of Theorem 2: Using the cutoff structure, we can express cohesion in both the independent and max-cohesion cases in terms of simple integrals. The theorem follows from re-arranging orders of integration and the law of total variance.

$$\begin{aligned}
f^s_{\text{independent}} &= \int_c \int_y \int_{y'} p_s(c,y) p_s(c,y') dy' dy dc \\
&= \int_c \left(\int_y p_s(c,y) dy \right)^2 dc \\
&= E_c[\bar{p_s}(c)^2] \\
&= E_c^2[\bar{p_s}] + \text{Var}_c(\bar{p_s}) \\
&= q_s^2 + \text{Var}_c(\bar{p_s}).
\end{aligned}$$

Where the last equality follows from our assumption that total supply exactly equals demand.

For max-cohesion, we first note that by giving everyone in the same community the same lottery number, we can achieve cohesion of $E_{c,z}[d_s^2(z|c)]$. This is because at lottery number z, the measure of students from community c that would be assigned to school s is $d_s(z|c)$, so the measure of pairs of students is $d_s^2(z|c)$. Because there is a continuum of communities, this lottery-correlation scheme is feasible, since the total

measure of students that would be assigned to school s remains equal to supply. Hence, maximum cohesion satisfies

$$f^s_{\max} \geq E_{c,z}[d_s^2].$$

On the other hand, the probability of any two students being co-assigned to school s is upper-bounded by $\min(p_s(i), p_s(i'))$. But if i and i' are in the same community, they see the same school cutoffs, because we assumed that everyone in the same community are in the same priority class, so

$$\min(p_s(i), p_s(i')) = \int_z \mathbb{1}_s(z|i)\mathbb{1}_s(z|i')dz.$$

Therefore,

$$\begin{aligned}
f^s_{\max} &\leq \int_c \int_y \int_{y'} \min(p_s(c,y), p_s(c,y'))dy'dydc \\
&= \int_c \int_y \int_{y'} \int_z \mathbb{1}_s(z|(c,y))\mathbb{1}_s(z|(c,y'))dzdy'dydc \\
&= \int_c \int_z \int_y \int_{y'} \mathbb{1}_s(z|(c,y))\mathbb{1}_s(z|(c,y'))dy'dydzdc \\
&= \int_c \int_z \left(\int_y \mathbb{1}_s(z|(c,y))dy \right)^2 dzdc \\
&= \int_c \int_z d_s^2(z|c)dzdc \\
&= E_{c,z}[d_s^2].
\end{aligned}$$

Combining the two, we get that

$$f^s_{\max} = E_{c,z}[d_s^2].$$

Note that the above correlation scheme would not be feasible in a finite market, but our having a continuum of communities allows the variation in demand from this level of extreme correlation to be averaged away so that we do not violate any capacity constraints. In a certain sense, the continuum of communities smooths away the NP-hardness of the max-cohesion problem in the finite case. □

Proof of Proposition 2:

$$\begin{aligned}
f^s_{max} &= \int_c \int_y \int_{y'} \min(p_s(c,y), p_s(c,y'))dy'dydc \\
&= \int_c \int_y \int_{y'} \frac{\max(p_s(c,y), p_s(c,y')) + \min(p_s(c,y), p_s(c,y'))}{2} dy'dydc \\
&- \int_c \int_y \int_{y'} \frac{\max(p_s(c,y), p_s(c,y')) - \min(p_s(c,y), p_s(c,y'))}{2} dy'dydc \\
&= \int_c \int_y p_s(c,y)dydc - \frac{1}{2}\int_c \int_y \int_{y'} |p_s(c,y) - p_s(c,y')|dy'dydc \\
&= q_s - \frac{1}{2}E_c[\bar{\Delta}_s(c)],
\end{aligned}$$

which is what we needed to show. □

Proof of Proposition 3:

$$\begin{aligned}
f^s_{\text{max}} - f^s_{\text{independent}} &= \int_c \int_y \int_{y'} \min(p_s(c,y), p_s(c,y')) - p_s(c,y)p_s(c,y')dy'dydc \\
&= \int_c \int_y \int_{y'} \frac{p_s(c,y)+p_s(c,y')}{2} - \frac{p_s^2(c,y)+p_s^2(c,y')}{2} dy'dydc \\
&- \int_c \int_y \int_{y'} \frac{|p_s(c,y)-p_s(c,y')|}{2} - \frac{(p_s(c,y)-p_s(c,y'))^2}{2} dy'dydc \\
&= E_i[p_s(i)(1-p_s(i))] - \frac{1}{2}E_{c(i)=c(i')}[\Delta_s(i,i')(1-\Delta_s(i,i'))],
\end{aligned}$$

which is what we needed. □

Proof of Proposition 4: The structural identity follows directly from Proposition 2 and Theorem 2.

