Approximately Stable, School Optimal, and Student-Truthful Many-to-One Matchings (via Differential Privacy)*

Sampath Kannan[1], Jamie Morgenstern[2], Aaron Roth[1], and Zhiwei Steven Wu[1]

[1]*Department of Computer and Information Science, University of Pennsylvania*
[2]*Computer Science Department, Carnegie Mellon University*

July 9, 2014

Abstract

We present a mechanism for computing asymptotically stable school optimal matchings, while guaranteeing that it is an asymptotic dominant strategy for every student to report their true preferences to the mechanism. Our main tool in this endeavor is *differential privacy*: we give an algorithm that coordinates a stable matching using differentially private signals, which lead to our truthfulness guarantee. This is the first setting in which it is known how to achieve nontrivial truthfulness guarantees for students when computing school optimal matchings, assuming *worst-case* preferences (for schools and students) in large markets.

*Kannan was partially supported by NSF grant NRI-1317788. email: kannan@cis.upenn.edu. Morgenstern was partially supported by NSF grants CCF-1116892 and CCF-1101215, as well as a Simons Award for Graduate Students in Theoretical Computer Science. Contact information: J. Morgenstern, Computer Science Department, Carnegie Mellon University, `jamiemmt@cs.cmu.edu`. Roth was partially supported by an NSF CAREER award, NSF Grants CCF-1101389 and CNS-1065060, and a Google Focused Research Award. Email: `aaroth@cis.upenn.edu`. Wu was supported in part by NSF Grants CCF-1101389. Email: `wuzhiwei@cis.upenn.edu`

1 Introduction

In this paper we consider the important problem of computing many-to-one stable matchings – a matching solution concept used for diverse applications, including matching students to schools [Abdulkadiroglu et al., 2009], and medical residents to hospitals [Roth, 1984, Roth and Peranson, 1999]. There are two sides of such a market, which we will without loss of generality refer to as the *students* and the *schools*. The goal is to find a feasible assignment μ of students to schools – each student a can be matched to at most 1 school, but each school u can be potentially matched to up to C_u students, where C_u is the *capacity* of school u. We would like to find a matching that is *stable*. Informally, when each student a has a preference ordering $\succ_a$ over schools, and each school u has a preference ordering $\succ_u$ over students, then an assignment μ forms a stable matching if it is feasible, and there is no student-school pair (a, u) such that they are unmatched ($\mu(a) \neq u$), but such that they would mutually prefer to deviate from the proposed matching μ and match with each other.

The set of stable many-to-one-matchings have a remarkable structural property: there exists a *school optimal* and a *student optimal* stable matching – i.e. a stable matching that all schools simultaneously prefer to all other stable matchings, and a stable matching that all students simultaneously prefer to all other stable matchings. Moreover, these matchings are easy to find, with the school-proposing (respectively, student proposing) version of the Gale-Shapley deferred acceptance algorithm [Gale and Shapley, 1962]. Unfortunately, the situation is not quite as nice when student incentives are taken into account. Even in the 1-to-1 matching case (i.e. when capacities $C_u = 1$ for all schools), there is no mechanism which makes truthful reporting of one's preferences a dominant strategy for both sides of the market [Roth, 1982]. In the many-to-one matchings case, things are even worse: an algorithm which finds the school optimal stable matching does not incentivize truthful reporting for either the students or the schools [Roth, 1984].

Because of this, a literature has emerged studying the incentive properties of stable matching algorithms under *large market assumptions* (e.g. [Immorlica and Mahdian, 2005, Lee, 2011, Kojima and Pathak, 2009]). In general, this literature has taken the following approach: make restrictive assumptions about the market (e.g. that students preference lists are only of constant length and are drawn uniformly at random), and under those assumptions, prove that an algorithm which computes exactly the school optimal stable matching makes truthful reporting a dominant strategy for a $1 - o(1)$ fraction of student participants (generally even under these assumptions, the schools still have incentive to misreport if the algorithm computes the school optimal stable matching).

In this paper we take a fundamentally different approach. We make absolutely no assumptions about student or school preferences, allowing them to be worst-case. We also insist on giving incentive guarantees to *every* student, not just most students. We compute (in a sense to be defined) an *approximately* stable and *approximately* school optimal matching using an algorithm with a particular insensitivity property (differential privacy), and show that truthful reporting is an *approximately* dominant strategy for *every* student in the market. These approximations become perfect as the size of the market grows large. Our notion of a "large market" requires only that the capacity of each school C_u grows with (the square root of) the number of schools, and (logarithmically) with the number of students, and does not require any assumption on how preferences are generated.

1.1 Our Results and Techniques

We recall the standard notion of stability in a many-to-one matching market with n students $a_i \in A$ and m schools $u_j \in U$, each with capacity C_j.

Definition 1. *A matching $\mu : A \to U$ is* feasible *and* stable *if:*
1. (Feasibility) For each $u_j \in U$, $|\{i : \mu(a_i) = u_j\}| \leq C_j$
2. (No Blocking Pairs with Filled Seats) For each $a_i \in A$, and each $u_j \in U$ such that $\mu(a_i) \neq u_j$, either $\mu(a_i) \succ_{a_i} u_j$ or for every student $a'_i \in \mu^{-1}(u_j)$, $a'_i \succ_{u_j} a_i$.
3. (No Blocking Pairs with Empty Seats) For every $u_j \in U$ such that $|\mu^{-1}(u_j)| < C_j$, and for every student $a_i \in A$ such that $a_i \succ_{u_j} \emptyset$, $\mu(a_i) \succ_{a_i} u_j$.

Our notion of approximate stability relaxes condition 3. Informally, we still require that there be no blocking pairs among students and *filled* seats, but we allow each school to possibly have a small number of empty seats. We view this as a mild condition, reflecting the reality that schools are not able to perfectly manage yield, and are often willing to accept a small degree of under-enrollment.

Definition 2 (Approximate Stability)**.** *A matching $\mu : A \to U$ is* feasible *and α-approximately* stable *if it satisfies conditions 1 and 2 (*Feasibility *and* No Blocking Pairs with Filled Seats*) and:*

3. (No Blocking pairs with Empty Seats at Under-Enrolled Schools) For every $u_j \in U$ such that $|\mu^{-1}(u_j)| < (1-\alpha)C_j$, and for every student $a_i \in A$ such that $a_i \succ_{u_j} \emptyset$, $\mu(a_i) \succ_{a_i} u_j$.

We also employ a strong notion of approximate dominant strategy truthfulness, related to first order stochastic dominance – informally, we say that a mechanism is η-approximately dominant strategy truthful if no agent can gain more than η in expectation (measured by *any* cardinal utility function consistent with his ordinal preferences) by misreporting his preferences to the mechanism.

Finally, we, define a notion of school optimality that applies to approximately stable matchings. Informally, we say that an approximately stable matching μ (in the above sense) is *school dominant* if when compared to the school optimal *exactly* stable matching μ', for every school u_j, every student a_i matched to u_j in μ is strictly preferred by u_j to any student matched to μ_j in μ' but not in μ.

We can now give an informal statement of our main result.

