Ortega, Josué; Ziegler, Gabriel; Arribillaga, R. Pablo; Zhao, Geng

Working Paper

Identifying and Quantifying (Un)Improvable Students

QBS Working Paper, No. 2025/05

Provided in Cooperation with:
Queen's University Belfast, Queen's Business School

Suggested Citation: Ortega, Josué; Ziegler, Gabriel; Arribillaga, R. Pablo; Zhao, Geng (2025) : Identifying and Quantifying (Un)Improvable Students, QBS Working Paper, No. 2025/05, Queen's University Belfast, Queen's Business School, Belfast

This Version is available at:
https://hdl.handle.net/10419/318333

QUEEN'S UNIVERSITY BELFAST
QUEEN'S BUSINESS SCHOOL

Working Paper Series - QBS Working Paper 2025/05

Identifying and Quantifying (Un)Improvable Students

Josué Ortega
Queen's University Belfast

Gabriel Ziegler
University of Edinburgh

R. Pablo Arribillaga
Instituto de Matemática Aplicada San Luis

Geng Zhao
University of California, Berkeley

26 May 2025

Series edited by Philip T. Fliers and Louise Moss.
To submit forward your paper to qbs.rps@qub.ac.uk.

Identifying and Quantifying (Un)Improvable Students

JOSUÉ ORTEGA
Queen's University Belfast

GABRIEL ZIEGLER
University of Edinburgh

R. PABLO ARRIBILLAGA
Instituto de Matemática Aplicada San Luis, Universidad Nacional de San Luis, and CONICET

GENG ZHAO
University of California, Berkeley

Abstract

The Deferred Acceptance (DA) mechanism can generate inefficient placements. Although Pareto-dominant mechanisms exist, it remains unclear which and how many students could improve their DA assignment. We characterize the set of unimprovable students and show that it includes those unassigned or matched with their least preferred schools. Nevertheless, by proving that in large markets DA's envy digraph contains a unique giant strongly connected component, we establish that almost all students are improvable, and furthermore, they can benefit simultaneously via disjoint trading cycles. Our findings reveal both the pervasiveness of DA's inefficiency and the remarkable effectiveness of Pareto-dominant mechanisms in addressing it, regardless of the specific mechanism chosen.

KEYWORDS. unimprovable students, school choice, random markets.

JEL CLASSIFICATION. C78, D47.

Emails: j.ortega@qub.ac.uk, ziegler@ed.ac.uk, rarribi@unsl.edu.ar, gengzhao@berkeley.edu.
We acknowledge helpful comments from Mariagiovanna Baccara, Li Chen, David Delacrétaz, Battal Doğan, Matt Elliott, Jörgen Kratz, SangMok Lee, Vikram Manjunath, Martin Meier, Alexander Nesterov, Szilvia Pápai, Fedor Sandomirskiy, Qianfeng Tang, Olivier Tercieux, Bertan Turhan, M. Utku Ünver, M. Bumin Yenmez, Alexander Westkamp and audiences at seminars at the Universities of Manchester, Moscow HSE, Naples, San Luis, Washington University in St. Louis, York, ZEW Mannheim and the SAET conference in Santiago, the Matching in Practice workshop in Zurich, the European Winter Meeting of the Econometric Society in Mallorca, the Royal Economic Society Conference in Belfast, the WINE conference in Shanghai, the Irish Economic Theory workshop, the Matching to Markets Workshop in Cargèse and the Lancaster game theory conference. *16/05/2025.*

1. Introduction

The student-proposing Deferred Acceptance (DA) algorithm is widely used to assign pupils to schools because it produces a stable allocation and incentivizes parents to rank schools truthfully. However, DA can result in outcomes that are Pareto-inefficient for students; in some cases, it may assign every student to one of their least-preferred schools (Kesten, 2010). Its inefficiencies have also been documented empirically. For instance, Abdulkadiroğlu et al. (2009) show that approximately 2% of students in the New York City high school match—which uses a version of DA—could have been assigned to a more preferred school without harming others.

To address this inefficiency while preserving weaker forms of stability and strategy-proofness, several mechanisms that Pareto dominate DA have been proposed. One such mechanism—Kesten's Efficiency-Adjusted Deferred Acceptance (EADA)—has shown exceptional promise in both theory and the lab. Yet empirical work suggests that EADA's efficiency gains may be unevenly distributed (see Section 2). Consequently, a natural theoretical question emerges: *which and how many students can improve their DA placement under EADA, and more generally, under efficient mechanisms that Pareto dominate DA?*

We begin by identifying *which* students are improvable. Our analysis introduces a novel analytical framework: DA's *envy digraph*—a directed graph in which nodes represent students who point to each other if they are envious of their DA assignment, irrespective of whether the envy is justified. Proposition 1 shows that a student is improvable if and only if they lie on a trading cycle in this graph. This characterization allows us to identify improvable students using a graphical approach and directly implies that sources and sink nodes in the envy digraph are unimprovable. While sink nodes correspond to those assigned to their top choice under DA (who trivially envy nobody), source nodes (those whom nobody envies) always include every unassigned student or those matched to their least preferred choice (Proposition 2). This finding has an important policy implication: students who are poorly assigned or unassigned cannot be helped by any mechanism that preserves DA's outcome as a baseline.

Next, we focus on *how many* students are unimprovable. Using the probabilistic framework of random matching markets with uniformly and independently distributed preferences, we quantify students who nobody envies (Proposition 3) and who envy nobody (Proposition 4). The proofs establish novel connections between matching and classical problems in probability—specifically, the distribution of singleton coupons in

the coupon collector problem and the rank distribution in the stable marriage problem. While valuable in their own right, these findings show that the proportion of students who are unimprovable through these easily identifiable channels converges to zero as the market grows large.

Our main result, Theorem 1, extends this intuition by proving that the proportion of unimprovable students becomes arbitrarily small as market size grows large. The proof leverages the density of DA's envy graph, where most students make and receive a logarithmic number of applications. This density implies that the probability of two sizeable groups of students not pointing to each other becomes vanishingly small in large markets. We show that the envy relations in school choice markets under DA exhibit a characteristic property of random directed graphs—the emergence of a unique giant strongly connected component where most nodes participate in at least one cycle.

Although DA's inefficiency was previously known, Theorem 1 reveals that its pervasiveness is not confined to pathological examples or specific datasets but emerges as a systematic, expected property in random markets. The consequences are so severe that the vast majority of students could find ways to improve their placement without harming others. Combined with Proposition 2, this result has important implications: while the most disadvantaged students cannot benefit from enhanced efficiency, the vast majority of students are improvable, highlighting the scope and necessity of welfare enhancements over DA.

Our final result, Theorem 2, addresses a related policy question: which efficient mechanism dominating DA should be chosen to maximize the number of improved students? By connecting student improvability to the classical cycle-packing problem in graph theory, we prove that in large markets, any mechanism that Pareto dominates DA will improve approximately the same number of students—almost all of them. This surprising asymptotic equivalence occurs because the high density of DA's envy digraph ensures that if a student is not included in one particular trading cycle, they will likely participate in an alternative trade. This result has important implications for market design: while different mechanisms that Pareto dominate DA may perform quite differently in small markets (Knipe and Ortega, 2025), they are virtually indistinguishable in terms of size of improvement in large markets. Thus, policymakers can focus on other desirable properties—such as minimal manipulation incentives or weak stability, both areas in which EADA excels—when selecting among mechanisms that Pareto dominate DA.

In summary, our contributions are threefold. First, we introduce the novel envy digraph framework that provides a precise characterization of improvable students in school choice. Second, we prove that almost all students are improvable in large markets, revealing the pervasive inefficiency of DA and the substantial scope for welfare improvements. Third, we establish the asymptotic equivalence of all efficient mechanisms that Pareto dominate DA, demonstrating that the choice among such mechanisms becomes increasingly irrelevant as markets grow.

2. Related Literature

Our work contributes to several strands of literature on matching mechanisms and random market models. We organize our discussion around four interconnected themes.

On DA's Pareto-inefficiency. The Pareto-inefficiency of DA (for students) is well-known (Abdulkadiroğlu and Sönmez, 2003). Theoretical bounds on DA's inefficiency were established by Kesten (2010). Numerous empirical studies have documented this inefficiency in practice (Abdulkadiroglu et al., 2005, Abdulkadiroğlu et al., 2009, Che and Tercieux, 2019, Ortega and Klein, 2023).

Efficient mechanisms that dominate DA. The most prominent mechanism in this class is Efficiency-Adjusted Deferred Acceptance (EADA, Kesten, 2010). Over the past decade, EADA's properties and implementation have been extensively studied (Bando, 2014, Tang and Yu, 2014, Dur et al., 2019, Troyan et al., 2020, Troyan and Morrill, 2020, Ehlers and Morrill, 2020, Tang and Zhang, 2021, Doğan and Ehlers, 2021, Reny, 2022, Chen and Möller, 2023), demonstrating that it is possible to achieve an efficient improvement over DA while maintaining relatively low instability and manipulability. Another well-known efficient mechanism that Pareto-dominates DA is DA+TTC, which applies the top trading cycles (TTC) procedure to the allocation obtained by DA (Alcalde and Romero-Medina, 2017).

