Obvious Manipulations

Peter Troyan
Department of Economics
University of Virginia

Thayer Morrill[*]
Department of Economics
North Carolina State University

October 3, 2018

Abstract

A mechanism is *strategy-proof* if agents can never profitably manipulate, in any state of the world; however, not all non-strategy-proof mechanisms are equally easy to manipulate - some are more "obviously" manipulable than others. We propose a formal definition of an *obvious manipulation* and argue that it may be advantageous for designers to tolerate some manipulations, so long as they are non-obvious. By doing so, improvements can be achieved on other key dimensions, such as efficiency and fairness, without significantly compromising incentives. We classify common non-strategy-proof mechanisms as either *obviously manipulable (OM)* or *not obviously manipulable (NOM)*, and show that this distinction is both tractable and in-line with empirical realities regarding the success of manipulable mechanisms in practical market design settings.

1 Introduction

When designing mechanisms for allocating resources, such as in auctions, matching, or other assignment problems, a key concern is the incentives given for agents to report their private information truthfully. By selecting a mechanism that is *strategy-proof*, the designer can eliminate any possibility for an

[*] Authors' names appear in random order. Emails: troyan@virginia.edu and thayer_morrill@ncsu.edu.

agent to manipulate the mechanism for her own individual advantage by lying. However, good incentives are also very costly, and may inhibit other important desiderata, such as efficiency, stability, or fairness. At the same time, not all non-strategy-proof mechanisms give the same incentives to manipulate. Indeed, for some mechanisms, there is clear evidence that agents are adept at identifying opportunities for manipulation in practice, while for others, though such opportunities exist, they are observed much less frequently. This paper seeks a simple and tractable method for determining when a mechanism is easy to manipulate.

To motivate our project, consider two widely-used manipulable algorithms. The first is the Boston mechanism, which is one of the most popular mechanisms for school choice.[1] Under this mechanism, if a student has high priority at a school that is her true second choice, she may be better off by lying and ranking this school first. By doing so, she can guarantee being assigned to it, whereas if she told the truth, she risks losing it to others who ranked it higher, and may end up at her third (or worse) choice. Not only is the Boston mechanism manipulable in the formal sense of failing to be strategy-proof, but further, the relevant manipulations are also very easy to identify and enact. Indeed, this has been discovered and used by both parents and policymakers. For instance, Pathak and Sönmez (2008) report on a well-organized parent group in Boston advising their members as follows:

> One school choice strategy is to find a school you like that is undersubscribed and put it as a top choice, OR, find a school that you like that is popular and put it as a first choice and find a school that is less popular for a "safe" second choice.

Using data from magnet school assignment in Wake County, NC, which used a version of the Boston mechanism, Dur et al. (2018) present empirical evidence that many students do in fact act strategically in line with the above advice. Indeed, the ability of students to both identify and profit from strategic opportunities (at the expense of non-strategic students) has been a leading factor in the abandonment of the mechanism in some jurisdictions.[2]

[1] The Boston mechanism has been used by school districts in many US cities, including Boston, Chicago, Seattle, Minneapolis, Charlotte-Mecklenburg, Denver, Miami, and Tampa, among others.

[2] Though the use of the Boston mechanism has been abandoned in some places (including

On the other hand, consider the (doctor-proposing) Deferred Acceptance, or DA, mechanism, which is used every year by the National Resident Matching Program (NRMP) to assign thousands of newly-graduated doctors to residency training positions in hospitals across the US (Roth and Peranson, 1999), as well as around the world. While this mechanism is often lauded for being strategy-proof for the doctors, it is also well-known that it is not strategy-proof for the hospitals. However, while it is possible for hospitals to manipulate their preferences and obtain a better assignment in some states of the world, to do so successfully is difficult, and requires a detailed understanding of the mechanics of the mechanism and of the preferences of the other agents. Without such knowledge, it is very possible that attempting such a manipulation can actually "backfire": the manipulating hospital may be not be assigned a doctor it would actually be happy to employ. This is in stark contrast to the Boston mechanism, where a student is able to guarantee a spot at her second-choice school, and thereby surely avoid a potentially worse outcome from reporting truthfully.