Theorem 1 (Informal)**.** *There is an algorithm for computing feasible and α-approximately stable* school dominant *matchings that makes truthful reporting an η-approximate dominant strategy for every student in the market, under the condition that for every school u, the capacity is sufficiently large, e.g.*

$$C_u \geq O\left(\frac{\sqrt{m}}{\eta\alpha} \cdot \text{polylog}(n)\right)$$

Remark 1. *Note that no assumptions are needed about either school or student preferences, which can be arbitrary. The only large market assumption needed is that the capacity C_u of each school is large. If, as the market grows, school capacities grow with the total number of schools at a rate of $\Omega(m^{1/2+\varepsilon})$ for any constant ε, then both η and α can be taken to tend to 0 in the limit.*

This result differs from the standard large market results in several ways. First, and perhaps most importantly, the result is worst-case over all possible preferences of both schools and students. Second, the guarantee states that *no* student may substantially gain by by misreporting her preferences; previous results [Immorlica and Mahdian, 2005, Kojima and Pathak, 2009] show that

only a subconstant fraction of students might have (substantial) incentive to deviate. In exchange for these strong guarantees, we relax our notion of stability and school optimality to approximate notions, which can be taken to be exact in the limit as the market grows large (under the condition that school capacities grow at a sufficiently fast rate).

When we do make one of the assumptions on student preferences made in previous work, we get stronger claims than the one above. For example, when the length of the preference lists of students are taken to be bounded, as they are in Immorlica and Mahdian [2005] and Kojima and Pathak [2009] we can remove our dependence on the number of schools:

Theorem 2 (Informal)**.** *Under the condition that all students have preference lists over at most k schools (and otherwise prefer to be unmatched), there is an algorithm for computing feasible and α-approximately stable* school dominant *matchings that makes truthful reporting an η-approximate dominant strategy for every student in the market, under the condition that for every school u, the capacity is sufficiently large:*

$$C_u \geq O\left(\frac{\sqrt{k}}{\eta\alpha} \cdot \text{polylog}(n)\right)$$

Remark 2. *Note that if k is considered to be a constant, then this result requires school capacity to grow only poly-logarithmically with the number of students n.*

Our results come from analyzing a differentially private variant of the classic deferred acceptance algorithm. Rather than having schools explicitly *propose* to students, we consider an equivalent variant in which schools u publish a set of "admissions thresholds" which allow any student a who is ranked higher than than the current threshold of school u (according to the preferences of u) to enroll. These thresholds naturally induce a matching when each student enrolls at their favorite school, given the thresholds. We first show that if the thresholds are computed under the constraint of differential privacy, then the algorithm is approximately dominant strategy truthful for the students. We then complete the picture by deriving a differentially private algorithm, and showing that with high probability, it produces an approximately stable, school dominant matching.

2 Related Work

2.1 Incentives in Stable Matching

Stable matching has long been known to be incompatible with truthfulness: no algorithm which produces a stable matching is truthful for both sides of the market [Roth, 1982], though Gale-Shapley is known to be truthful for the side of the market which is proposing in the 1-to-1 setting. Several lines of work have investigated stable matching in *large* markets, where players' preferences are drawn from some distribution, and considering properties of the market as n, the number of players, grows large. Much of this work reduces to arguing that few players will have more than one stable match; since only those players who have multiple stable matchings ever have incentive to misreport, this directly implies most players will have no incentive to misreport. For a tabular representation of the related work in terms of the expected number of stable matches an individual has under various assumptions, see Table 1.

Let $\mathcal{D}$ be a fixed distribution over the set of n women. Consider the following process of generating length-k preference lists over women. Draw some $w_1 \sim \mathcal{D}$, and let w_1 be the first woman in a preference list. Now, let $(w_1, \ldots, w_{i-1})$ be the first $(i-1)$ women, in order, drawn from

Reference	Assumptions	# of possible stable matches
Immorlica and Mahdian [2005]	Random, i.i.d. preferences on male side	$1 - o(1)$-fraction of women have ≤ 1 stable match
Kojima and Pathak [2009]	Random, i.i.d. preferences on student side	$1 - o(1)$-fraction of schools have more than 1 stable set of students
Pittel [1992]	Uniform random preferences on both sides	Average rank of parter for side optimized for is $\log(n)$, $\frac{n}{\log(n)}$ for the non-optimized side
Ashlagi et al. [2013]	Uniform random preferences on both sides, n men, $n-1$ women	Average rank for men's match $\frac{n}{3\log(n)}$, for women's match $3\log(n)$, in any stable matching

Figure 1: Related works with distributional assumptions on the preferences of one or both sides of the market. Under these assumptions, it is often possible to show that many agents have few (or one) stable partners. If an agent has zero or one stable partner, then she has no incentive to misreport.

$\mathcal{D}$. Draw $w_i \sim \mathcal{D}$ until $w_i \notin \{w_1, \ldots, w_{i-1}\}$. We denote such a distribution over preference lists by $\mathcal{D}^k$. Immorlica and Mahdian [2005] prove a generalization of a conjecture of Roth and Peranson [1999], showing if the men draw their preference lists according to $\mathcal{D}^k$, the expected number of women with more than one stable match is $o(n)$ (as n grows, for fixed k). Since it is known that a person has incentive to misreport only if they have more than one stable partner, this implies that only a vanishingly small fraction of the women will have incentive to misreport to any stable matching process. They also show that any stable matching algorithm induces a Nash equilibrium for which a $1 - o(1)$ fraction of players behave truthfully. These results are extended by Kojima and Pathak [2009] to the many-to-one matching setting. They show that student-optimal stable matchings, where colleges have arbitrary preferences, and the students have random preference lists of fixed length drawn as above, will have similar a subconstant fraction of schools which have incentive to misreport. Lee [2011] considers a slightly different distributional assumption about preferences in the one-to-one setting, where he shows that only a small number of players have large incentive to misreport.

Azevedo and Budish [2012] introduce the notion of "strategyproofness in the large", and show that the Gale-Shapley algorithm satisfies this definition. Roughly, this means that fixing any (constant sized) typespace, and any distribution over that typespace, if player preferences are sampled i.i.d. from the typespace, then for any fixed η, as the number of players n tends to infinity, truthful reporting becomes an η-approximate Bayes Nash equilibrium. These assumptions can be restrictive however – note that this kind of result requires that there are many more players than there are "types" of preferences, which in particular (together with the full support assumption on the type distribution) requires that in the limit, there are infinitely many identical agents of each type. In contrast, our results do not require a condition like this.

To the best of our knowledge, our results are the first to give truthfulness guarantees in settings where both sides of the market have worst-case preference orderings. Unlike some prior work, even without distributional assumptions, we are able to give truthfulness guarantees to *every* student,

not only a $1 - o(1)$ fraction of students. Under one of the assumptions used in prior work (namely, that the length of the preference lists is short), our results can be sharpened as well.

2.2 Differential Privacy as a Tool for Truthfulness

The study of differentially private algorithms [Dwork et al., 2006] has blossomed in recent years. A comprehensive survey of the work in this area is beyond the scope of this paper; here, we mention the work which relates directly to the use of differential privacy in constructing truthful mechanisms.

McSherry and Talwar [2007] were the first to identify privacy as a tool for designing approximately truthful mechanisms. Nissim et al. [2012] showed how privacy could be used as a tool to design *exactly* truthful mechanisms without needing monetary payments (in certain settings). Huang and Kannan [2012] proved that the exponential mechanism, a basic tool in differential privacy introduced in McSherry and Talwar [2007] is maximal in distributional range, which implies that there exist payments which make it exactly truthful. Kearns et al. [2014] demonstrated a connection between private equilibrium computation and the design of truthful *mediators* (and also showed how to privately compute approximate correlated equilibria in large games). This work was extended by Rogers and Roth [2014] who show how to privately compute *Nash* equilibria in large congestion games.

The paper most related to our own is Hsu et al. [2014] which shows how to compute approximate Walrasian equilibria privately, when bidders have quasi-linear utility for money and the supply of each good is sufficiently large. In that paper, in the final allocation, every agent is matched to their *approximately* most preferred goods at the final prices. In our setting, there are several significant differences: first, in the Walrasian equilibrium setting, only agents have preferences over goods (i.e. goods have no preferences of their own), but in our setting, both sides of the market have preferences. Second, although there is a conceptual relationship between "threshold scores" in stable matching problems and prices in Walrasian equilibria, the thresholds do not play the role of money in matching problems, and there is no notion of being matched to an "approximately" most preferred school.