Two seminal papers have studied the family of mechanisms that Pareto dominate DA: Alva and Manjunath (2019) provide a comprehensive study of stable-dominating rules, which include any efficient mechanism that improves on student-proposing DA, and Tang and Yu (2014), who introduce the concept of improvable students and underdemanded schools. This paper is heavily influenced by their work, which we use as

building blocks for our results. However, these studies focus on identifying improvable students without quantifying their prevalence—a gap we fill by deriving asymptotic bounds.

Empirical evaluations of mechanisms that dominate DA. Diebold and Bichler (2017) conduct an extensive comparison across course allocation datasets, finding that the difference between DA and EADA rank distributions is not statistically significant in 7 out of 19 datasets. They also show that EADA and DA always match the same number of students. Ortega and Klein (2023) compute counterfactual EADA allocations for Budapest's school choice system and find that, while EADA improves the average student placement, it does not help unassigned students or the worst-placed pupil (assigned to his 13th ranked school under both mechanisms) or those unassigned. In laboratory experiments, Cerrone et al. (2024) find that EADA generates Pareto-efficient allocations more frequently than DA in the lab but note that the distribution of efficiency improvements is uneven among participants.

Random matching markets. Our analysis of the proportion of unimprovable students builds on a rich literature on random matching markets at the intersection of economics and computer science (Wilson, 1972, Knuth, 1976, Pittel, 1989, 1992, Immorlica and Mahdian, 2005, Kojima and Pathak, 2009, Che and Kojima, 2010, Lee, 2016, Liu and Pycia, 2016, Ashlagi et al., 2017, Che and Tercieux, 2018, 2019, Pycia, 2019, Ashlagi et al., 2021, 2023, Nikzad, 2022, Ortega and Klein, 2023, Ronen et al., 2025). Our paper advances this literature by providing a characterization and quantification of unimprovable students. Finally, our Theorem 2, establishing the equivalence of all efficient mechanisms that Pareto dominate DA with regard to the number of students improved, has some resemblance to a number of equivalence theorems for large matching markets by Che and Kojima (2010), Liu and Pycia (2016), Pycia (2019) and Che and Tercieux (2018). Among these, Pycia (2019)'s result bears the closest resemblance to our finding. However, since we do not focus on strategy-proof mechanisms, his result does not subsume our Theorem 2.

3. Model

Following Abdulkadiroğlu and Sönmez (2003), a school choice problem P consists of a finite set of students I and a finite set of schools S. Each student i has a strict preference $\succ_i$ over the schools. Each school s has a quota of available seats q_s and a strict

priority over the students $\triangleright_s$, determined by local educational regulations. To allow for unassigned students, we allow the existence of a null school (denoted $s_\emptyset$), which has unlimited capacity. We use $n = |S|$ to denote the number of schools.

For a given school choice problem P, a *matching* μ is a mapping from I to S such that no school is matched to more students than its quota. We denote by μ_i the school to which student i is assigned and by μ_s^{-1} the set of students assigned to school s. With this notation, a matching needs to satisfy $|\mu_s^{-1}| \leq q_s$ for every $s \in S$. We call every student i with $\mu_i = s_\emptyset$ *unassigned*.

The function $\text{rk}_i : S \to \{1, \ldots, n\}$ specifies the rank of school s according to the preference profile $\succ_i$ of student i:

$$\text{rk}_i(s) = |\{s' \in S : s' \succ_i s\}| + 1,$$

so that the most desirable option gets a rank of 1, whereas the least desirable gets a rank of n. With some abuse of notation, we use the same rank function to specify the students' rank per the priority profile of schools, i.e. $\text{rk}_s(i) = |\{j \in I : i \triangleright_s j\}| + 1$.

A matching μ *weakly Pareto-dominates* matching ν if, for every student $i \in I$, $\text{rk}_i(\mu_i) \leq \text{rk}_i(\nu_i)$. A matching μ *Pareto-dominates* matching ν if μ weakly dominates ν and there exists a student $j \in I$ with $\text{rk}_j(\mu_j) < \text{rk}_j(\nu_j)$. A matching is *Pareto-dominated* if there exists a matching that Pareto-dominates it and is *Pareto-efficient* if it is not Pareto-dominated.

Student i *desires* school s in matching μ if $\text{rk}_i(s) < \text{rk}_i(\mu_i)$ and he *envies* student j at matching μ if $\text{rk}_i(\mu_j) < \text{rk}_i(\mu_i)$. We say that student j *violates* student i's priority at school s in matching μ if i desires s, $\mu_j = s$, and $\text{rk}_s(i) < \text{rk}_s(j)$. A matching μ is *non-wasteful* if every school s that is desired by some student in μ satisfies $|\mu_s^{-1}| = q_s$. A matching μ is *stable* if it is non-wasteful and no student's priority at any school is violated in μ.

A *mechanism* associates a matching to every school choice problem. We are mainly interested in the *student-proposing Deferred Acceptance (DA) mechanism* (Gale and Shapley, 1962), which works as follows:

Round 1: Every student applies to her most preferred school. Every school tentatively accepts the best students according to its priority, up to its capacity, and rejects the rest.

Round k: Every student rejected in the previous round applies to her next best school. Among both new applicants and previously accepted students, every school accepts the best students according to its priority, up to its capacity, and rejects the rest.

The procedure stops when there is a round without any new rejection. We use $\texttt{DA}(P)$ to denote the unique resulting matching generated by DA in school choice problem P. $\texttt{DA}_i(P)$ denotes the school to which student i is assigned. $\texttt{DA}_s^{-1}(P)$ denotes the set of students assigned to school s in $\texttt{DA}(P)$.

Similarly, we will use $\texttt{M}$ to denote an arbitrary mechanism that returns a Pareto-efficient allocation that dominates DA. We use $\mathcal{M}(P)$ to denote the class of such mechanisms for school choice problem P, which include Efficiency-Adjusted Deferred Acceptance (EADA) and Top Trading Cycles using DA's allocation as endowments (DA+TTC).

We now introduce a key concept in this paper: unimprovable students.

DEFINITION 1. A student $i \in I$ is *unimprovable* if, for every matching $\texttt{M}(P) \in \mathcal{M}(P)$, we have $\texttt{M}_i(P) = \texttt{DA}_i(P)$.

Unimprovable students do not necessarily need to be in an adverse situation. For example, every student who is assigned to their most preferred school in DA is unimprovable. We denote by $U(P)$ the set of unimprovable students in school choice problem P. It is well-known that, for any school choice problem, there is always at least one unimprovable student (Tang and Yu, 2014).

4. IDENTIFYING (UN)IMPROVABLE STUDENTS

Recall that a digraph G is given by a set of nodes V and a set of directed edges E. We define DA's envy graph as follows.

DEFINITION 2. Given a school choice problem P, we define DA's *envy digraph* $G^{\texttt{DA}(P)} = (I, E(\texttt{DA}(P)))$ where each node corresponds to a student and edge (i, j) exists if i envies j in DA.

DA's envy digraph gives us a full characterization of unimprovable students by identifying cycles in $G^{\texttt{DA}(P)}$. A *cycle* of a digraph is a sequence of nodes $i_0, i_1, \ldots, i_j, i_0$ such that there is an edge between any consecutive nodes and no edge is repeated. A *trading cycle* is a cycle in which every node appears only once, except for i_0.

PROPOSITION 1. *A student is unimprovable if and only if he does not belong to any cycle in* $G^{\mathtt{DA}(P)}$.

PROOF. The if direction, stating that a student is unimprovable if he does not belong to any cycle, is well-known in the literature (Kesten, 2010, Tang and Yu, 2014, Erdil, 2014).

For the only if part: if the cycle in which he participates is a trading cycle, then simply execute such exchange so that each student node is now assigned to the school assigned to the student node he points to. By construction of the graph, every student node involved in the swap is better off, while all other students remain unaffected.

If the cycle is not a trading cycle, then at least one node besides i_0 appears multiple times in the sequence. Let i_k be any such node. Then, rewrite the sequence so that every node appearing before i_k first appears remains, and every node appearing after i_k last appears remains (in other words, remove everything between i_k first and last appearance). Note that the modified cycle that obtains is still a cycle in the original graph. If any repeated node remains, repeat the previous step until there are no other repeated nodes left. Because the length of the cycle is finite, we end up with a trading cycle in which students may execute the exchange prescribed by the edge's direction to improve their placement without affecting others, concluding the proof. □

This characterization will be key for us to quantify unimprovable students in a probabilistic framework in the next section using a graph-theoretical approach. We can use a similar approach as before to find students who are always unimprovable, as follows.