$$\begin{aligned}
q_s(1-q_s) &= q_s - f^s_{\text{independent}} + \text{Var}_c(\bar{p_s}) \\
&= \frac{1}{2}E_c[\bar{\Delta}_s(c)] + f^s_{\text{max}} - f^s_{\text{independent}} + \text{Var}_c(\bar{p_s}).
\end{aligned}$$

□

Proof of Proposition 5: The first line follows directly from Theorem 2, and observing that there is no between community variation in assignment probabilities (since communities have identical preferences and sizes).

The second line follow from when $\beta \to \infty$, we get the case in which everyone goes to their closest school.

The third and fourth line follows from almost everyone's preferences in the same community become the same as $\alpha \to \infty$ or $\beta \to \infty$, so the demand function $d_s(z|c)$ in any community for any school at any lottery number become close to 1. The result follows from the identity in Theorem 2.

Note that in the above, for independent lottery to achieve perfect community cohesion, we needed school capacities to be just right. No such assumption is needed in the proof for the correlated lottery case. □

Proof of Theorem 3: If $\beta = 0$, then there cannot be between community variation in assignment probabilities, so by Theorem 2, $f^s_{\text{independent}} = q_s^2$.

Label the schools so $r_1 > r_2 > \cdots$. (Recall that $r_s = e^{\nu_s}$.) Because of the logit utility structure, in any community, the probability that school s is chosen first is

$$\frac{r_s}{\sum_{j=1}^m r_j}.$$

Because we assumed no priority, the mechanism is equivalent to probabilistic serial (or lottery-plus-cutoff with a single priority class). Define run-out time $t^*_s = 1 - z^*_s$. The cutoff for the most highly demanded school, school 1, is such that

$$t^*_1 = q_1(\frac{\sum_{j=1}^m r_j}{r_1}).$$

By the independence of irrelevant alternative properties of logit, these students, given a lower lottery number, would choose school s with probability $\frac{r_s}{\sum_{j=2}^m r_j}$. The cutoff for the 2nd most demanded school is such that

$$\begin{aligned} t_2^* - t_1^* &= \left(q_2 - t_1^* \frac{r_2}{\sum_j^m r_j} \right) \frac{\sum_{j=2}^m r_j}{r_2} \\ &= \left(\frac{q_2}{r_2} - \frac{q_1}{r_1} \right) \sum_{j=2}^m r_j. \end{aligned}$$

Define tail sum $R_s = \sum_{j=s}^m r_s$. Continuing in this way by induction we gets

$$t_s^* = \sum_{j=1}^{s-1} q_j + \frac{R_s}{r_s} q_s,$$

and

$$t_s^* - t_{s-1}^* = \left(\frac{q_s}{r_s} - \frac{q_{s-1}}{r_{s-1}} \right) R_s,$$

which implies the formula for max-cohesion,

$$f_{\max}^s = \sum_{k=1}^s (t_k^* - t_{k-1}^*)(\frac{r_s}{R_k})^2 = r_s^2 \sum_{k=1}^s \left(\frac{q_k}{r_k} - \frac{q_{k-1}}{r_{k-1}} \right) \frac{1}{R_k}.$$

Since increasing α increases the ratios $\frac{r_1}{r_2}, \frac{r_2}{r_3}, \cdots, \frac{r_{m-1}}{r_m}$, it suffices to show that if one of these ratios increase, while the others stay the same, maximum cohesion increases. In other words, it suffices to show that for any l, if we perform the transformation

$$\begin{gathered} r_1, r_2, \cdots, r_l, r_{l+1}, \cdots, r_m \\ \downarrow \\ \gamma r_1, \gamma r_2, \cdots, \gamma r_l, r_{l+1}, \cdots, r_m \end{gathered}$$

with $\gamma > 1$, we increase maximum cohesion $f_{\max}^s$.