As a final example, consider the competitive equilibrium from equal incomes mechanism (CEEI), which has recently been implemented for course assignment at Wharton business school. Eric Budish (the inventor of CEEI) and Judd Kessler say the following about trying to manipulate the mechanism:

> ...the kinds of profitable manipulations that were found in extensive computational exploration...are non-intuitive. Since there is a risk to misreporting – one is no longer guaranteed one's most-preferred affordable schedule at the realized prices – and the benefits of misreporting are difficult, if not impossible, to realize, we felt comfortable advising students to report truthfully. If either of the authors of this paper were participating in this market design, even in a small economy like the ones used in the laboratory, we would report truthfully. (Budish and Kessler, 2017)

These examples all highlight that some mechanisms provide opportunities for manipulation that are much easier for agents to recognize and execute successfully than others; in other words, some manipulations are more "obvious"

its namesake city and a total legislative ban in England), it still remains one of the most popular assignment mechanisms overall. Pathak and Sönmez (2013) provide a comprehensive list of school choice mechanisms used in various localities, both past and present.

than others. The main contribution of this paper is a formalization of the word "obvious", which we then use to classify non-strategy-proof mechanisms as either *obviously manipulable* or *not obviously manipulable*.

Behaviorally, our formal definition of an obvious manipulation is motivated by the influential paper of Li (2017) on *obvious strategy-proofness.* Li (2017) starts from the observation that real-world agents are often unable to engage in the intricate contingent reasoning to fully understand the implications of a given course of action on a state-by-state basis.[3] Instead, he presumes that they are only able do something much simpler, which is to determine the set of possible outcomes, $\pi(s)$, from any given strategy s (but, crucially, not how the realized outcome depends contingently on the actions of others). He then looks for a strategy s such that *every* outcome in $\pi(s)$ is (weakly) better than *every* outcome in $\pi(s')$, for all s'. While a robust solution concept, a drawback is that $\pi(s)$ and $\pi(s')$ will usually be incomparable according to this strong criterion, and so very few mechanisms will be obviously strategy-proof.[4,5] Indeed, as noted above, even strategy-proofness itself can be very limiting, and so our approach is to relax this restriction. We do so by considering agents who compare the worst (best) outcomes in $\pi(s)$ to the worst (best) outcomes in $\pi(s')$. We then define a manipulation s' as *obvious* if either $\min \pi(s')$ is strictly preferred to $\min \pi(s)$ or $\max \pi(s')$ is strictly preferred to $\max \pi(s)$.[6]

The motivation for this definition is three-fold. First, the min and max are typically very salient outcomes, and are thus likely to be very important in decision-making. Second, it is often much less complicated to calculate the min and the max than to compare contingent outcomes under all possible scenarios, consistent with the motivation of Li (2017), and thus it is natural for people to use other, simpler criteria when evaluating outcomes from various

[3] Indeed, there is increasing evidence that many people have difficulties with hypothetical reasoning even in single-agent decision problems (Charness and Levin, 2009; Esponda and Vespa, 2014), let alone environments with strategic interactions among many agents.

[4] Also, obvious strategy-proofness is defined on the extensive form of a game, as almost no normal-form games will be obviously strategy-proof. Many real-world applications like those we are concerned with (school choice, NRMP) have tens or even hundreds of thousands of agents, making it very impractical to run an extensive-form mechanism.

[5] Ashlagi and Gonczarowski (2018), Troyan (2016), Pycia and Troyan (2016), Arribillaga et al. (2017), and Bade and Gonczarowski (2016) fully characterize obviously strategy-proof mechanisms in various environments, including matching, voting, and auctions, among others.