PROPOSITION 2. *For any school choice problem* P*, if there is a student* i *such that*

i) $\mathrm{rk}_i[\mathtt{DA}_i(P)] = n$ *and there is at least one school that is under-demanded, or*

ii) $\mathtt{DA}_i(P) = s_\emptyset$,

then i *is unimprovable.*

PROOF. For part i): if a student i is assigned to his least preferred school and some school is under-demanded, then $\mathtt{DA}_i(P)$ must be the only under-demanded school because every other institution rejected student i, ad since any students assigned to an under-demanded school is unimprovable (Tang and Yu, 2014), the conclusion follows. Part ii) is proven by Alva and Manjunath (2019), but we provide a short proof nonetheless. If a student is not assigned, then the fictitious school $s_\emptyset$ is under-demanded, as it did not

reject any student due to its infinite capacity. Consequently, per the previous argument, such student is unimprovable. □

The characterization of unimprovable students through cycles of the envy digraph provides us with a clear identification of specific categories of students who cannot benefit from any mechanism that Pareto dominates DA. However, it remains unclear how many students fall into these categories in a typical school choice problem. In the next section, we turn to quantifying the proportion of unimprovable students using a probabilistic framework, which will allow us to understand the scope and distribution of potential welfare improvements through mechanisms that Pareto dominate DA.

5. Quantifying (Un)Improvable Students

Now we turn to the question of identifying how many students are expected to be unimprovable on an average school choice problem. We consider a *random school choice problem* where strict preferences and priorities are drawn independently and uniformly at random, with an equal number of students and schools (and each school's quota is one). We will use this framework assuming (as the majority of the aforementioned literature) that the number of students and schools is equal and given by n. The benefit of imposing this strong assumption is that it allows us to obtain tractable results. We denote by P_n a random instance of a school choice problem of size n and by $G^{\mathtt{DA}(P_n)}$ the corresponding random envy digraph generated by DA in such problem.[1]

First we will quantify the fraction of students who nobody envies or envy nobody, as we have argued these are unimprovable. Then we will proceed to asymptotically compute the fraction of students who are improvable.

5.1 *Students who Nobody Envies (NE).*

We use $\mathbb{E}[|NE_n|]$ to denote the expected number of students assigned to under-demanded schools, or, equivalently, the students who nobody envies. Proposition 3 shows that the expected number of under-demanded schools equals the n-th Harmonic number $H_n := \sum_{k=1}^{n} \frac{1}{k}$.

Proposition 3. *In a random school choice problem of size* n, $\mathbb{E}[|NE_n|] = H_n$.

[1] We discuss many-to-one matching and correlated preferences in the Appendix.

PROOF. DA can alternatively be implemented by a sequential algorithm in which a single student applies to a school in each step, proposed by McVitie and Wilson (1971). Whenever a student is rejected, he goes back to the queue of students who need to apply to a school. McVitie and Wilson show that, for any queuing method of unassigned and rejected students, this sequential algorithm generates the same matching as DA. The McVitie-Wilson algorithm has a close resemblance to the *coupon collector problem*, in which n coupons are drawn with replacement over a number of rounds until all coupons have appeared at least once. The analogy between both problems is that coupon drawing is equivalent to a random applications to schools, and both algorithms end once all coupons have been collected, i.e. once all schools have received at least one application.

The number of *singleton* coupons that appear exactly once in the coupon collector problem is equivalent to the number of under-demanded schools, as this is the number of schools who received exactly one application during the execution of the McVitie-Wilson algorithm. The last school to receive an application (resp. the last coupon to appear) is always under-demanded (resp. a singleton), but there might be some more. Myers and Wilf (2006) derive the joint probability distribution that there are s singleton coupons and the problem stops after r draws, and use it to compute the expected number of singleton coupons, which equals H_n. It follows that the expected number of under-demanded schools removed in the first round exactly equals H_n. □

In Proposition 5 in the Appendix, we extend this result by computing the complete distribution of envy, *i.e.* of in-degrees of $G^{\mathtt{DA}(P_n)}$.

5.2 *Students who Envy Nobody (EN).*

We use $\mathbb{E}[|EN_n|]$ to denote the expected number of students who are assigned to their most preferred schools in a school choice problem of size n. We first derive the probability distribution of the rank of students in DA as follows.

From the coupon collector problem, we know that when the number of students becomes arbitrary large there are $n\,H_n$ applications made in the execution of DA. This means that, on average, each student makes H_n applications, and on average, each school receives H_n applications, and this variable is tightly concentrated around its mean (Motwani and Raghavan, 1995, Theorem 3.8). For an average student, the probability that his application is accepted by some school to which he applies equals $\frac{1}{H_n}$, whereas the probability that his application is rejected at an arbitrary school equals

$1 - \frac{1}{H_n}$ (note that the probability of a rejection at a school is independent of whether a student has been rejected by another school, because of our assumption of independent schools' priorities). Consequently, the probability that he ends up in his k-th ranked school is the probability of $k-1$ failures and one success:

$$\frac{1}{H_n}\left(1 - \frac{1}{H_n}\right)^{k-1} \tag{1}$$

Given that $\frac{1}{H_n}$ is the probability of an application being successful, it follows that the distribution of students' ranks in DA is Geometric with parameter $\frac{1}{H_n}$. Thus, we established the following asymptotic result.

PROPOSITION 4. *For sufficiently large n and $k \leq n$,* $\Pr(\text{rk}_i = k) \approx \frac{1}{H_n}\left(1 - \frac{1}{H_n}\right)^{k-1}$.

Ashlagi et al. (2021) establish a similar result to Proposition 4, showing that ranks in a more general tiered matching model follow a geometric distribution. Because the students' rank distribution is Geometric, we immediately obtain the expected number of students who are assigned to their most preferred school.

COROLLARY 1. *For sufficiently large n,* $\mathbb{E}[|EN_n|] \approx \frac{n}{H_n}$.

From Propositions 3 and 4, we can establish a lower bound for the number of unimprovable students. For example, in a market with $n = 100$, we expect at least 5 students whom nobody envies and 19 students who envy nobody (with possible overlap between these categories). Importantly, these results demonstrate that the fraction of students who are unenvied or who envy nobody becomes negligible as n grows large. This observation naturally leads us to a broader question: what is the asymptotic behavior of the total fraction of unimprovable students? To answer this, we introduce a novel approach using strongly connected components in graph theory, which enable us to characterize the global structure of DA's envy digraph as market size grows.

5.3 *The Giant Strongly Connected Component*

A digraph is *strongly connected* if there is a path in each direction between each pair of nodes. That is, a path exists from the first node in the pair to the second, and another path exists from the second node to the first. A *strongly connected component (SCC)* of a directed graph $G = (I, E)$ is a subgraph $G' = (I' \subseteq I, E' \subseteq E)$ that is strongly connected and is maximal with this property: no additional edges or vertices from G can

be included in the subgraph without breaking its property of being strongly connected. The *size* of an SCC is the number of nodes in it. An SCC is *trivial* if it only includes one node. An SCC is *giant* when the fraction of nodes it contains is bounded away from zero asymptotically.

In our envy diagraph, improvable students and non-trivial SCCs are related by Proposition 1. First, note that if a student belongs to a non-trivial SCC, it must be part of a cycle with other nodes in the SCC and therefore is improvable. Conversely, if a student is improvable, they must belong to a cycle and every node in this cycle can reach the others and be reached by them, thereby forming (part of) a non-trivial SCC, which the student we started out with belongs to.

While the geometric out-degree distribution of $G^{\mathtt{DA}(P_n)}$ differs from the Poisson distribution characteristic of the standard Erdős-Rényi random digraph model, we demonstrate below that a unique giant SCC also emerges in DA's random envy digraph with high probability.[2]

THEOREM 1. *Fix an arbitrary $\epsilon > 0$. With high probability, $G^{\mathtt{DA}(P_n)}$ has a unique giant SCC containing at least $(1-\epsilon)n$ students.*

5.4 *Proof of Theorem 1*

The proof strategy hinges on establishing that DA's envy random digraph is sufficiently dense to preclude fragmentation into two large disjoint components lacking directed edges between them. Such a partition would necessitate that no student in the first component had applied to any school matched to students in the second component during the execution of DA—an event whose probability decreases super-polynomially as n approaches infinity. We formalize this intuition through a sequence of rigorous lemmas. First, we demonstrate that with high probability, any substantial subset of students generates a significant number of applications during the allocation process. Subsequently, we establish that the probability of absence of cross-component applications between large partitions is vanishingly small. The proof concludes by applying the union bound across all possible partitions, establishing the desired result.

Bounding the Number of Applications. We use a simple fact about DA in large markets: with high probability, the total number of applications made by a sizable subset of

[2] An event occurs with high probability if its probability approaches 1 as n grows, i.e., if the probability is at least $1 - \frac{1}{n^c}$ for some constant $c > 0$.

students is at least on the order of $n \log n$. This is because any single student, with high probability, needs $\Theta(\log n)$ applications to become matched. We formalize the claim that *every* set of students of size at least ϵn collectively makes at least $\delta n \log n$ applications for some constant $\delta > 0$ depending on ϵ.

LEMMA 1. *Fix $\epsilon > 0$ and $L \leq \infty$. With high probability, all but at most ϵn of the students each make at least L applications in the DA process.*

PROOF. It suffices to show that a *single* student i makes at least L applications with probability at least $1 - \delta$, for some δ that can be arbitrarily small as $n \to \infty$. Once this has been established, we can let X_i be the indicator that student i makes fewer than L applications. Then $\mathbb{E}[\sum_i X_i] \leq \delta n$. By Markov's inequality, the probability that more than ϵn students make fewer than L applications is bounded above by δ/ϵ for large enough n. Taking δ to be sufficiently small relative to ϵ implies that with high probability, at most ϵn students make fewer than L applications.