3 Preliminaries

3.1 Many-to-one Matching

A many-to-one stable matching problem consist of m schools $U = \{u_1, \ldots, u_m\}$ and n students $A = \{a_1, \ldots, a_n\}$. Every student a has a preference ordering $\succ_a$ over all the schools, and each school u has a preference ordering $\succ_u$ over the students. Let $\mathcal{P}$ denote the domain of all preference orderings over schools (so each $\succ_a \in \mathcal{P}$).

It will be useful for us to think of a school u's ordering over students A as assigning a unique[1] *score* $\texttt{score}(u, a)$ to every student, in descending order (for example, these could be student scores on an entrance exam). Every school u has a capacity C_u, the maximum number of students the school can accommodate. A feasible matching μ is a mapping $\mu : A \to U \cup \emptyset$, which has the property each student a is paired with at most one school $\mu(a)$, and each school u is matched with

[1] It is essentially without loss of generality that students are assigned unique scores. If not, we could break ties by a simple pre-processing step: add noise $\sum_{k=1}^{l} 2^{-k} b_k$ to each student's score, where each b_k is a random bit; if the scores are integral, the probability of having ties is $1/\operatorname{poly}(n)$ as long as $l \geq O(\log(n))$.

at most C_u students: $|\mu^{-1}(u)| \leq C_u$. For notational simplicity, we will sometimes simply write $\mu(u)$ to denote the set of students assigned to school u.

A matching is α-approximately *stable* if it satisfies Definition 2. When computing matchings, it will be helpful for us to think instead about computing *admission thresholds* t_u for each school. A set of admission thresholds $t \in \mathbb{R}^m_{\geq 0}$ induces a matching μ in a natural way: every student $a \in A$ is matched to her most preferred school amongst those whose admissions thresholds are below her score at the school. Formally, for a set of admissions thresholds t, the induced matching μ^t is defined by:

$$\mu^t(a) = \arg\max_{\succ_a} \{u \mid \texttt{score}(u, a) \geq t_u\}$$

We say that a set of admission thresholds s is feasible and α-approximately stable if its induced matching $\mu^t(a)$ is feasible and α-approximately stable. Note that an α-stable matching is an exactly stable matching in a market in which schools have reduced capacity (where the capacity at each school is reduced by at most a $(1-\alpha)$ factor).

Remark Definition 2 also implies that if a school u is under-enrolled by more than αC_u, its admission score $t_u = 0$. This means such a school is very unpopular and could not recruit enough students even without any admission criterion.

We now introduce a notion of approximate school optimality, which our algorithm guarantees.

Definition 3. *A matching μ is* **school-dominant** *if, for each school u, for all $a \in \mu(u) \setminus \mu'(u)$ and all $a' \in \mu'(u) \setminus \mu(u)$, $a \succ_u a'$, where μ' is the school-optimal matching.*

In words, a matching μ is school-dominant if for every school u, when comparing the set of students S_1 that u is matched to in μ but not in the school optimal matching μ', and the set of students S_2 that u is matched to in the school optimal matching μ', but is not matched to in μ, u strictly prefers every student in S_1 to every student in S_2. (i.e. compared to the school optimal matching, a school may be matched to *fewer* students, but not to *worse* students.) We note that school-dominance alone is trivial to guarantee: in particular, the empty matching is school dominant. Only together with an upper bound on the number of empty seats allowed per school (for example, as guaranteed by α-approximate stability) is this a meaningful concept.

We want to give mechanisms that make it an *approximately dominant strategy* for students to report truthfully. We have to be careful about what we mean by this, since students a have ordinal preferences $\succ_a$, rather than cardinal utility functions $v_a : U \to [0, 1]$. We say that a cardinal utility function v_a is *consistent* with a preference ordering $\succ_a$ if for every $u, u' \in U$, $u \succ_a u'$ if and only if $v_a(u) \geq v_a(u')$. We will say that a mechanism is η-approximately truthful for students if for every student, and *every cardinal utility function v_a consistent with truthful $\succ_a$*, truthful reporting is an η-approximate dominant strategy as measured by v_a.

Definition 4. *Consider any randomized mapping $\mathcal{M} : \mathcal{P}^n \to U^n$. We say that $\mathcal{M}$ is η-approximately dominant strategy truthful (or truthful) if for any vector of student preferences $\succ \in \mathcal{P}^n$, any student a, any utility function $v_a : U \to [0, 1]$ that is consistent with $\succ_a$, and any $\succ'_a \neq \succ_a$, we have:*

$$\mathrm{E}_{\mu \sim M(\succ)}[v_a(\mu(a))] \geq \mathrm{E}_{\mu \sim M(\succ'_a, \succ_{-a})}[v_a(\mu(a))] - \eta.$$

Note that this definition is very strong, since it holds simultaneously for every utility function consistent with student preferences. When $\eta = 0$ it corresponds to first order stochastic dominance.

3.2 Differential Privacy Preliminaries

Our tool for obtaining approximate truthfulness is *differential privacy*, which we define in this section. We say that the "private data" of each student a consists of both her preference ordering $\succ_a \in \mathcal{P}$ over the schools and the scores $\texttt{score}(u, a) \in \mathcal{V}$ assigned by the schools. A private database $D \in (\mathcal{P} \times \mathcal{V})^n$ is a vector of n student profiles, and D and D' are neighboring databases if they differ in no more than one student record. In particular, our matching algorithms take n student profiles as input and produce a set of admission scores as output (i.e., range $\mathcal{R} = \mathcal{V}^m$).

Definition 5 (Dwork et al. [2006])**.** *An (randomized) algorithm $\mathcal{A}$: $(\mathcal{P} \times \mathcal{V})^n \to \mathcal{R}$ is* (ε, δ)-differentially private *if for every pair of neighboring databases $D, D' \in (\mathcal{P} \times \mathcal{V})^n$ and for every set of subset of outputs $S \subseteq \mathcal{R}$,*

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D') \in S] + \delta.$$

If $\delta = 0$, we say that $\mathcal{M}$ is ε-differentially private*.*

3.3 Differentially Private Counters

The central privacy tool in our matching algorithm is the private streaming counter(for a more detailed discussion of differential privacy under continual observation, see Chan et al. [2011] and Dwork et al. [2010a]) proposed by Chan et al. [2011] and Dwork et al. [2010a]. Given a bit stream $\sigma = (\sigma_1, \ldots, \sigma_T) \in \{-1, 0, 1\}^T$, a streaming counter $\mathcal{M}(\sigma)$ releases an approximation to $c_\sigma(t) = \sum_{i=1}^{t} \sigma_i$ at every time step t. Below, we define an accuracy property we will then use to describe the usefulness of these counters.

Definition 6. *A streaming counter $\mathcal{M}$ is* (τ, β)-useful *if with probability at least $1 - \beta$, for each time $t \in [T]$,*

$$|\mathcal{M}(\sigma)(t) - c_\sigma(t)| \leq \tau.$$

For the rest of this paper, let **Counter**(ε, T) denote the Binary Mechanism of Chan et al. [2011], instantiated with parameters ε and T. **Counter**(ε, T) satisfies the following accuracy guarantee (further details may be found in Appendix A.2). Our mechanism uses m different **Counter**s to maintain the counts of temporally enrolled students for all schools. The following theorem allows us to bound the error of each **Counter** through the collective sensitivity across all **Counter**s.