[6] Additionally, any s' that weakly dominates s is also classified as an obvious manipulation. See Section 2 for a formal definition.

courses of action. And third, there are many more non-obviously manipulable mechanisms than there are strategy-proof mechanisms, which means that a designer will be able to implement a far greater range of allocations. This will allow for improvements on dimensions such as efficiency or fairness, while still ensuring that there are no obvious manipulations.

After introducing the definition formally in Section 2, we proceed to apply it to various environments, starting with school choice in Section 3. We first formalize the above discussion regarding the Boston mechanism and show it is indeed obviously manipulable (Proposition 1). This result is particularly relevant because, while some school districts have moved away from the Boston mechanism, it still remains one of the most popular mechanisms in practice, despite being so obviously susceptible to manipulations.

While some districts have moved away from Boston and towards DA (often on the advice of economists), DA itself suffers from a shortcoming that it may produce a Pareto inefficient assignment. To correct this, many have proposed new mechanisms that Pareto improve on DA, but all of these mechanisms give up DA's strategy-proofness. While it is known that any mechanism that Pareto dominates deferred acceptance is manipulable (Abdulkadiroğlu et al., 2009; Kesten, 2010; Alva and Manjunath, 2017), we show a striking result: while they may be manipulable, any mechanism that Pareto dominates DA is not *obviously* manipulable (Theorem 1). This has particularly important implications for the efficiency-adjusted deferred acceptance (EADA) mechanism of Kesten (2010). EADA has received renewed attention recently, as several papers have shown that EADA is the unique Pareto efficient mechanism that also satisfies natural fairness axioms (Dur et al., 2015; Ehlers and Morrill, 2017; Tang and Zhang, 2017; Troyan et al., 2018). The only shortcoming of the EADA assignment is its implementation, given the fact that it is a manipulable mechanism. However, Theorem 1 implies that EADA is not obviously manipulable, and thus this is less likely to be an issue in practice.

While several of our results are stated specifically in a school choice context, the definition of (non)-obvious manipulability can be applied much more broadly, and in Section 4, we explore several of these applications as well. In the context of two-sided matching, we show that while DA is manipulable for the receiving side (e.g., hospitals can manipulate doctor-proposing DA), it is not obviously so (Theorem 2). In the context of multi-unit auctions, we show that pay-as-bid multi-unit auctions (a generalization of the first-price auction) are obviously manipulable (Corollary 2), while the $(K+1)$-price auction is not (Theorem 4).[7] Last, we consider the classic bilateral trade setting with one buyer and one seller. We first show directly that double auctions (Chatterjee and Samuelson, 1983) are obviously manipulable. We then ask whether there is any NOM mechanism that also satisfies other common desirable axioms. Our final main result is an impossibility result in the spirit of Myerson and Satterthwaite (1983): every efficient, individually rational and weakly budget balanced mechanism is obviously manipulable (Theorem 5).

We highlight that in our model, agents have standard preferences over outcomes, and we make no assumptions about prior probability distributions over the types or reports of other agents; rather, we presume that the ability of agents to recognize certain deviations as profitable may vary across mechanisms. Thus, our approach is consistent with the Wilson doctrine (Wilson, 1987), in the sense that determining whether a mechanism is obviously manipulable requires no assumptions about common knowledge or an agents' prior beliefs. For instance, in the bilateral trade setting, it is difficult for the buyer to determine her optimal bid in a double auction mechanism, because it is highly sensitive to his beliefs about the seller's ask (and vice-versa). Our definition captures this difficultly by classifying this mechanism as obviously manipulable.[8]

The most common alternative approach to relaxing strategy-proofness in market design (without moving all of the way to Bayesian incentive compatibility) is to consider large markets. For instance, Immorlica and Mahdian (2005) and Kojima and Pathak (2009) show that the incentives to manipulate DA vanish as the size of the market approaches infinity.[9] Azevedo and Budish (2018) define a related but more general concept of strategy-proofness in the large (SPL) that applies to a wide variety of settings. While similar in motiva-

[7] K denotes the number of identical units to be sold. While strategy-proofness holds for $K = 1$ (a second-price auction), the $(K + 1)$-price auction is manipulable for $K > 1$.