That any student makes $\Omega(\log n)$ applications in DA in probability is a well-known result in random matching markets (Knuth, 1976, Pittel, 1989).[3] □

Partition Argument and Probability of No Cross-Applications. Suppose that $G^{\mathtt{DA}(P_n)}$ has two disjoint subsets $I_0, I_1 \subseteq I$, each of size at least ϵn, such that there is no directed edge from I_0 to I_1 in $G^{\mathtt{DA}(P_n)}$. By definition, $(i \to j)$ means student i prefers $\mathtt{DA}_j(P_n)$ to $\mathtt{DA}_i(P_n)$. So if there is no edge from I_0 to I_1, it means no student in I_0 prefers the school matched to any student in I_1 over its own match. Equivalently, no student in I_0 applied to any school $\mathtt{DA}_j(P_n)$ with $j \in I_1$ in the DA algorithm; otherwise, that application would produce an envy edge $i \to j$ if that school ultimately ended up matched to j.

Hence, an equivalent statement is: If $G^{\mathtt{DA}(P_n)}$ has an SCC of size *at most* $(1 - 2\epsilon)n$, we can split I into two blocks I_0, I_1, each at least ϵn, so that no student in I_0 ever applied to any school in $S_1 = \{\mathtt{DA}_i(P_n) : i \in I_1\}$. We first show that for a *fixed* partition (I_0, S_1), the probability of no-application across blocks is very small.

LEMMA 2. *Let $I_0 \subseteq I$ and $S_1 \subseteq S$ be subsets of students and schools with $|I_0|, |S_1| \geq \epsilon n$. Then the probability that the students in I_0 in combination make at least Ln applications yet no one applies to any school in S_1 is at most $(1 - \epsilon)^{Ln}$.*

[3] See also Ashlagi et al. (2023, Theorem 3.2) for a characterization of rank distribution in any stable outcome.

PROOF. If applications were all mutually independent, the claim would be a simple consequence of standard binomial concentration. In DA, however, the applications depend on the history of rejections and hence see slight correlation. Instead, we consider a version of deferred acceptance with redundant applications (what Knuth (1976) calls the *amnesiac algorithm*) where we generate students' preferences in a deferred manner and direct each application-to-be-made to a uniformly random school independently. That is, we allow a student to re-apply to a school that has already rejected her (and hence will do so again). To get her true preferences, we simply remove the duplicates from the sequence of applications she will ever make. It is easy to see that this version of amnesiac DA can be coupled exactly to the conventional DA to yield the same outcome.

In the amnesiac algorithm, each student makes weakly more applications. Thus, it suffices to consider the event that the students in I_0 in combination make at least Ln applications in the amnesiac algorithm yet no one applies to any school in S_1.

To understand this probability, we use the following coupling trick to handle the preferences (and applications) of students in I_0: First, generate an (infinite) sequence of i.i.d. uniform samples $\sigma = (\sigma_1, \sigma_2, \ldots)$ from the set of schools S independent of the preferences of all schools and all students outside I; Then, run DA with redundant applications, and whenever an application is to be sampled from a student $i \in I_0$, supply with the next element from σ. We can verify inductively that, at any stage, this coupling yields the same distribution (as in the amnesiac algorithm) for the next applications to make—simply uniform and independent of all previous events. Our target no-cross-application event in the lemma statement implies that (1) at least Ln elements from σ are read before DA terminates and (2) none of these elements belong to the set S_1; these two combined imply that $\sigma_j \notin S_1$ for all $j = 1, \ldots, Ln$. By construction, the sequence $\sigma_1, \ldots, \sigma_{Ln}$ are i.i.d. samples independent of all other sources of randomness, and thus $\Pr(\sigma_j \notin S_1 \ \forall j \in [Ln]) = (\Pr(\sigma_1 \notin S_1))^{Ln}$. Since $|S_1| \geq \epsilon n$, we have $\Pr(\sigma_1 \notin S_1) \leq 1 - \epsilon$, and our desired upper bound of $(1-\epsilon)^{Ln}$ follows immediately. □

Union Bound over All Subsets We now combine Lemma 2 with a union bound over all possible realizations of I_0 and S_1 to obtain the following corollary, from which the proof of Theorem 1 follows naturally.

COROLLARY 2. *Fix $\epsilon > 0$. With high probability, there does not exist a subset of students $I_0 \subseteq I$ and a subset of schools $S_1 \subseteq S$ such that (1) both subsets are of size at least ϵn, and (2) no one in I_0 ever applied to any school in S_1 in DA.*

PROOF. For any $I_0 \subseteq I$ and $S_1 \subseteq S$ satisfying the cardinality condition, let $\mathcal{E}_{I_0,S_1}$ denote the event that no one in I_0 ever applied to any school in S_1 in DA. Proving the corollary reduces to bounding the probability of the union event

$$p := \Pr\left(\bigcup_{|I_0|,|S_1|\geq \epsilon n} \mathcal{E}_{I_0,S_1}\right).$$

Let $\mathcal{A}_{I_0}$ denote the event that students in I_0 makes a combined number of at least Ln applications (including redundant ones) for some constant L to be specified later. By a union bound, we have

$$p \leq \Pr\left(\bigcup_{|I_0|,|S_1|\geq \epsilon n} (\mathcal{E}_{I_0,S_1} \cap \mathcal{A}_{I_0}) \cup \mathcal{A}_{I_0}^c\right) \leq \sum_{|I_0|,|S_1|\geq \epsilon n} \Pr(\mathcal{E}_{I_0,S_1} \cap \mathcal{A}_{I_0}) + \Pr\left(\bigcup_{|I_0|\geq \epsilon n} \mathcal{A}_{I_0}^c\right). \tag{2}$$

By Lemma 2, $\Pr(\mathcal{E}_{I_0,S_1} \cap \mathcal{A}_{I_0}) \leq (1-\epsilon)^{Ln}$. The number of ways to choose any $I_0 \subseteq I$ is at most 2^n and same for $S_1 \subseteq S$. We only need to consider partitions where $|I_0| \geq \epsilon n$ and $|S_1| \geq \epsilon n$, so the total number of relevant partitions is strictly smaller. Thus, the summation term in the right-hand side of expression (2) is at most

$$(2^n)^2 \cdot (1-\epsilon)^{Ln} = \exp\big((2\log 2 + L\log(1-\epsilon))n\big) \to 0$$

as $n \to \infty$, granted that L is sufficiently large compared to ϵ. By Lemma 1, with high probability, all but $\epsilon n/2$ students each make at least $2L/\epsilon$ applications, and as a result any ϵn students make in total at least $2L/\epsilon \cdot \epsilon n/2 = Ln$ applications, implying the event $\bigcap_{|I_0|\geq \epsilon n} \mathcal{A}_{I_0}$. Thus, $\Pr\left(\bigcup_{|I_0|\geq \epsilon n} \mathcal{A}_{I_0}^c\right) \to 0$ as $n \to \infty$. Combining the bounds on the two terms gives an upper bound on p that tends to zero as $n \to \infty$, finishing the proof. □

We now prove Theorem 1 using Corollary 2.

PROOF OF THEOREM 1. Note that DA with redundant applications produces the exact same matching and envy graph (modulo duplicate edges due to redundancy). Thus, we may use DA with redundant applications as our underlying data generation algorithm.

Given that ϵ is arbitrary, it suffices to show that the giant SCCs contains at least $(1-2\epsilon)n$ students with high probability. When this is not the case, there must exist a set of SCCs of combined size at least ϵn that either cannot reach or cannot be reached from the largest SCC. In either case, there exist disjoint subsets $I_0, I_1 \subseteq I$ both of size at least ϵn where there are no edges from I_0 to I_1, and equivalently no students in I_0 have applied

to any school in $S_1 := \{\texttt{DA}_i(P) : i \in I_1\}$. This is, however, unlikely: By Corollary 2, the probability of being able to find such sizable subsets I_0 and S_1 with no applications between them tends to zero. Hence, the size of the giant SCCs is at least $(1-2\epsilon)n$ with high probability. □

Theorem 1 yields a surprising corollary, namely:

COROLLARY 3. *The fraction of students who are unimprovable converges to zero in large markets.*

Theorem 1 implies that the set of improvable students forms a substantial majority of the entire student population, but is silent regarding how many of them can be improved by the same mechanism, nor does it specify which mechanism chooses the maximum improvement over DA. Our next Theorem tackles both questions.

5.5 *Equivalence among Mechanisms that Dominate DA*

We define a *cycle packing* H as a union of pairwise-disjoint cycles in G, with the property that no additional cycles exist in the subgraph $G \backslash H$. $V(H)$ denotes the vertex set of H, or equivalently, the set of nodes that are in some cycle in H. Note that, by Proposition 1, every efficient mechanism corresponds to a cycle packing of DA's envy digraph. Our second Theorem shows that, with high probability, every efficient mechanism improves almost all nodes.