Theorem 3. *Suppose we have m bit streams such that the change of an agent's data affects at most k streams, and alters at most c bits in each stream. For any $\beta > 0$, the composition of m distinct* **Counter** $\left(\varepsilon/2c\sqrt{2kc\ln(1/\delta)}, T\right)$*s is (ε, δ)-differentially private, and (α, β)-useful for*

$$\alpha = \frac{16c\sqrt{kc\ln(1/\delta)}}{\varepsilon} \ln\left(\frac{2m}{\beta}\right)\left(\sqrt{\log(T)}\right)^5.$$

4 Algorithms Computing Private Matchings are Approximately Truthful

In this section, we prove the theorem which motivates the rest of our paper. Consider an algorithm M which takes as input student preferences $\succ$ and computes school thresholds s. If M is ε-differentially private, then the algorithm $A(\succ)$ which computes thresholds $s = M(\succ)$ and then

outputs the induced matching μ^s is ε-approximately dominant strategy truthful. Note that this guarantee holds independent of stability. For lack of space, we relegate the proof to Appendix B.

Theorem 4. *Let $M : \mathcal{P}^n \to \mathbb{R}^m_{\geq 0}$ be any (ε, δ)-differentially private mechanism which takes as input n student profiles and outputs m school thresholds. Let $A : \mathbb{R}^m_{\geq 0} \to U^n$ be the mechanism which takes as input m school thresholds s, and outputs the corresponding matching $A(s) = \mu^s$. Then the mechanism $A \circ M : \mathcal{P}^n \to U^n$ is $(\varepsilon + \delta)$-approximately dominant strategy truthful.*

The intuition behind this theorem is simple: if a mechanism is private, a student's report has almost no effect on the realization of the school thresholds; Given a fixed set of school thresholds, he can do no better than reporting truthfully, which causes him to be matched to his most preferred school that will take him.

Proof of Theorem 4. Fix any vector of student preferences $\succ$, any player a, any utility function v_a consistent with $\succ_a$, and any deviation $\succ'_a \neq \succ_a$. Now consider player a's utility for truth-telling. For $\varepsilon \leq 1$, we have

$$\begin{aligned}
\mathbb{E}_{\mu \sim A \circ M(\succ)}[v_a(\mu(a))] &= \mathrm{E}_{s \sim M(\succ)}[v_a(\arg\max_{\succ_a}\{u \mid \texttt{score}(u,a) \geq s_u\})] \\
&= \sum_s \Pr[M(\succ) = s] \cdot v_a(\arg\max_{\succ_a}\{u \mid \texttt{score}(u,a) \geq s_u\}) \\
&\geq \sum_s \exp(-\varepsilon) \Pr[M(\succ'_a, \succ_{-a}) = s] \cdot v_a(\arg\max_{\succ_a}\{u \mid \texttt{score}(u,a) \geq s_u\}) - \delta \\
&= \exp(-\varepsilon)\mathrm{E}_{s \sim M(\succ'_a, \succ_{-a})}[v_a(\arg\max_{\succ_a}\{u \mid \texttt{score}(u,a) \geq s_u\})] - \delta \\
&\geq \exp(-\varepsilon)\mathrm{E}_{s \sim M(\succ'_a, \succ_{-a})}[v_a(\arg\max_{\succ'_a}\{u \mid \texttt{score}(u,a) \geq s_u\})] - \delta \\
&\geq (1-\varepsilon)\mathrm{E}_{s \sim M(\succ'_a, \succ_{-a})}[v_a(\arg\max_{\succ'_a}\{u \mid \texttt{score}(u,a) \geq s_u\})] - \delta \\
&\geq \mathrm{E}_{\mu \sim M(\succ'_a, \succ_{-a})}[v_a(\mu(a))] - (\varepsilon + \delta).
\end{aligned}$$

where the first and last equalities follow from the definition of the induced matching μ^s, the first inequality follows from the differential privacy condition, and the second follows from the consistency of v_a with $\succ_a$. □

5 Truthful School-Optimal Mechanism

In this section, we present the algorithm which proves our main result Theorem 1. Algorithm 1 computes an α-approximately stable and school-dominant matching, and enjoys approximate dominant strategy truthfulness for the student side. We assume the reader is familiar with `DA-School`, the well-known school-proposing version of the deferred acceptance algorithm [Gale and Shapley, 1962]. For a brief overview of `DA-School` within our context of score thresholds, see Section D. We now state a useful fact about deferred acceptance, whose proof can be found in Appendix C.

Lemma 1. *Let μ_t be some matching which is an intermediate matching in a run of the school-proposing deferred acceptance algorithm. Then μ_t is school-dominant.*

Our algorithm, `Private-DA-School`(ε, δ), is a private version of `DA-School`. At each time t, each school will publish a threshold score (initially, for each school, this will be the maximum

possible score for that school). Schools will lower their thresholds when they are under capacity; as they do so, some students will tentatively accept admission and some will reject or leave for other schools. Initially, all students will be unmatched. For a given student a, as soon as a school lowers its threshold below the score a has there, a will signal to the mechanism which school is her favorite of those for which her score passes their threshold. Then, as the schools continue to lower their thresholds to fill seats, if a school that a likes better than her current match lowers its threshold below her score, a will inform the mechanism that she wishes to switch to her new favorite.

Each school maintains a private counter of the number of students tentatively matched to the school. We let E be the additive error bound of the counters. The schools will reserve E number of seats from their initial capacity to avoid being over-enrolled, so the algorithm is run as if the capacity at each school is $C_u - E$. Then each school can be potentially under-enrolled by $2E$ seats, but they would take no more than α fraction of all the seats as long as the capacity $C_u \geq 2E/\alpha$.

Algorithm 1: `Private-DA-School`(ε, δ)

Input: school capacities $\{C_u\}$, student preferences $\{\succ_a\}$ and scores $\{\texttt{score}(u, a)\}$, range of scores $[0, J]$

Output: a set of score thresholds $\{t_j\}$

initialize: **for each** school u_j and each student a_i

$T = mnJ, \qquad \varepsilon' = \frac{\varepsilon}{16\sqrt{2m\ln(1/\delta)}}, \qquad E = \frac{128\sqrt{m\ln(1/\delta)}}{\varepsilon}\ln\left(\frac{2m}{\beta}\right)\left(\sqrt{\log(nT)}\right)^5,$

$\text{counter}(u_j) = \textbf{Counter}(\varepsilon', nT), \qquad t_j = J, \qquad \mu(a_i) = \emptyset, \qquad \widehat{C}_{u_j} = C_{u_j} - E.$

while there is some under-enrolled school u_j: $\text{counter}(u_j) < \widehat{C}_{u_j}$ and $t_{u_j} > 0$ **do**

  $t_{u_j} = t_{u_j} - 1$

  for all student a_i **do**

    if $\mu(a_i) \neq \text{argmax}_{\succ_{a_i}}\{u_j \mid \texttt{score}(u_j, a_i) \geq t_{u_j}\}$ **then**

      Send (-1) to counter$(\mu(a_i))$

      let $\mu(a_i) = \text{argmax}_{\succ_{a_i}}\{u_j \mid \texttt{score}(u_j, a_i) \geq t_{u_j}\}$

      Send 1 to counter$(\mu(a_i))$

      Send 0 to all other counters

    else

      Send 0 to all counters

return Final threshold scores $\{t_j\}$ and matching μ^t

Now, we state the formal version of Theorem 1.