[8] Our results thus provide a contrast to the recent literature on mechanism design with maximin expected utility agents (MEU, Gilboa and Schmeidler, 1989), which also has agents comparing worst-case outcomes under any two reports. For instance, De Castro and Yannelis (2018) (see also Wolitzky, 2016) show how ambiguity can be used to "solve" the impossibility of Myerson and Satterthwaite (1983), whereas our Theorem 5 reinforces Myerson and Satterthwaite's negative result. This is discussed further in Section 4.

[9] Regarding DA in particular, also related are Barberà and Dutta (1995) and Fernandez (2018), who define particular classes of strategies (protective strategies and regret-free truth-telling, respectively), and use them to rationalize truthful reporting under DA.

tion, our approach is distinct in several respects. Most notably, we require no assumptions on how preferences are drawn or agent beliefs, nor do we require the market size to approach infinity. Another recent strand of literature tries to quantify a mechanism's manipulability using particular metrics. This includes Carroll (2011), who defines a mechanism's susceptibility to manipulation as the maximum cardinal utility any agent can gain from lying, and Pathak and Sönmez (2013), who use a profile-counting metric to define one mechanism as "more" manipulable than another if, for any preference profile where the latter is manipulable, the former is as well. We do not attempt to rank mechanisms by their degree of manipulability, but instead want to eliminate all obvious manipulations.

In summary, we believe that imposing only non-obvious manipulability can be a useful design objective in many settings, as it will allow improvements on other important dimensions such as fairness or efficiency, while eliminating the most clear opportunities for manipulation. Obvious manipulations are easier for cognitively limited agents to recognize than others, and are less risky (in the sense of downside risk) than telling the truth. Further, from a pragmatic standpoint, our classification is tractable and is inline with empirical realities with regard to successful practical market design across a range of applications. This suggests that not only is obvious manipulability capturing an important feature of incentives in existing mechanisms, but can also be applied when considering implementing new mechanisms that have not yet been used in practice.

2 Definitions

We consider an environment with a finite set of N agents, denoted $I = \{i_1, \ldots, i_N\}$, and a set of outcomes, denoted X. The set Θ_i denotes the possible **types** for agent i, with generic element $\theta_i \in \Theta_i$. The function $u_i(x; \theta_i)$ denotes agent i's utility for outcome x when his type is θ_i. Let $\Theta_I = \times_{i \in I} \Theta_i$. A (direct) **mechanism** is a function $\phi : \Theta_I \to X$ that maps type profiles to outcomes. When convenient, in some applications we will use the notation $\phi_i(\theta)$ to denote i's individual allocation (e.g., in school choice, $\phi(\theta)$ is the entire assignment of all students to schools when the type profile is θ, while $\phi_i(\theta)$ is i's school and $u_i(\phi_i(\theta); \theta_i)$ is i's utility for school $\phi_i(\theta)$ when of type θ_i).

An important concern when choosing a mechanism is the incentives given to the agents to report their preferences truthfully. The standard definition of strategy-proofness requires that truthful reporting be a weakly dominant strategy in the preference revelation game induced by the mechanism; formally, mechanism ϕ is **strategy-proof** if $u_i(\phi(\theta_i, \theta_{-i}); \theta_i) \geq u_i(\phi(\theta'_i, \theta_{-i}); \theta_i)$ for all i, all $\theta_i, \theta'_i \in \Theta_i$, and all $\theta_{-i} \in \Theta_{-i}$. While desirable as an incentive property, strategy-proofness is also a demanding condition, and may restrict a mechanism designer's ability to achieve other desirable goals. Indeed, many practical market design settings use non-strategy-proof, or manipulable, mechanisms (see the Introduction). It is these mechanisms that will be the focus of our paper.