Now, for $l \leq s-1$, we have

$$\begin{aligned} f_{\max}^s &= r_s^2 \sum_{k=1}^s \left(\frac{q_k}{r_k} - \frac{q_{k-1}}{r_{k-1}} \right) \frac{1}{R_k} \\ &= \frac{q_s r_s}{R_s} - r_s^2 \sum_{k=1}^{s-1} \frac{q_k}{R_k R_{k+1}}. \end{aligned}$$

so $f_{\max}^s$ increase because the transformation fixes r_s, R_s and the q's, while only increasing some of the R_k's in the denominator of the negative term. (Recall that $R_k = \sum_{j=k}^m r_s$.)

For $l \geq s$, we write $f_{\max}^s$ in a different way

$$\begin{aligned} f_{\max}^s &= r_s^2 \sum_{k=1}^s \left(\frac{q_k}{r_k} - \frac{q_{k-1}}{r_{k-1}} \right) \frac{1}{R_k} \\ &= \sum_{k=1}^s \left[q_k \frac{r_s}{r_k} - q_{k-1} \frac{r_s}{r_{k-1}} \right] \frac{r_s}{R_k}. \end{aligned}$$

so $f^s_{\max}$ also increases by the transformation, since each of $\frac{r_s}{r_k}$ and $\frac{r_s}{r_{k-1}}$ stays fixed, while $\frac{r_s}{R_k}$ increase because the numerator increase by a factor of γ while the denominator $R_k = \sum_{j=k}^{m} r_j$ increase by a factor less than or equal to γ (some of the summands increase by factor γ while others stay the same). This completes the proof. □

EC.4. Detailed analysis with two schools

Continuing from the notation in Section 2.1 and proof of Proposition 1, in the 2 school case a mechanism using independent lotteries simply chooses a q^*-subset of E uniformly randomly, and assign them to school 1. We call the cohesion obtained by independent implementation of the lottery $f_{\text{independent}}$. It is straightforward to work out

$$f_{\text{independent}} = \Gamma + \left(1 - \frac{2q^*(|E| - q^*)}{|E|(|E|-1)}\right)(2\sum_c \binom{|E|}{2}) \tag{EC.3}$$

We want to bound the ratio $\frac{f_{\max}}{f_{\text{independent}}}$, which is how much correlated lottery improves cohesion in this setting. However, $f_{\max}$ is computationally hard to solve.

However, we can come up with fairly good bounds on $f_{\max}$ when the number of communities is large. Imagine that we arrange students in E in a circle, with members of the same community being adjacent to one another. We can randomly split this circle into a group of q^* and a group of $|E| - q^*$ by uniformly randomly selecting a student and counting q^* students clockwise from her. This keeps communities together except for at most 2 communities. The expected number of same-community pairs split is at most $\frac{2\sum_c |E_c|\left(\frac{1}{|E_c|}\sum_{a=1}^{|E_c|} a(|E_c|-a)\right)}{\sum_c |E_c|} = \frac{2\sum_c |E_c|\frac{(|E_c|-1)(|E_c|+1)}{6}}{\sum_c |E_c|} \le \frac{2\sum_c |E_c|\binom{|E_c|}{2}}{\sum_c |E_c|} \le \max_c 2\binom{|E_c|}{2}$.

So this randomization, while not necessarily keeping all communities together, gives up at most the largest community. Thus,

$$\Gamma + 2\sum_c \binom{|E_c|}{2} - 2\max_c \binom{|E_c|}{2} \le f(\max) \le \Gamma + 2\sum_c \binom{|E_c|}{2} \tag{EC.4}$$

Combining this with previous expression for $f_{\text{independent}}$,

$$1 \le \frac{f_{\max}}{f_{\text{independent}}} \le \frac{\Gamma + 2\sum_c \binom{|E_c|}{2}}{\Gamma + (1 - \frac{2q^*(|E|-q^*)}{|E|(|E|-1)})(2\sum_c \binom{|E|}{2})} \tag{EC.5}$$

When the number of communities k is large and $|E_c|$'s are approximately equal, $\frac{\max_c \binom{|E_c|}{2}}{\sum \binom{|E_c|}{2}} \to 0$, so the upper-bound is asymptotically achievable using the lottery correlation scheme defined above.