THEOREM 2. *With high probability, every cycle packing of the envy digraph $G^{\texttt{DA}(P_n)}$ covers at least $(1-\epsilon)n$ nodes, for any arbitrary constant $\epsilon > 0$.*

PROOF. Let $H \subseteq G^{\texttt{DA}(P_n)}$ be a cycle packing. Consider the induced subgraph $G^{\texttt{DA}(P_n)} \backslash H = G^{\texttt{DA}(P_n)}[I \backslash V(H)]$, which by definition, contains no cycles.

Suppose, for contradiction, that $|V(H)| < (1-\epsilon)n$. Then, the directed acyclic subgraph $G^{\texttt{DA}(P_n)} \backslash H$ contains $N = |I \backslash V(H)| > \epsilon n$ vertices. Hence, there exists a topological ordering $i_1, \ldots, i_N$ of vertices in $I \backslash V(H)$ such that no directed edge from i_k to i_ℓ for $1 \leq k < \ell \leq N$ is present in $G^{\texttt{DA}(P_n)} \backslash H$ (and hence $G^{\texttt{DA}(P_n)}$).

Partitioning these nodes into $I_0 = \{i_1, \ldots, i_{\lfloor N/2 \rfloor}\}$ and $I_1 = \{i_{\lfloor N/2 \rfloor + 1}, \ldots, i_N\}$ yields two subsets, each with size at least $\lfloor \epsilon n/2 \rfloor$, and crucially, no edges in $G^{\texttt{DA}(P_n)}$ go from I_0 into I_1. This scenario is identical to the one in Theorem 1, which is improbable due to Corollary 2. Consequently, we must have $|V(H)| \geq (1-\epsilon)n$ with high probability. □

Theorem 2 establishes not merely the prevalence of improvable students but also demonstrates that nearly all such students can concurrently benefit from improvement via non-intersecting trading cycles. Furthermore, Theorem 2 establishes the asymptotic equivalence across all such mechanisms regarding the proportion of students they improve. This result has significant implications for policy implementation. When policymakers deliberate on which efficiency-enhancing mechanism to adopt as an alternative to DA, concerns about differential distributions of efficiency gains become largely irrelevant, as all such mechanisms ultimately benefit approximately the same number of students in large markets. This equivalence result is particularly striking given that in carefully constructed small instances, different efficient mechanisms that dominate DA may significantly differ in their corresponding number of improved students (Knipe and Ortega, 2025).

5.6 *Simulations and Extensions*

The simulations depicted in Figure 1 show the emergence of a unique giant component in $G^{\mathtt{DA}(P_n)}$ for even relatively small values of n.

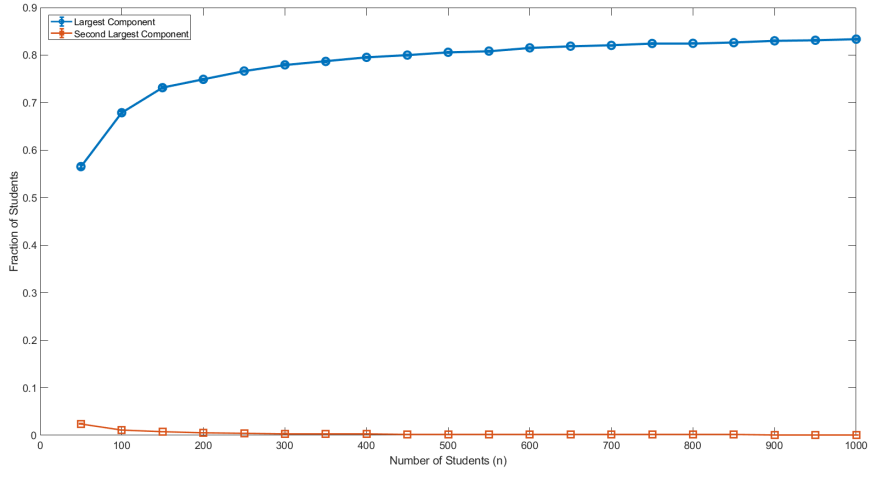


FIGURE 1. Average fraction of nodes in the largest (blue) and second-largest (red) SCCs (average over 2,000 random envy digraphs for each n).

The simulations in Figure 2 illustrate the convergence of unimprovable students to zero as market size increases. The convergence rate is approximately $O(1/\log n)$, directly

corresponding to the rate at which students receive their top choice under DA (as shown in Corollary 1). This theoretical relationship explains the gradual nature of the decline.

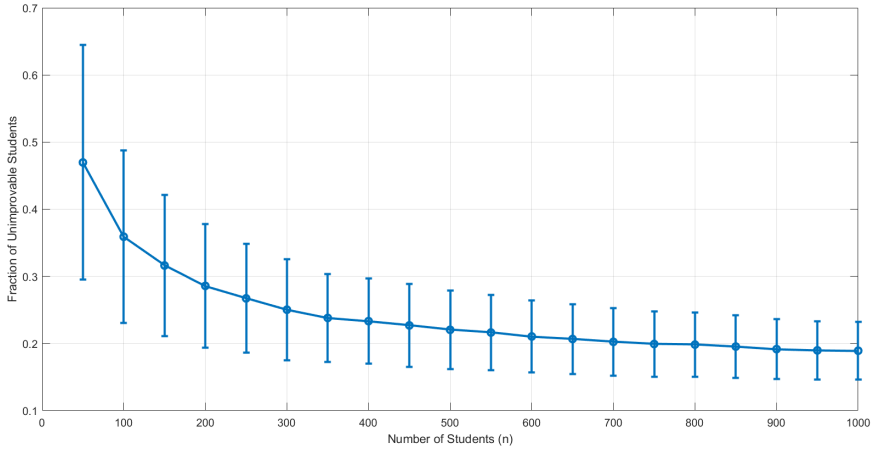


FIGURE 2. Average fraction of unimprovable students (average over 2,000 random problems for each n).

In the Appendix, we discuss two extensions of our work.

1. Many-to-one Matching: Our probabilistic analysis has assumed that the number of schools and students is the same, yet our main results (Theorems 1 and 2)) can extended to a model with n schools, each with a quota q, and qn students. In this model, each student makes fewer applications, but each rejection makes him envy q instead of only 1 student. This trade-off between application frequency and envy multiplicity results in an envy digraph that becomes even more densely connected. Consequently, the emergence of a unique giant strongly connected component remains a robust phenomenon, as well as the asymptotic size equivalence among efficient mechanisms that dominate DA. However, convergence occurs more slowly as q increases. We provide proofs and simulations in the Appendix
2. Correlated Preferences: In the Appendix, we introduce correlation on students' preferences by dividing the set of schools into tiers, as in Ashlagi et al. (2021). We find that several giant SCCs form, in fact as many as tiers, including a sizeable fraction of students. Simulations suggest that the fraction of nodes in a giant SCC converges to one asymptotically.

6. Conclusion

DA's inefficiency affects almost all students in large markets, with nearly all students becoming improvable through trading cycles in the envy digraph as market size grows. While our findings suggest that an efficiency adjustment should be implemented, our equivalence theorem says such debate should be guided by other properties and not by the number of students who benefit from a better school placement.

References

Abdulkadiroglu, Atila, Parag Pathak, and Alvin Roth (2005), "The new york city high school match." *American Economic Review*, 95(2), 364–367. [4]

Abdulkadiroğlu, Atila, Parag A Pathak, and Alvin E Roth (2009), "Strategy-proofness versus efficiency in matching with indifferences: Redesigning the nyc high school match." *American Economic Review*, 99(5), 1954–78. [2, 4]

Abdulkadiroğlu, Atila and Tayfun Sönmez (2003), "School choice: A mechanism design approach." *American Economic Review*, 93(3), 729–747. [4, 5]

Alcalde, José and Antonio Romero-Medina (2017), "Fair student placement." *Theory and Decision*, 83(2), 293–307. [4]

Alva, Samson and Vikram Manjunath (2019), "Stable-dominating rules." Technical report, Working paper, University of Ottawa. [4, 8]

Ashlagi, Itai, Mark Braverman, Amin Saberi, Clayton Thomas, and Geng Zhao (2021), "Tiered random matching markets: Rank is proportional to popularity." In *12th Innovations in Theoretical Computer Science Conference, ITCS 2021*, 46, Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing. [5, 11, 18, 24, 26]

Ashlagi, Itai, Mark Braverman, and Geng Zhao (2023), "Welfare distribution in two-sided random matching markets." *arXiv preprint arXiv:2302.08599*. [5, 13]

Ashlagi, Itai, Yash Kanoria, and Jacob Leshno (2017), "Unbalanced random matching markets: The stark effect of competition." *Journal of Political Economy*, 125(1), 69–98. [5]

Bando, Keisuke (2014), "On the existence of a strictly strong nash equilibrium under the student-optimal deferred acceptance algorithm." *Games and Economic Behavior*, 87, 269–287. [4]

Brown, Mark, Erol A Peköz, and Sheldon M Ross (2008), "Coupon collecting." *Probability in the Engineering and Informational Sciences*, 22(2), 221–229. [24]

Cerrone, Claudia, Yoan Hermstrüwer, and Onur Kesten (2024), "School choice with consent: An experiment." *The Economic Journal*. [5]

Che, Yeon-Koo and Fuhito Kojima (2010), "Asymptotic equivalence of probabilistic serial and random priority mechanisms." *Econometrica*, 78(5), 1625–1672. [5]