Definition 1. Report θ'_i is a (profitable) **manipulation** of mechanism ϕ for agent i of type θ_i if there exists some $\theta_{-i} \in \Theta_{-i}$ such that $u_i(\phi(\theta'_i, \theta_{-i}); \theta_i) > u_i(\phi(\theta_i, \theta_{-i}); \theta_i)$. If some type of some agent i has a profitable manipulation θ'_i, then we say that mechanism ϕ is **manipulable**.

Note that for a mechanism to be classified as manipulable, there must simply exist some profile of the other agents, θ_{-i}, such that when they report θ_{-i}, agent i prefers to report θ'_i over the truth θ_i. However, in other instances, reporting θ'_i may actually be worse for agent i than reporting truthfully. Thus, to any agent who must report her own type before she knows the types of others, it may be very unclear whether such a manipulation will be profitable in practice. The standard way of dealing with this in game theory is to assume that there is a common prior distribution over Θ from which each agent's type is drawn. Each agent then calculates their own outcome for each possible action profile and is assumed to choose the action that maximizes her own expected utility.

In general, this can be a difficult process. First, as articulated by Wilson (1987), it is unclear where such a common prior comes from in the first place, which opens up the possibility for significant mis-coordination. A second, and even more basic, issue is that simply determining even just the possibilities from a potential report θ'_i may be infeasible for many real-world agents, as doing so requires extensive contingent reasoning on various hypothetical scenarios. In defining obvious dominance, for example, Li (2017)'s motivation is an agent who "knows all the possible outcomes that might result from [a particular] strategy...[but] does not know the possible outcomes *contingent on some unobserved event*" (emphasis as in the original), and then looks for mechanisms where all possibilites from one strategy are (weakly) better than all possibilities from any other. At the same time, calculating the worst possible (or best possible) outcome is typically much simpler than calculating all possible outcomes, and even if it is possible to do the latter, it is still unclear how to compare the resulting sets of possibilities (at least without making assumptions on prior distributions and beliefs). Motivated by these observations, and by the examples given in the introduction, we look for a weakening of strategy-proofness that does not rely on the assumption of a common prior, but rather focuses on simpler best/worst case analyses.

Definition 2. Mechanism $\phi(\cdot)$ is **not obviously manipulable (NOM)** if, for any profitable manipulation θ_i', the following are true:

(i) $\min_{\theta_{-i}} u_i(\phi(\theta_i', \theta_{-i}); \theta_i) \leq \min_{\theta_{-i}} u_i(\phi(\theta_i, \theta_{-i}); \theta_i)$

(ii) $\max_{\theta_{-i}} u_i(\phi(\theta_i', \theta_{-i}); \theta_i) \leq \max_{\theta_{-i}} u_i(\phi(\theta_i, \theta_{-i}); \theta_i)$

(iii) there exists some θ_{-i} such that $u_i(\phi(\theta_i', \theta_{-i}); \theta_i) < u_i(\phi(\theta_i, \theta_{-i}); \theta_i)$.

If any of (i), (ii), or (iii) do not hold for some manipulation θ_i', then θ_i' is said to be an **obvious manipulation** for agent i of type θ_i, and mechanism ϕ is said to be **obviously manipulable (OM)**.

Intuitively, a manipulation θ_i' is classified as "obvious" if it either makes the agent strictly better off in the worst case (i.e., $\min_{\theta_{-i}} u_i(\phi(\theta_i', \theta_{-i}); \theta_i) > \min_{\theta_{-i}} u_i(\phi(\theta_i, \theta_{-i}); \theta_i)$) or it makes the agent strictly better off in the best case (i.e., $\max_{\theta_{-i}} u_i(\phi(\theta_i', \theta_{-i}); \theta_i) > \max_{\theta_{-i}} u_i(\phi(\theta_i, \theta_{-i}); \theta_i)$). Part (iii) of Definition 2 allows us to compare mechanisms for which the best and worst cases from manipulating and truth-telling are equivalent (e.g., some auctions).[10] If any of (i)-(iii) are violated for a manipulation θ_i', then we say θ_i' is a **non-obvious manipulation**. In other words, a manipulation is non-obvious if the best and worst case outcomes from truth-telling are always weakly better, and in the case of equality, the agent can construct at least one situation in which truth-telling is strictly better.