Theorem 1 *Algorithm 1 is (ε, δ)-differentially private, and hence $(\varepsilon + \delta)$-approximately dominant strategy truthful. With probability at least $1 - \beta$, it outputs an α-approximately stable, school dominant matching, as long as the capacity at each school u satisfies $C_u \geq R = O\left(\frac{\sqrt{m}}{\varepsilon\alpha}\,\text{polylog}\left(n, m, \frac{1}{\delta}, \frac{1}{\beta}\right)\right)$.*

We prove Theorem 1 in two parts. Lemma 2 shows that `Private-DA-School`(i, s) (ε, δ)-differentially private in the preferences of the students. Lemma 3 shows that the resulting matching is school-dominant so long as the capacity at each school is large enough. These two together imply Theorem 1 directly.

Lemma 2. *`Private-DA-School`(ε, δ) is (ε, δ)-differentially private.*

We relegate the proof of Lemma 2 to Appendix B for lack of space. The intuition behind the proof is as follows. Once a student leaves a school, it will never return to that school; thus, the sensitivity of all m counters to a given student is at most $2m$. The proof formalizes the sense in which `Private-DA-School`(ε, δ) has sensitivity at most $2m$ to a particular student.

Lemma 3. *With probability at least $1-\beta$, `Private-DA-School`(ε, δ) outputs an α-approximately stable, school-dominant matching, as long as the capacity at each school u satisfies $C_u \geq R = O\left(\frac{\sqrt{m}}{\varepsilon\alpha}\operatorname{polylog}\left(n, m, \frac{1}{\delta}, \frac{1}{\beta}\right)\right)$.*
If the maximum error of each of the collection of counters is bounded by x with probability at least, $1-\beta$, we need only $C_u \geq O(x)$

Proof of Lemma 3. We prove that the output thresholds $\{t_{u_j}\}$ induce an α-approximately stable, school-dominant matching μ^t.

We claim that there can be no blocking pairs with filled seats in μ^t. Suppose some student a_i wishes to attend u_j. Then, it is either the case that $\texttt{score}(u_j, a_i) \geq t_{u_j}$ or $\texttt{score}(u_j, a_i) < t_{u_j}$. In the first case, a_i cannot block with u_j: she could have gone to u_j and chose a school she preferred to u_j. In the second case, consider some $a_{i'}$ such that $u_j = \mu^t(a_{i'})$; this implies $\texttt{score}(u_j, a_{i'}) \geq t_{u_j}$. Thus, $\texttt{score}(u_j, a_{i'}) > \texttt{score}(u_j, a_i)$, so $a_{i'} \succ_{u_j} a_i$, and a_i doesn't block with $a_{i'}$, so there are no blocking pairs with filled seats.

By Theorem 3 and union bound, we know that the error of all m counters over all time steps is bounded by E except with probability β, where

$$E = \frac{128\sqrt{m\ln(1/\delta)}}{\varepsilon}\ln\left(\frac{2m}{\beta}\right)\left(\sqrt{\log(nmJ)}\right)^5.$$

So, we condition on the event that all schools' counters are accurate within E throughout the run of `Private-DA-School`(ε, δ) for the remainder of our argument.

We first claim that no school is over-enrolled in μ^t. Consider the last time u_j lowered its threshold to t_{u_j}. Let n_{u_j} denote the number of students tentatively matched to u_j just prior to the final lowering of t_{u_j}. By definition, u_j only lowers its threshold when $\textbf{Counter}(u_j) < \widehat{C}_{u_j} = C_{u_j} - E$, so

$$C_{u_j} - E = \widehat{C}_{u_j} > \textbf{Counter}(u_j) \geq n_{u_j} - E \geq |\mu^t(u_j)| - E - 1$$

where the first equality is by definition, the first inequality comes from the fact that u_j lowered its threshold, the third from the accuracy we've conditioned on from the counters, and the final from the fact that u_j never again lowers its threshold. Thus, $C_{u_j} \geq |\mu^t(u_j)|$, and u_j is not over-enrolled.

Now, we show no school is under-enrolled by more than $2E$, unless $t_{u_j} = 0$. When the algorithm terminates, each school u_j either has a threshold $t_{u_j} = 0$ or

$$|\mu^t(u_j)| + E \geq \textbf{Counter}(u_j) \geq \widehat{C}_{u_j} = C_{u_j} - E$$

where the first equality comes from the conditional bound on the error of the counters, the second from the fact that the algorithm terminated, and the final one from the definition of $\widehat{C}_{u_j}$. Thus, $|\mu(u_j)| \geq C_{u_j} - 2E$ whenever $t_{u_j} > 0$, so no school is under-enrolled by more than an α-fraction of its seats so long as $C_{u_j} \geq 2E/\alpha$.

Finally, we show school dominance. We will now show that μ, the matching corresponding to the thresholds output by `Private-DA-School`(ε, δ), is also achieved by running `DA-School` on the same instance, and halting early. No school is over-enrolled, by our argument above, *at any point*

during the run of the algorithm. So, each proposal made by u_j would be a valid proposal to make in `DA-School` with full capacity. Thus, `Private-DA-School`(t, e) rminates with each school having made (weakly) fewer proposals than it would have in `DA-School`. Since each school makes its proposals in the same order (according to $\succ_{u_j}$), this implies that μ^t is a matching that corresponds to some intermediate point in `DA-School` using with the same ordering of proposals. Thus, by Lemma 6, μ is school-dominant (and our argument is entirely parametric in E, so the second part of the claim follows directly). □

Now, we present the formal version Theorem 2. Without loss of generality, we can assume algorithm ignores students after they have accepted k or more schools' proposals.

Theorem 2. *Suppose each student has a preference list of length at most k. Then, School-Proposing$\left(\frac{\varepsilon\sqrt{m}}{2\sqrt{k}}, \delta\right)$ is (ε, δ)-private and thus $\varepsilon + \delta$-approximately truthful. With probability at least $1 - \beta$, it outputs an α-approximately stablem school-dominant matching, as long as the capacity at each school $C_u \geq O\left(\frac{\sqrt{k}}{\alpha\varepsilon} \text{polylog}\left(n, m, \frac{1}{\delta}, \frac{1}{\beta}\right)\right)$.*

We relegate the formal proof of Theorem 2 to Section B; the intuition is simple enough after the proof of Theorem 1. When students have much shorter preference lists, this greatly reduces the sensitivity of the counters to a single student's responses (from $\Theta(m)$ to $\Theta(k)$, where k is the maximum length of the preference list). This allows us to more tightly concentrate the error from the counters while maintaining differential privacy.

6 Conclusions

In this paper we applied differential privacy as a tool to design a many-to-one stable matching algorithm with strong incentive guarantees for the student side of the market. To the best of our knowledge, our work is the first work to show positive truthfulness results for the non-optimal side of the market, under *worst-case* preferences, for *all* participants on the non-optimal side of the market.

Additionally, although we have not focused on this, our algorithm also provides strong *privacy* guarantees to the students. Each student, upon learning the school thresholds (and hence the school that she herself is matched to) can learn almost nothing about either the preferences or scores of the other students (i.e. almost nothing about the preferences that the other students hold over schools, or the preferences that schools hold over the other students). Here "almost nothing" is the formal guarantee of differential privacy, which in particular implies that for every student a, no matter what her prior belief over the private data of some other student a' is, her posterior belief over a's data would be almost the same in the two worlds in which a' participates in the mechanism, and in which she does not. These guarantees might themselves be valuable in settings in which the matching being computed is sensitive – e.g. when computing a matching between patients and drug trials, for example.

References

Atila Abdulkadiroglu, Parag A Pathak, and Alvin E Roth. Strategy-proofness versus efficiency in matching with indifferences: redesigning the new york city high school match. Technical report, National Bureau of Economic Research, 2009.

Itai Ashlagi, Yashodhan Kanoria, and Jacob D Leshno. Unbalanced random matching markets. In *EC*, pages 27–28, 2013.