Che, Yeon-Koo and Olivier Tercieux (2018), "Payoff equivalence of efficient mechanisms in large matching markets." *Theoretical Economics*, 13(1), 239–271. [5]

Che, Yeon-Koo and Olivier Tercieux (2019), "Efficiency and stability in large matching markets." *Journal of Political Economy*, 127(5), 2301–2342. [4, 5]

Chen, Yiqiu and Markus Möller (2023), "Regret-free truth-telling in school choice with consent." *Theoretical Economics*. [4]

Diebold, Franz and Martin Bichler (2017), "Matching with indifferences: A comparison of algorithms in the context of course allocation." *European Journal of Operational Research*, 260(1), 268–282. [5]

Doğan, Battal and Lars Ehlers (2021), "Minimally unstable pareto improvements over deferred acceptance." *Theoretical Economics*, 16(4), 1249–1279. [4]

Dur, Umut, A Arda Gitmez, and Özgür Yılmaz (2019), "School choice under partial fairness." *Theoretical Economics*, 14(4), 1309–1346. [4]

Ehlers, Lars and Thayer Morrill (2020), "(il) legal assignments in school choice." *The Review of Economic Studies*, 87(4), 1837–1875. [4]

Erdil, Aytek (2014), "Strategy-proof stochastic assignment." *Journal of Economic Theory*, 151, 146–162. [8]

Erdős, Pál and Alfréd Rényi (1961), "On a classical problem of probability theory." *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 6(1-2), 215–220. [26]

Gale, David and Lloyd S Shapley (1962), "College admissions and the stability of marriage." *The American Mathematical Monthly*, 69(1), 9–15. [6]

Immorlica, Nicole and Mohammad Mahdian (2005), "Marriage, honesty, and stability." In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, 53–62. [5]

Kesten, Onur (2010), "School choice with consent." *The Quarterly Journal of Economics*, 125(3), 1297–1348. [2, 4, 8]

Knipe, Taylor and Josué Ortega (2025), "Improvable students in school choice." *arXiv preprint arXiv:2504.12871*. [3, 17]

Knuth, Donald (1976), *Mariages stables et leurs relations avec d'autres problèmes combinatoires: introduction à l'analyse mathématique des algorithmes*. Presses de l'Université de Montréal. [5, 13, 14, 26]

Kojima, Fuhito and Parag A. Pathak (2009), "Incentives and stability in large two-sided matching markets." *American Economic Review*, 99(3), 608–27. [5]

Le Cam, Lucien (1960), "An approximation theorem for the poisson binomial distribution." *Pacific Journal of Mathematics*, 10(4), 1181–1197. [25]

Lee, SangMok (2016), "Incentive compatibility of large centralized matching markets." *The Review of Economic Studies*, 84(1), 444–463. [5]

Liu, Qingmin and Marek Pycia (2016), "Ordinal efficiency, fairness, and incentives in large markets." *SSRN preprint*. [5]

McVitie, D. and L. Wilson (1971), "The stable marriage problem." *Communications of the ACM*, 14(7), 486–490. [10]

Motwani, Rajeev and Prabhakar Raghavan (1995), *Randomized algorithms*. Cambridge University Press, New York, NY, USA. [10, 23]

Myers, Amy and Herbert Wilf (2006), "Some new aspects of the coupon collector's problem." *SIAM Review*, 48(3), 549–565. [10]

Newman, Donald J and Lawrence Shepp (1960), "The double dixie cup problem." *The American Mathematical Monthly*, 67(1), 58–61. [26]

Nikzad, Afshin (2022), "Rank-optimal assignments in uniform markets." *Theoretical Economics*, 17(1), 25–55. [5]

Ortega, Josué and Thilo Klein (2023), "The cost of strategy-proofness in school choice." *Games and Economic Behavior*, 141, 515–528. [4, 5]

Pittel, Boris (1989), "The average number of stable matchings." *SIAM Journal on Discrete Mathematics*, 2(4), 530–549. [5, 13]

Pittel, Boris (1992), "On likely solutions of a stable marriage problem." *The Annals of Applied Probability*, 358–401. [5]

Pycia, Marek (2019), "Evaluating with statistics: Which outcome measures differentiate among matching mechanisms?" *Unpublished paper, University of Zurich*. [5]

Reny, Philip J (2022), "Efficient matching in the school choice problem." *American Economic Review*, 112(6), 2025–43. [4]

Ronen, Amit, Jonah Evan Hess, Yael Belfer, Simon Mauras, and Alon Eden (2025), "Stable marriage: Loyalty vs. competition." *arXiv preprint arXiv:2501.18442*. [5]

Tang, Qianfeng and Jingsheng Yu (2014), "A new perspective on kesten's school choice with consent idea." *Journal of Economic Theory*, 154, 543–561. [4, 7, 8]

Tang, Qianfeng and Yongchao Zhang (2021), "Weak stability and pareto efficiency in school choice." *Economic Theory*, 71, 533–552. [4]

Troyan, Peter, David Delacrétaz, and Andrew Kloosterman (2020), "Essentially stable matchings." *Games and Economic Behavior*, 120, 370–390. [4]

Troyan, Peter and Thayer Morrill (2020), "Obvious manipulations." *Journal of Economic Theory*, 185, 104970. [4]

Wilson, L (1972), "An analysis of the stable marriage assignment algorithm." *BIT Numerical Mathematics*, 12(4), 569–575. [5]

APPENDIX A: ADDITIONAL RESULTS

A.1 *In-degree Distribution*

In the main text, we have derived the distribution of students' ranks in DA, obtaining the out-degree distribution of $G^{\mathtt{DA}(P_n)}$ as a corollary. Now, we derive the in-degree distribution as follows.

PROPOSITION 5 (The Distribution of Envy). *For any fixed integer k, there exists sufficiently large n, such that* $\Pr(\overline{deg}(i) = k-1) \approx \frac{(H_n)^k}{k!} e^{-H_n}$.

PROOF. The coupon collector analogy will be again useful for us. If a student i is assigned to school s, and school s is drawn x times in the coupon collector problem, then there are $x-1$ students who envy student i, as they were all rejected by school x. So, to obtain how frequently a student is envied, we will study the number of times that a coupon appears over the time that it takes for all coupons to be collected T, which is well-known to be tightly concentrated around nH_n (Motwani and Raghavan, 1995). Therefore, in what follows we fix $T = nH_n$ (this assumption simplifies the proof, but we provide a complete proof in Proposition 6 below that does not impose it).

Over the course of the nH_n applications that it takes for the algorithm to finish, each day the coupon/school s is drawn with probability $1/n$. This means that the number of times that coupon s is drawn, denoted by X_i, is a $X_i \sim \text{Binomial}(nH_n, 1/n)$. When the number of trials is large and the success probability is small (both of which occur when n grows), the binomial is well-approximated by a Poisson distribution (Motwani and Raghavan, 1995, section 3.6.2). Hence, we may approximate $X_i \approx \text{Poisson}(\lambda)$, with the parameter $\lambda = H_n$. In consequence, the chance that coupon/school s appears k times is approximately:

$$\Pr(X_i = k) = e^{-\lambda} \frac{\lambda^k}{k!} \tag{3}$$

Substituting $\lambda = H_n$, we obtain:

$$\Pr(X_i = k) = e^{-H_n} \frac{(H_n)^k}{k!} \tag{4}$$

□

The previous proof directly assume that the coupon collector process ends in nH_n iterations, making the argument easy to follow. Now we present a formal proof that avoids such assumption, but that is more complex.

PROPOSITION 6 (Poisson Approximation for In-Degrees). *In the coupon collector problem with n distinct coupons, let X_i denote the number of times coupon i appears during the collection process. Then, for each fixed non-negative integer k,*

$$\Pr(X_i = k) = e^{-H_n} \frac{(H_n)^k}{k!} + o(1) \tag{5}$$

as $n \to \infty$. That is, the distribution of X_i is asymptomatically well-approximated by a distribution of a Poisson random variable with parameter H_n.

PROOF. We use the *Poissonization* technique to simplify the analysis by replacing the fixed (but random) total number of trials with a Poisson-distributed number. This technique has been used to analyze the coupon collector problem (Brown et al., 2008, Ashlagi et al., 2021). We proceed in five steps.

Step 1: Poissonized Model. Let $N \sim \text{Poisson}(\lambda)$ be the number of trials, where $\lambda = nH_n$. This choice aligns the expected number of trials with the original problem, since $\mathbb{E}[T] = nH_n$, where T is the time to collect all n coupons.

In the Poissonized model:

- Each trial independently yields coupon i with probability $1/n$.
- The number of times coupon i appears, X_i, follows a Poisson(λ/n) distribution.

Thus, for any fixed k,

$$\Pr(X_i = k \mid N \sim \text{Poisson}(\lambda)) = e^{-\lambda/n} \frac{(\lambda/n)^k}{k!}. \tag{6}$$

Substituting $\lambda = nH_n$, we get:

$$\Pr(X_i = k \mid N \sim \text{Poisson}(nH_n)) = e^{-H_n} \frac{(H_n)^k}{k!}. \tag{7}$$

Here, H_n naturally emerges as the expected intensity of coupon i's appearances over nH_n trials.