The maximum improvement occurs if Γ is small, and $\frac{2q^*(|E|-q^*)}{|E|(|E|-1)} \approx p(1-p)$ is large, in which case the improvement ratio $\frac{f_{\max}}{f_{\text{independent}}} \approx \frac{1}{1-\frac{2q^*(|E|-q^*)}{|E|(|E|-1)}} \le 2 + \frac{2}{|E|-2}$. So the maximum possible improvement from correlated lottery is about a factor of 2. (This is achieved if for example there are no priorities and the lottery is most random, so $\Gamma = 0, p = \frac{1}{2}$, in which case an independent lottery splits any given same-community pair about $\frac{1}{2}$ the time, and a correlated lottery, using the scheme described above, keeps communities together almost all the time.)

EC.4.1. Large market approximation with two schools

The NP-hardness in the finite case precludes precisely pinning down max-cohesion f_{max}, but as seen in the above this difficulty disappears when the number of communities is large. We examine this "large-market" case more closely to yield more intuition on when exactly we can expect correlated lottery to improve cohesion the most.

We approximate the expressions for $f_{\text{independent}}$ and f_{max} in Equations EC.3 and EC.4 in this large market case. For each community c, define $n_c = |I_c|$ and define u_c, v_c, w_c to be the fraction of this community in sets D, E, F respectively. So $u_c = \frac{|D_c|}{|n_c|}$, $v_c = \frac{|E_c|}{|n_c|}$, $w_c = \frac{|F_c|}{|n_c|}$, and we can approximate $2\binom{|D_c|}{2} \approx n_c^2 u_c^2$, $2\binom{|E_c|}{2} \approx n_c^2 v_c^2$ and $2\binom{|F_c|}{2} \approx n_c^2 w_c^2$. Moreover, we approximate the term in Equation EC.3, $\frac{q^*(|E|-q^*)}{|E|(|E|-1)} \approx \frac{q^*(|E|-q^*)}{|E|^2} = p(1-p)$. To simplify notation, treat c as a random variable that takes a particular value c_0 with probability $\frac{n_{c_0}^2}{\sum_{c'} n_{c'}^2}$. Rescale cohesion to be between 0 and 1 by defining $\tilde{f} = \frac{f}{\sum_{c'} n_{c'}^2}$. We have

$$\begin{aligned}
\tilde{f}_{\text{independent}} \cong & E[u_c^2 + v_c^2 + 2pu_c v_c + 2(1-p)v_c w_c + (1-2p(1-p))v_c^2] \\
= & E[(u_c + pv_c)^2] + E[(w_c + (1-p)v_c)^2] \\
= & E^2[u_c + pv_c] + E^2[w_c + (1-p)v_c)] + Var(u_c + pv_c) + Var(w_c + (1-p)v_c) \\
= & \bar{q_1}^2 + \bar{q_2}^2 + 2Var(u_c + pv_c).
\end{aligned}$$

Where $\bar{q_1} = \frac{q_1}{n}$, $\bar{q_2} = \frac{q_2}{n}$ are the proportion of seats that are in school 1 and 2 respectively. Moreover, when the number of communities is large and when the relative size of the largest community is small, we have that the upper-bound in Equation EC.4 becomes equality, so

$$\begin{aligned}
\tilde{f}_{\text{max}} \cong & E[u_c^2 + v_c^2 + 2pu_c v_c + 2(1-p)v_c w_c + v_c^2] \\
\cong & \tilde{f}_{\text{independent}} + 2E[p(1-p)v_c^2].
\end{aligned}$$

We can interpret the terms as follows:

- $\bar{q_1}^2 + \bar{q_2}^2$: Herfindahl index of school size. This is a measure of school size concentration. It is higher when school sizes are more unequal.
- $2Var(u_c + pv_c)$: Between community variation in assignment probabilities. $u_c + pv_c$ is exactly the expected proportion of people from community c who go to school 1. This variation is 0 if assignment probabilities in communities are identical, and increases if preferences and/or priorities result in more varied proportion of people assigned to each school in different communities.
- $E[2p(1-p)v_c^2]$: How much independent lottery hurts cohesion. This is the expected number of same community pairs split up by independent lottery. It is higher when communities have more people affected by the lottery ($|E|$ is higher) and when the lottery is more "uncertain" (closer to $p = \frac{1}{2}$). For fixed p, it is higher when the variation across communities of the number of people affected by the lottery ($Var(v_c)$) is higher (or equivalently when within-community heterogeneity is low).