Eduardo Azevedo and Eric Budish. Strategyproofness in the large. Technical report, Working paper, 2012.

T-H Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. *ACM Transactions on Information and System Security (TISSEC)*, 14(3):26, 2011.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer, 2006.

Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N Rothblum. Differential privacy under continual observation. In *Proceedings of the 42nd ACM symposium on Theory of computing*, pages 715–724. ACM, 2010a.

Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *FOCS*, pages 51–60, 2010b.

David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *American Mathematical Monthly*, pages 9–15, 1962.

Justin Hsu, Zhiyi Huang, Aaron Roth, Tim Roughgarden, and Zhiwei Steven Wu. Private matchings and allocations. In *STOC*, pages 21–30. ACM, 2014.

Zhiyi Huang and Sampath Kannan. The exponential mechanism for social welfare: Private, truthful, and nearly optimal. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 140–149. IEEE, 2012.

Nicole Immorlica and Mohammad Mahdian. Marriage, honesty, and stability. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 53–62. Society for Industrial and Applied Mathematics, 2005.

Michael Kearns, Mallesh Pai, Aaron Roth, and Jonathan Ullman. Mechanism design in large games: Incentives and privacy. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 403–410. ACM, 2014.

Fuhito Kojima and Parag A Pathak. Incentives and stability in large two-sided matching markets. *The American Economic Review*, pages 608–627, 2009.

SangMok Lee. Incentive compatibility of large centralized matching markets. *Job market paper*, 2011.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 94–103. IEEE, 2007.

Kobbi Nissim, Claudio Orlandi, and Rann Smorodinsky. Privacy-aware mechanism design. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 774–789. ACM, 2012.

Boris Pittel. On likely solutions of a stable marriage problem. *The Annals of Applied Probability*, pages 358–401, 1992.

Ryan M Rogers and Aaron Roth. Asymptotically truthful equilibrium selection in large congestion games. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 771–782. ACM, 2014.

Alvin E Roth. The economics of matching: Stability and incentives. *Mathematics of Operations Research*, 7(4):617–628, 1982.

Alvin E Roth. The evolution of the labor market for medical interns and residents: a case study in game theory. *The Journal of Political Economy*, 92(6):991, 1984.

Alvin E Roth. The college admissions problem is not equivalent to the marriage problem. *Journal of economic Theory*, 36(2):277–288, 1985.

Alvin E Roth and Elliott Peranson. The redesign of the matching market for american physicians: Some engineering aspects of economic design. Technical report, National bureau of economic research, 1999.

A Privacy Analysis for Counters

Theorem 5 (Chan et al. [2011])**.** *For $\beta > 0$,* **Counter**(ε, T) *is ε-differentially private with respect to a single bit change in the stream, and (α, β)-useful for*

$$\alpha = \frac{4\sqrt{2}}{\varepsilon} \ln\left(\frac{2}{\beta}\right) \left(\sqrt{\log(T)}\right)^5.$$

Chan et al. [2011] show that **Counter**(ε, T) is ε-differentially private with respect to single changes in the input stream, when the stream is generated non-adaptively. For our application, we require privacy to hold for a large number of streams whose joint-sensitivity can nevertheless be bounded, and whose entries can be chosen adaptively. To show that **Counter** is also private in this setting (when ε is set appropriately), we first present a slightly more refined composition theorem.

A.1 Composition

An important property of differential privacy is that it degrades gracefully when private mechanisms are composed together, even adaptively. We recall the definition of an adaptive composition experiment due to Dwork et al. [2010b].

Definition 7 (Adaptive composition experiment)**.**

- *Fix a bit $b \in \{0, 1\}$ and a class of mechanisms $\mathcal{M}$.*
- *For $t = 1 \ldots T$:*
  - *The adversary selects two databases $D^{t,0}, D^{t,1}$ and a mechanism $\mathcal{M}_t \in \mathcal{M}$.*
  - *The adversary receives $y_t = \mathcal{M}_t(D^{t,b})$*

The "output" of an adaptive composition experiment is the view of the adversary over the course of the experiment. The experiment is said to be ε-differentially private if

$$\max_{S \subseteq \mathcal{R}} \frac{\Pr[V^0 \in S]}{\Pr[V^1 \in S]} \leq \exp(\varepsilon),$$

and (ε, δ)-differentially private if

$$\max_{S \subset \mathcal{R}, \Pr[V^0 \in S] \geq \delta} \frac{\Pr[V^0 \in S] - \delta}{\Pr[V^1 \in S]} \leq \exp(\varepsilon),$$

where V^0 is the view of the adversary with $b = 0$, V^1 is the view of the adversary with $b = 1$, and $\mathcal{R}$ is the range of outputs.

Any algorithm that can be described as an instance of this adaptive composition experiment (for an appropriately defined adversary) is said to be an instance of the class of mechanisms $\mathcal{M}$ under *adaptive T-fold composition.*

A very useful tool to analyze private algorithms is the following theorem that allows us to analyze the "composition" of private algorithms.

Theorem 6 (Adaptive Composition Dwork et al. [2010b])**.** *Let $\mathcal{A} \colon \mathcal{U} \to \mathcal{R}^T$ be a T-fold adaptive composition*[2] *of (ε, δ)-differentially private algorithms. Then $\mathcal{A}$ satisfies $(\varepsilon', T\delta + \delta')$-differential privacy for*

$$\varepsilon' = \varepsilon\sqrt{2T\ln(1/\delta')} + T\varepsilon(e^{\varepsilon} - 1).$$

In particular, for any $\varepsilon \leq 1$, if $\mathcal{A}$ is a T-fold adaptive composition of $(\varepsilon/\sqrt{8T\ln(1/\delta)}, 0)$-differentially private mechanisms, then $\mathcal{A}$ satisfies (ε, δ)-differential privacy.

For a more refined analysis in our setting, we now state a straightforward consequence of a composition theorem of Dwork et al. [2010b].

Lemma 4 (Dwork et al. [2010b])**.** *Let $\Delta \geq 0$. Under adaptive composition, the class of $\frac{\varepsilon}{\Delta}$-private mechanisms satisfies ε-differential privacy and the class of $\frac{\varepsilon}{2c\sqrt{2\Delta\ln(1/\delta)}}$-private mechanisms satisfies (ε, δ)-differential privacy, if the adversary always selects databases satisfying*

$$\textit{for all } t \ \left|D^{t,0} - D^{t,1}\right| \leq c, \textit{ and also } \ \sum_{t=1}^{T} \left|D^{t,0} - D^{t,1}\right| \leq \Delta.$$

In other words, the privacy parameter of each mechanism should be calibrated for the total distance between the databases, over the whole composition. This is useful for analyzing the privacy of the counters in our algorithm, which collectively have bounded sensitivity.

A.2 Details for Counters

We reproduce Binary mechanism here in order to refer to its internal workings in our privacy proof.

First, it is worth explaining the intuition of the **Counter**. Given a bit stream $\sigma \colon [T] \to \{-1, 0, 1\}$, the algorithm releases the counts $\sum_{i=1}^{t} \sigma(i)$ for each t by maintaining a set of partial

[2] See Appendix A.1 and [Dwork et al., 2010b] for further discussion.

sums $\sum[i,j] \coloneqq \sum_{t=i}^{j} \sigma(t)$. More precisely, each partial sum has the form $\sum[2^i+1, 2^i+2^{i-1}]$, corresponding to powers of 2.