Step 2: Relating Poissonized Model to Original Problem. In the original problem, T is the number of trials until all n coupons are collected. We know $\mathbb{E}[T] = nH_n$ and $\text{Var}(T) \leq n^2\pi^2/6 = O(n^2)$. To bridge the models, we use concentration of T around its mean. By Chebyshev's inequality:

$$\Pr\big(|T - nH_n| > \varepsilon nH_n\big) \leq \frac{\text{Var}(T)}{(\varepsilon nH_n)^2} \leq \frac{\pi^2 n^2/6}{\varepsilon^2 n^2 (H_n)^2} = \frac{\pi^2}{6\varepsilon^2 (H_n)^2}. \tag{8}$$

Since $H_n \approx \log n + \gamma$, we have $\Pr\big(|T - nH_n| > \varepsilon nH_n\big) = O(1/(\log n)^2) \to 0$ as $n \to \infty$. Thus, T is sharply concentrated around nH_n, justifying the Poissonized approximation.

Step 3: De-Poissonization. To connect back to the original problem, consider the true distribution of X_i:

$$\Pr(X_i = k) = \sum_{t=0}^{\infty} \Pr(X_i = k \mid T = t) \Pr(T = t). \tag{9}$$

Given $T = t$, each of the t trials independently yield coupon i with probability $1/n$, so $X_i \sim \text{Binomial}(t, 1/n)$. For large t, this approximates a Poisson distribution: if $t = nH_n(1 + \delta_n)$ with $|\delta_n| \to 0$, then Binomial$(t, 1/n) \approx$ Poisson$(t/n) \approx$ Poisson$(H_n(1 + \delta_n))$.

Define the "typical" set $A_n = \{t : |t - nH_n| \le \varepsilon n H_n\}$. Then:

$$\Pr(X_i = k) = \sum_{t \in A_n} \Pr(X_i = k \mid T = t) \Pr(T = t) + \sum_{t \notin A_n} \Pr(X_i = k \mid T = t) \Pr(T = t). \quad (10)$$

- For $t \in A_n$, $t/n = H_n(1 + \delta_n)$ with $|\delta_n| \le \varepsilon$, and $\Pr(X_i = k \mid T = t) = \binom{t}{k}(1/n)^k(1 - 1/n)^{t-k}$. Using the Poisson approximation for binomials (Le Cam 1960, valid since t/n is finite and n is large), this is:

$$\Pr(X_i = k \mid T = t) \approx e^{-t/n} \frac{(t/n)^k}{k!} = e^{-H_n(1+\delta_n)} \frac{(H_n(1 + \delta_n))^k}{k!}. \quad (11)$$

  As $\varepsilon \to 0$, this approaches $e^{-H_n}(H_n)^k/k!$.
- For $t \notin A_n$, $\Pr(T = t) \le \Pr(|T - nH_n| > \varepsilon n H_n) = O(1/(\log n)^2)$, so this term contributes $o(1)$.

Since $\Pr(T \in A_n) \to 1$ and the conditional probability stabilizes, the error is $o(1)$.

Step 4: Poisson Approximation. In the Poissonized model, $X_i \sim \text{Poisson}(H_n)$, giving:

$$\Pr(X_i = k \mid N \sim \text{Poisson}(nH_n)) = e^{-H_n} \frac{(H_n)^k}{k!}. \quad (12)$$

By the de-Poissonization argument, this holds in the original problem with error $o(1)$:

$$\Pr(X_i = k) = e^{-H_n} \frac{(H_n)^k}{k!} + o(1). \quad (13)$$

Step 5: Conclusion. We have shown that, for each fixed k, $\Pr(X_i = k)$ matches the Poisson(H_n) probability mass function up to $o(1)$, so X_i converges in distribution to Poisson(H_n) as $n \to \infty$. (Note that $H_n \to \infty$, but convergence is assessed pointwise for each k.) This completes the proof. □

A.2 *Many-to-One Matching*

In the main text, we have shown that a giant unique SCC appears in $G^{\mathtt{DA}(P_n)}$ if there are as many schools as students. We relax such assumption here, allowing for a school to be able to admit many students, as follows.

THEOREM 3 (Giant SCC in Many-to-One Matching). *Consider a random school choice problem P_n with n schools, where each school has the same fixed quota q, and nq students. Fix an arbitrary $\epsilon > 0$. With high probability, the envy digraph $G^{\mathtt{DA}(P_n)}$ has a unique giant strongly connected component containing at least $(1 - \epsilon)nq$ students.*

PROOF. We follow a strategy similar to Theorem 1, adapting it to the many-to-one setting. First, we establish bounds on the number of applications in the DA algorithm, then analyze partitions of the resulting envy digraph.

Step 1: Bounds on applications. We first establish the following bound on the number of applications a typical student makes, similar to Lemma 1:

LEMMA 3. *Fix $\epsilon > 0$. With high probability, all but at most ϵnq students each make at least $\sqrt{\log n}$ applications.*

PROOF. Again, it suffices to show that for a fixed student i, the number of applications she makes is at least $\sqrt{\log n}$ with high probability; the lemma follows as a consequence of Markov's inequality applied to the number of students making fewer applications than this, same as we saw in the proof of Lemma 1.

The standard approach for one-to-one matchings using the coupon collector problem as an approximation can be adapted to this case with slight modifications. In particular, one can establish that the total number of applications throughout DA is of order $\Theta(n \log n)$. Since the order of applications in DA is irrelevant, one can hold off i and run DA on the rest of the students $I \backslash \{i\}$, and when this first stage terminates (i.e., when all the students except for i are stably matched) there should be already $\Theta(n \log n)$ applications. At this point, all but one school has their capacity filled; Without loss of generality, let it be school s_n. Using terminologies from Ashlagi et al. (2021), we say that this intermediate state of DA is *smooth* if (1) at most $O(n \log n)$ applications are made until this point and (2) all but at most n^a schools have received applications from $\Omega(\log n)$ distinct students with constant $a < 1$. With nearly identical analysis as in Ashlagi et al. (2021), we can show that a smooth state is achieved with high probability; the details are repetitive and hence omitted.[4] We now let i start making applications until the full DA terminates (i.e., some one applies to this last unfilled school s_n). Note that, by symmetry, an application to a school who has received $m \geq q$ applications will be accepted with probability $q/(m+1)$. Denote the number of applications to school s_k at the beginning of this second stage by M_k for $k = 1, \ldots, n-1$. Under a smooth state, we may assume without loss of generality that $M_k \geq \Theta(\log n)$ for $k = 1, \ldots, \lfloor n - n^a \rfloor$; as the second stage of DA unfolds, the number of applications to each school can only increase. The probability that an application from i is accepted is at most

$$\frac{1}{n}\Big(1 + \sum_{k=1}^{n-1} \frac{q}{M_k + 1}\Big) \leq \frac{n^a}{n} + (n - n^a)\frac{q}{\Theta(\log n)} \leq \Theta(q/\log n).$$

Thus, conditional on the likely smooth state before i starts to apply, the number of applications that i makes until receiving a (tentative) acceptance is stochastically lower bounded by a geometric distribution with success probability $\Theta(q/\log n)$, and therefore,

[4] The first bound on the number of applications can be derived from classic results on the generalized coupon collector problem; see, e.g., Newman and Shepp (1960), Erdős and Rényi (1961). The second bound on the number of applications received by the schools is slightly more involved. To derive it, we run the amnesiac DA (Knuth, 1976), and using concentration arguments on the number of draws in (generalized) coupon collector problem, one can show that the number of applications, counting duplicates, is of order $\Omega(n \log n)$ up to this stage; further, duplicates are rare among them, so the total number of applications in the classic DA is again $\Omega(n \log n)$ with high probability prior to the second stage. The number of applications each school receives can then be bounded using a Poissonization argument or with direct Chernoff-type bounds. See, e.g., Ashlagi et al. (2021, Proposition 4.3) for a formal statement and analysis in the one-to-one case.

the conditional probability (under a smooth state) that i makes fewer than $\sqrt{\log n}$ applications is at most $1-(1-\Theta(q/\log n))^{\sqrt{\log n}} \to 0$. Since a smooth state is achieved with high probability, the marginal probability of i making less than $\sqrt{\log n}$ applications vanishes as $n \to \infty$, finishing our proof. □

Step 2: Partition Argument. Suppose $G^{\mathtt{DA}(P_n)}$ has two disjoint subsets $I_0, I_1 \subseteq I$, each of size $\geq \epsilon nq$, with no edges from I_0 to I_1. Let $S_1 = \{s \in S : \exists j \in I_1 \text{ with } \mathtt{DA}_j(P_n) = s\}$, so $|S_1| \geq \epsilon n$. We now show that for any fixed subsets I_0 and S_1, it is extremely unlikely that no one in I_0 ever applies to schools in S_1.