An interesting special case is when there are no priorities and school sizes equal, in which the above expressions immediately yield.

PROPOSITION EC.1. *When there are no priorities ($u_c = 0$) and $q_1 = q_2 = \frac{1}{2}$, we have that the proportional improvement from correlated lottery is*

$$\frac{\tilde{f}_{max} - \tilde{f}_{independent}}{\tilde{f}_{independent}} = \frac{1-p}{p} = 2E[v_c] - 1,$$

which can be interpreted as a measure of overall preference correlation or competition.

Proof of Proposition EC.1: When $u_c = 0$ and $q_1 = q_2 = \frac{1}{2}$, we have $p = \frac{1}{2E[v_c]}$, so

$$\begin{aligned}
\tilde{f}_{\text{independent}} &= E[(u_c + pv_c)^2] + E[(w_c + (1-p)v_c)^2] \\
&= p^2 E[v_c^2] + E[(1 - pv_c)^2] \\
&= 2p^2 E[v_c^2] + 1 - 2pE[v_c] \\
&= 2p^2 E[v_c^2].
\end{aligned}$$

So

$$\begin{aligned}
\frac{\tilde{f}_{\text{max}} - \tilde{f}_{\text{independent}}}{\tilde{f}_{\text{independent}}} &= \frac{2p(1-p)E[v_c^2]}{2p^2 E[v_c^2]} \\
&= \frac{1-p}{p}.
\end{aligned}$$

□

Based on this, we gather the following qualitative insights: to have higher cohesion when the lottery is implemented independently, one needs to have more varied school sizes, more between-community variations in preferences, or have the over-demanded school give more priority to communities that most demand it. Furthermore, the potential to improve cohesion via correlated lottery is greater when more people are affected by the lottery, when within-community preference correlation is high, or when the lottery is more uncertain. When school sizes are equal, the cohesion improvement is greater when more people desire the over-demanded school.

EC.5. Inability of lottery correlation to increase diversity

Besides community cohesion, another desirable outcome might be racial or socio-economic diversity. Unfortunately, we show in this section that running an independent lottery already achieves near optimal diversity, so any significant gain requires changing the assignment probabilities.

To see why this is true, consider the assignment of students of different races to a school of capacity q. Suppose there are r different races. In the random assignment, let $n_1, n_2, \cdots n_r$ be random variables denoting the number of students of each race assigned to the school. Suppose that the school is always

assigned at capacity, so $\sum_{j=1}^{r} n_j = q$. A metric for diversity is the number of pairs of students of different races assigned to the same school, so maximizing diversity at this school is to maximize

$$E[\sum_{j_1 \neq j_2} n_{j_1} n_{j_2}].$$

This is the same as minimizing

$$q^2 - 2E[\sum_{j_1 \neq j_2} n_{j_1} n_{j_2}] = E[\sum_{j=1}^{n} n_j^2].$$

However, by Jensen's inequality,

$$E[\sum_{j=1}^{n} n_j^2] \geq \sum_{j=1}^{n} E[n_j]^2.$$

This lower bound is approximately achieved by running an independent lottery, because in that case n_j is almost always close to $E[n_j]$ by Chernoff bound. In fact, in the large market model of Section 3, n_j would be equal to $E[n_j]$ always, and this lowerbound is exactly achieved by independent lottery implementation.

In a finite market, one can achieve small diversity gains by constraining n_j to be always between $\lfloor E[n_j] \rfloor$ and $\lceil E[n_j] \rceil$. More precisely, one would express the original assignment probabilities as a convex combination of such constrained deterministic assignments. This idea has been previously explored in Budish et al. (2013). However, by the above argument, one would expect gains in diversity to be small.