In this way, we can calculate the count $\sum_{i=1}^{t} \sigma(i)$ by summing at most $\log t$ partial sums: let $i_1 < i_2 \ldots < i_m$ be the indices of non-zero bits in the binary representation of t, so that

$$\sum_{i=1}^{t} \sigma(i) = \sum[1, 2^{i_m}] + \sum[2^{i_m}+1, 2^{i_m}+2^{i_{m-1}}] + \ldots + \sum[t-2^{i_1}+1, t].$$

Therefore, we can view the algorithm as releasing partial sums of different ranges at each time step t and computing the counts is simply a post-processing of the partial sums. The core algorithm is presented in Algorithm 2.

Algorithm 2: Counter(ε, T)

Input: A stream $\sigma \in \{-1,1\}^T$
Output: $B(t)$ as estimate for $\sum_{i=1}^{t} \sigma(i)$ for each time $t \in [T]$
for all $t \in [T]$ **do**
  Express $t = \sum_{j=0}^{\log t} 2^j \mathrm{Bin}_j(t)$.
  Let $i \leftarrow \min_j \{\mathrm{Bin}_j(t) \neq 0\}$
  $a_i \leftarrow \sum_{j<i} a_j + \sigma(t)$, $(a_i = \sum[t-2^i+1, t])$
  for $0 \le j \le i-1$ **do**
    Let $a_j \leftarrow 0$ and $\widehat{a_j} \leftarrow 0$
  Let $\widehat{a_j} = a_j + \mathrm{Lap}(\log(T)/\varepsilon)$
  Let $B(t) = \sum_{i:\mathrm{Bin}_i(t)\neq 0} \widehat{a_i}$

A.3 Counter Privacy under Adaptive Composition

Theorem 3. *Suppose we have m bit streams such that the change of an agent's data affects at most k streams, and alters at most c bits in each stream. For any $\beta > 0$, the composition of m distinct* **Counter** $\left(\varepsilon/2c\sqrt{2kc\ln(1/\delta)}, T\right)$*s is (ε,δ)-differentially private, and (α,β)-useful for*

$$\alpha = \frac{16c\sqrt{kc\ln(1/\delta)}}{\varepsilon} \ln\left(\frac{2m}{\beta}\right) \left(\sqrt{\log(T)}\right)^5.$$

Proof. The composition of m counters is essentially releasing a collection of noisy partial sums adaptively. We need to first frame this setting as an advanced composition experiment defined in Definition 7. First, we treat each segment $\sigma[a,b]$ in a stream as a database. For each such database, we are releasing the sum by adding noise sampled from the Laplace distribution:

$$\mathrm{Lap}\left(\frac{2c\sqrt{2kc\ln(1/\delta)}\log(T)}{\varepsilon}\right),$$

which is $\frac{\varepsilon}{2c\sqrt{2kc\ln(1/\delta)\log(T)}}$-private mechanism (w.r.t. a single bit change). We know that changing an agent's data changes at most c bits in each stream, and affects at most k streams, and also each

bit change can result in $\log(T)$ bits changes across different stream-segment databases. Therefore, we can bound the total distance between all pairs stream-segment databases by

$$\Delta \leq kc \log(T).$$

By Lemma 4, we know that the composition of all m counters under our condition satisfies (ε, δ)-differential privacy.

Plugging in our choice of ε to the accuracy proof for **Counter** in Chan et al. [2011], we obtain our accuracy guarantee by applying union bound. □

B Omitted Proofs

Proof of Lemma 2. `Private-DA-School`(ε, δ) outputs a sequence of sets of thresholds and nothing else. We will construct a mechanism $\mathcal{M}$, which will output the same sequence of thresholds as `Private-DA-School`(ε, δ), for which it is more obvious to prove (ε, δ)-differential privacy. This will imply (ε, δ)-differential privacy of `Private-DA-School`(ε, δ). Here is the definition of $\mathcal{M}$:

Algorithm 3: $\mathcal{M}$

Publish threshold $t_{u_j} = J$ for each school u_j;
$\varepsilon' = \frac{\varepsilon}{12\sqrt{2m \ln \frac{1}{\delta}}}$;
Initialize counter(u_j) = **Counter**(ε', nmJ);
Let $\widehat{C}_{u_j} = C_{u_j} - E$;
while *there is some under-enrolled school* u_j*: counter*$(u_j) \leq \widehat{C}_{u_j}$ *and* $t_{u_j} > 0$ **do**
    Let $t'_{u_j} = t_{u_j} - 1$;
    Publish thresholds $(t_{u_1}, \ldots, t'_{u_j}, \ldots, t(u_m))$;
    Receive bits $b_{u'_j} \in \{-1, 0, 1\}$ for each $u_{j'}$;
    Send b_{u_j} to counter(u_j);

We define the input bits to the algorithm $\mathcal{M}$ as follows. For a fixed execution of the while loop, we will define the bits $b_{u_{j'}}$ to give to $\mathcal{M}$. Let u_j be the school which lowered its threshold in this timestep. Let $b_{u_j} = 1$ if and only if, for the unique student a_i such that `score`$(u_j, a_i) = t'_{u_j}$, it is true that $u_j = \text{argmax}_{\succ_{a_i}}\{u \mid \texttt{score}(u, a_i) \geq t_{u_j}\}$ (a_i prefers u_j to all other schools for which her score surpasses the threshold). Let $b_{u_{j'}} = -1$ if and only if $b_{u_j} = 1$ and also $u_{j'} = \text{argsecondmax}_{\succ_{a_i}}\{u \mid \texttt{score}(u, a_i) \geq t_{u_j}\}$ (u_j is a_i's favorite available school and $u_{j'}$ is her second favorite). For all other j'', let $b_{u_{j''}} = 0$.

Then, there are at most $2m$ nonzero bits sent to $\mathcal{M}$ about a particular student a_i, and at most 2 nonzero bits sent by a particular a_i to any school u_j. These bits are the only interface $\mathcal{M}$ has with private data. Furthermore, $\mathcal{M}$ and `Private-DA-School`(ε, δ) have the same distribution over output data. So it suffices to show that $\mathcal{M}$ is (ε, δ)-differentially private.

Let $f : \{J\}^m \times [n] \to \{J\}^m$ be the function that, as a function of the previous thresholds and counter values, outputs the new set of thresholds at each time t. Then, the thresholds published by $\mathcal{M}$ are a composition of f, m instantiations of **Counter** (ε', nmJ), and previously computed data. Thus, it suffices to show the composition of the m counters satisfy (ε, δ)-differential privacy. By

construction, each school u_j receives at most 2 nonzero bits from a given student, and no student's data creates more than $2m$ nonzero bits in all streams together. By Theorem 6 and Lemma 4, the composition of m **Counter**(ε', nmJ) satisfy (ε, δ)-differential privacy when no stream has more than 2 bits affected by a single agent's data and no student has more than $2m$ total nonzero bits in any stream. Thus, $\mathcal{M}$ (and also `Private-DA-School`(ε, δ)) satisfies (ε, δ) differential privacy. □

Proof of Theorem 2. We prove (ε, δ)-privacy, which again reduces to proving (ε, δ)-differential privacy of the set of m counters. By a simple calculation, School-Proposing$(\frac{\varepsilon\sqrt{m}}{2\sqrt{k}}, \delta)$ uses

$$\varepsilon' = \frac{\varepsilon}{4\sqrt{k \ln \frac{1}{\delta}}}$$

as the privacy parameter for the m counters it uses. Theorem 6 states that a collection of m **Counter**(ε', nT) with total sensitivity Δ (and individual sensitivity c) satisfy (ε, δ)-differential privacy so long as

$$\varepsilon' \leq \frac{\varepsilon}{2c\sqrt{\Delta \ln \frac{1}{\delta}}}.$$

Remark 2 limits the total amount of sensitivity School-Propose will have; a student will be able to affect at most $2k$ bits in the input stream, so $\Delta \leq 2k$, and at most 2 per school, so $c \leq 2$. Thus, it suffices to use privacy parameter

$$\varepsilon' \leq \frac{\varepsilon}{4\sqrt{k \ln \frac{1}{\delta}}},$$

so our algorithm is (ε, δ)-differentially private. Furthermore, Theorem 6 and a union bound imply the maximum error any one of the counters will have at any time during the execution of the algorithm is

$$E \leq \frac{128\sqrt{k \ln(1/\delta)}}{\varepsilon} \ln\left(\frac{2m}{\beta}\right) \left(\sqrt{\log(nmJ)}\right)^5.$$

with probability $1 - \beta$. Thus, by Lemma 3, we get the desired guarantee for α-approximate stability and school-dominance. □

C Proofs of Matching Lemmas

We state one more lemma which we will use in the proofs of Lemmas 6.