LEMMA 4. *For any fixed $I_0 \subseteq I$ and $S_1 \subseteq S$ with $|I_0| \geq \epsilon nq$ and $|S_1| \geq \epsilon n$, the probability that students in I_0 make at least $n\sqrt{\log n}$ applications yet none apply to S_1 is at most $(1-\epsilon)^{n\sqrt{\log n}}$.*

PROOF. We use the similar idea as in the proof of Lemma 1. In the amnesiac DA model, each application independently targets a school in S_1 with probability $\geq \epsilon$. Thus, the probability of the event in the lemma is at most $(1-\epsilon)^{n\sqrt{\log n}}$ as claimed. □

Step 3: Union Bound over Partitions. The event that the giant SCC contains strictly less than $(1-2\epsilon)nq$ students implies the existence of linear-sized $I_0, I_1 \subseteq I$ with no edges going from I_0 to I_1, and thus no one in I_0 ever applies to the set of schools $S_1 := \{s \in S : \exists j \in I_1, \mathtt{DA}_j(P) = s\}$. Denote such an event by $\mathcal{E}_{I_0,S_1}$. By a similar union bound argument as in the proof of Theorem 1, the probability p the giant SCC contains $< (1-2\epsilon)nq$ students is bounded by:

$$p \leq \sum_{|I_0|\geq \epsilon nq, |S_1|\geq \epsilon n} \Pr(\mathcal{E}_{I_0,S_1} \cap \mathcal{A}_{I_0}) + \Pr\left(\bigcup_{|I_0|\geq \epsilon nq} \mathcal{A}^c_{I_0}\right), \tag{14}$$

where $\mathcal{A}_{I_0}$ denotes the event that the subset of students I_0 make at least $n\sqrt{\log n}$ applications in total.

Note that $\mathcal{E}_{I_0,S_1} \cap \mathcal{A}_{I_0}$ is precisely the kind of events characterized in Lemma 4, and hence the first summation term (which has fewer than $(2^{nq})^2$ terms) is at most

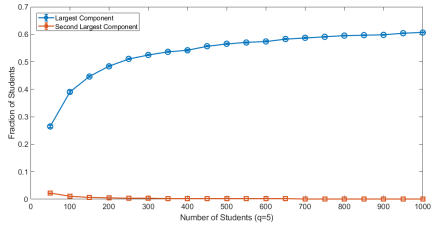
$$2^{2nq} \cdot (1-\epsilon)^{n\sqrt{\log n}} = \exp\left(2nq\ln 2 + n\sqrt{\log n}\log(1-\epsilon)\right) \to 0 \tag{15}$$

as $n \to \infty$. Lemma 3 further ensures $\Pr(\bigcup_{|I_0|\geq \epsilon nq} \mathcal{A}^c_{I_0}) \to 0$ as $n \to \infty$. Thus, the giant SCC contains $\geq (1-\epsilon)nq$ students with high probability. □

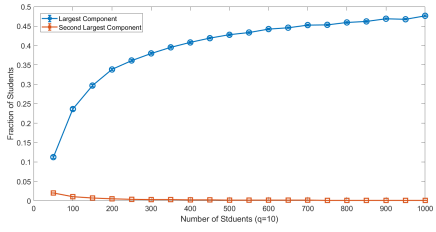
Despite the different structure of the many-to-one market with uniform quotas, the fundamental conclusion remains: in large random markets, a unique giant SCC emerges, and thus most students are improvable. However, we emphasize that the rate of convergence in the many-to-one setting differs from the one-to-one case, and depends on the size of the constant q: As q increases, each school forms larger "envy-free clusters" (students assigned to the same school), potentially slowing convergence.

In Figures A.1a and A.1b, we observe exactly that: the unique giant SCC emerges but the fraction of nodes in it for fixed n becomes smaller for larger constant q.

Note that although the convergence appears slow, scaling like $1/\log n$ and consistent with our analysis,the main driver behind this slow rate is the highly dispersed distribution of applications submitted by students. In particular, an $O(1/\log n)$ fraction of students submit only one application and thus envy no one. Empirical simulations suggest that, aside from these trivially unimprovable students, the giant strongly connected component (SCC) contains nearly all of the remaining students.



(a) $n = [10 : 10 : 200], q = 5$.



(b) $n = [5 : 5 : 100], q = 10$.

FIGURE A.1. Average fraction of nodes in the largest (blue) and second-largest (red) SCCs (average over 1,000 random problems for each n), preferences and priorities uniform i.i.d.

Equivalence in Many-to-One Matching. The cycle packing result in Theorem 2 also extends naturally to the many-to-one setting with fixed quota q for each school. We present this generalization as follows:

THEOREM 4 (Cycle Packing in Many-to-One Matching). *Consider a random school choice problem P_n with n schools, where each school has the same fixed quota q, and nq students. Fix an arbitrary $\epsilon > 0$. With high probability, every cycle packing of the envy digraph $G^{\mathtt{DA}(P_n)}$ covers at least $(1-\epsilon)nq$ students.*

PROOF. The proof follows the approach of Theorem 2, adapted to the many-to-one setting. Let $H \subseteq G^{\mathtt{DA}(P_n)}$ be a cycle packing, and consider the induced subgraph $G^{\mathtt{DA}(P_n)} \backslash H = G^{\mathtt{DA}(P_n)}[I \backslash V(H)]$, which contains no cycles by definition.

Suppose, for contradiction, that $|V(H)| < (1-\epsilon)nq$. Then, the directed acyclic subgraph $G^{\mathtt{DA}(P_n)} \backslash H$ contains $N = |I \backslash V(H)| > \epsilon nq$ vertices. We can find a topological ordering $i_1, \ldots, i_N$ of vertices in $I \backslash V(H)$ such that no directed edge from i_k to i_ℓ for $1 \le k < \ell \le N$ is present in $G^{\mathtt{DA}(P_n)} \backslash H$.

Partitioning these nodes into $I_0 = \{i_1, \ldots, i_{\lfloor N/2 \rfloor}\}$ and $I_1 = \{i_{\lfloor N/2 \rfloor + 1}, \ldots, i_N\}$ yields two subsets, each with size at least $\lfloor \epsilon nq/2 \rfloor$, with no edges going from I_0 to I_1 in $G^{\mathtt{DA}(P_n)}$.

This scenario implies that no student in I_0 has applied to any school matched to students in I_1 during the execution of DA. By Lemma 4, such a partition is extremely unlikely for sufficiently large n, with probability vanishing as n grows.

Thus, we must have $|V(H)| \ge (1-\epsilon)nq$ with high probability. □

These findings collectively demonstrate that our main conclusions—both about the pervasiveness of improvable students and the asymptotic equivalence of mechanisms

that Pareto dominate DA—are robust to varying market structures, including the practically important case of schools with multiple seats.

A.3 *Correlation in Preferences*

In the main text, we assume that students' preferences are independently drawn. Here we relax such assumption by dividing schools into two tiers of equal size, and assume that each school in Tier 1 is more preferred than any school in Tier 2. This tiered model aims to capture the empirical observation that some schools are known to be of high quality and preferred by most students. Within tiers, preferences are drawn uniformly at random.

Interestingly, in this scenario we observe the emergence of two (rather than one) giant SCCs. Simulations suggest that the fraction of nodes in one of the two giant SCCs converge to 1 as the market grows, as follows. The emergence of two giant SCCs is ex-

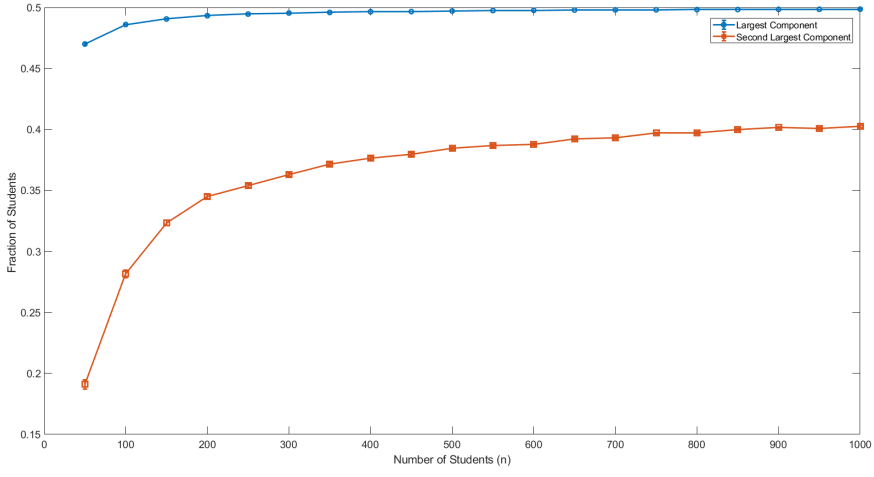


FIGURE A.2. Average fraction of nodes in the largest (blue) and second-largest (red) SCCs in a two-tiered model (average over 1,000 random envy digraphs for each n).

pected. Every student matched with a school in Tier 2 is not envied by any student in Tier 1, so therefore any SCCs can contain either only Tier 1 students (i.e. matched to schools from Tier 1) or Tier 2 students, which implies no SCC includes more than half of the students. But, since each subgraph is dense enough, we find that a giant SCC obtains for each tier of the market, and thus almost all students in each tier are improvable via within-tier exchanges. Based on these observations, we conjecture that in a model that allows for a constant c number of tiers, c giant SCCs grow and they contain asymptotically almost all nodes as the market grows.