Lemma 5. *Consider a set P_u of proposals made by each school u according to some prefix of school-proposing DA. Let $P \subseteq A \times U$ be the set of proposals made by all schools. Then, the matching μ which results from P is unique (and independent of the order in which proposals are made), assuming students are truthful.*

Proof. Each student ultimately accepts her most preferred proposal among the set of proposals she has received, independent of their ordering. (i.e. admissions thresholds only descend, and she picks her most preferred school amongst those schools with thresholds below her scores). Thus, each

school u will be matched to the subset of P_u which finds u to be their favorite offer, independent of the order in which proposals were made. □

Lemma 6. *Let μ_t be some matching which is an intermediate matching in a run of the school-proposing deferred acceptance algorithm. Then μ_t is school-dominant.*

Proof. The school-proposing deferred acceptance algorithm is somewhat underspecified. In particular, if multiple schools have space remaining, the *order* in which those schools make proposals isn't predetermined. But, by Lemma 5 shows that reordering of the same proposals from the schools will arrive at the same matching. Thus, it suffices to show, for a fixed ordering of the entire set of proposals made by DA, that each intermediate matching is school-dominant.

Let t denote the time at which we wish to halt a run of DA. Let $P_{u,t}$ denote the set of proposals which school u has made according some fixed ordering up to time t, and P_u denote set of proposals made by school u according to the entire run of DA. Let μ_t denote the "current" matching according to the first run of DA stopped at time t and μ denote the final outcome of DA.

Consider any school A. Notice that, since $|P_u| \geq |P_{u,t}|$, by the definition of DA,

$$P_{u,t} \subseteq P_u \tag{1}$$

since, for a given school, the proposing order is just working down their preference list.

Now consider a particular school u. We must show that for each $a \in \mu_t(u) \backslash \mu(u), a' \in \mu(u) \backslash \mu_t(u)$, $a \succ_u a'$. If a was proposed to by u in P_t and rejects u, then u will be rejected by a when she receives a superset P of proposals. Thus, the only students u has according to μ but not μ_t are students $z \in P_u \backslash P_{u,t}$ (students who are proposed to after time t). But, by the definition of DA, if u proposes to two students a and a', and proposes to a before a', $a \succ_u a'$, as desired. □

D A Reference to `DA-School`

In this section, we present `DA-School`, the well-known school-proposing deferred acceptance algorithm [Gale and Shapley, 1962]. In this setting, schools which are not at capacity *propose* to students one at a time, starting from their favorite students and moving down their preference list. When a student gets a proposal, if she is tentatively matched to some other school, she will reject the offer from whichever school she likes less and accept the offer from the school she likes better. At this point, she is tentatively matched to the school she likes better, and the other school will continue to make proposals to fill the seat offered to her. The version of the algorithm we present here is non-standard – it operates by having each school set an admissions threshold, which it decreases slowly – but is easily seen to be equivalent to the deferred acceptance algorithm. This version of the algorithm will be much more amenable to a private implementation, which we give next. When a school u lowers its threshold t_u below the score of a student a at school u ($\texttt{score}(u, a)$), we say that school u has *proposed* to student a.

It is well-known that `DA-School` will output a school-optimal stable matching (in our notation, a 0-approximate school-dominant stable matching) [Roth, 1985], assuming all players are truthful.

Algorithm 4: `DA-School`, the deferred acceptance algorithm with schools proposing

Input: school capacities $\{C_u\}$, student preferences $\{\succ_a\}$ and scores $\{\texttt{score}(u, a)\}$, range of scores $[0, J]$

Output: a set of score thresholds $\{t_j\}$

initialize: for each school u_j and each student a_i

$\text{counter}(u_j) = 0 \qquad t_j = J, \qquad \mu(a_i) = \emptyset$

while there is some under-enrolled school u_j: $\text{counter}(u_j) \leq \widehat{C}_{u_j}$ and $t_{u_j} > 0$ **do**

  $t_{u_j} = t_{u_j} - 1$

  for all student a_i **do**

    if $\mu(a_i) \neq \text{argmax}_{\succ_{a_i}}\{u_j \mid \texttt{score}(u_j, a_i) \geq t_{u_j}\}$ **then**

      $\text{counter}(\mu(a_i)) = \text{counter}(\mu(a_i)) - 1;$

      let $\mu(a_i) = \text{argmax}_{\succ_{a_i}}\{u_j \mid \texttt{score}(u_j, a_i) \geq t_{u_j}\}$

      $\text{counter}(\mu(a_i)) = \text{counter}(\mu(a_i)) + 1;$

return Final threshold scores $\{t_j\}$

E Private Matching Algorithms Must Allow Empty Seats

In this paper, we gave an algorithm with strong worst-case incentive properties in large markets, without needing to make distributional assumptions about the agents preferences, or requiring any other "large market" condition other than that the capacities of the schools be sufficiently large. However, in exchange, we had to relax our notion of stability to an approximate notion which allows a small number of empty seats per school. We here give an example demonstrating why this relaxation is necessary for any differentially private matching algorithm. An algorithm that must return an -exactly- stable matching must have extremely high sensitivity to the change in preferences of any single agent, if preferences are allowed to be worst case.

Example 1. *Suppose there are n students and 2 schools, H and Y. Suppose, for students $1 \leq a \leq \frac{n}{2}$, $H \succ_a Y$, and for $\frac{n}{2} < a \leq n$, $Y \succ_a H$. Each school has capacity for exactly half of the students: $C_H = C_Y = \frac{n}{2}$. Suppose Y has preference ordering $\succ_Y$, $s_1 \succ_Y s_2 \succ_Y \ldots \succ_y s_n$; H has preference ordering $s_{\frac{n}{2}+1} \succ_H s_{\frac{n}{2}+2} \succ_H \ldots s_n \succ_H s_1 \succ_H \ldots \succ_H s_{\frac{n}{2}}$. The school-optimal matching matches students $s_1, \ldots, s_{\frac{n}{2}}$ to Y and $s_{\frac{n}{2}+1}, \ldots, s_n$ to H. Now consider the market with any single student removed. The school-optimal stable matching changes entirely (i.e. every single student is matched to a different school). For example, if s_1 is removed, Y will admit $s_{\frac{n}{2}+1}$ (who will accept), H will admit s_2 (who will accept), Y will admit $s_{\frac{n}{2}+2}$ and so on. In the end, each student will get her favorite school, and the schools will swap students. The same effect is achieved by having a single student* change *her preferences, by reporting that she prefers to be unmatched than to be matched to her second choice school. This example shows that the exact school-optimal matching is* highly *sensitive to the addition, removal, or alteration of preferences of a single student and hence impossible to achieve under differential privacy. Our algorithms blunt this kind of sensitivity via the use of a small budget of seats that we may leave empty.*