The rest of the paper will consist of applying our definition to several canonical environments and classifying well-known manipulable mechanisms as either obviously manipulable (OM) or not obviously manipulable (NOM).

[10] Equivalently, part (iii) says that truth-telling is not a weakly dominated strategy.

However, before moving on to such applications in the next section, we provide some remarks on the interpretation of (non-)obviously manipulability.

As explained in the introduction, our motivation for this definition is three-fold. First, the min and max are typically very salient outcomes to an agent. Second, it is often much less complicated to calculate the min and the max than to compare outcomes under all possible scenarios, and so one can interpret our results as contributions to a renewed interest in "simplicity" in mechanism design in the spirit of Li (2017). While he considers a mechanism to be simple if is easy for a cognitively limited agent to identify a course of action that is better than all others (in the sense of being obviously dominant), NOM mechanisms are simple in a different sense: in an NOM mechanism, it is hard for a cognitively limited agent to identify an action that is better than some given action. Thus, the idea applies most naturally to things like direct mechanisms, where there is a natural reference point (i.e., truth-telling).[11] And third, Definition 2 admits many more mechanisms as not obviously manipulable relative to classifications based on other criteria such as (obvious) strategy-proofness, thereby expanding the designer's toolbox.

Also, note that while we will see that many natural and commonly-used manipulable mechanisms are NOM, this does not mean that we necessarily expect 100% truth-telling in such a mechanism; similarly, just because a mechanism is OM does not necessarily imply that everyone will try to manipulate it. Rather, our definition is intended to capture the idea that if a mechanism is obviously manipulable, it is easy for agents to recognize how they *could* manipulate the mechanism; whether or not they *should* try to manipulate is a different question and will depend on their subjective probabilities of how likely it is that such a manipulation will be successful and their preference intensities for various outcomes. For instance, in the Boston mechanism, most participants are easily able to recognize the manipulation of inflating the ranking of a school at which they have a high priority; however, whether they want to submit such a report depends on their expectation of demand for that school and how strongly they prefer it to alternatives. The point is that a first step in submitting a manipulation is to be *aware* of it as potentially profitable. By choosing a mechanism that is NOM, the designer can rule out the most obvious types of manipulations, and in so doing at least mitigate the likelihood that agents

[11] However, we do partially generalize Definition 2 beyond direct mechanisms in Section 4.

will be able to identify and execute manipulations in practice.

3 School choice

We begin by applying the ideas of (non)-obvious manipulability to school choice. We consider a canonical model of school choice, as in the seminal paper of Abdulkadiroğlu and Sönmez (2003). Let S be a set of schools. Each school has a capacity q_s and a strict priority ranking $\succ_s$ over I. A **matching** is a function $\mu : I \cup S \to I \cup S \cup \{\emptyset\}$ such that (i) $\mu_i \in S \cup \{\emptyset\}$ for all $i \in I$ (ii) $\mu_s \subset I$ and $|\mu_s| \leq q_s$ for all $s \in S$ and (iii) $\mu_i = s$ if and only if $i \in \mu_s$. If $\mu_i = \emptyset$, then a student remains unmatched (which can also be interpreted as taking some outside option).

In the notation of the previous section, X would be the set of all matchings and θ_i would parameterize each agent's utility function over matchings. However, in school choice models, students only care about their own assignment, and usually only ordinal preferences are considered. It is thus standard to identify an agent's type θ_i with her (strict) ordinal preference ranking over $S \cup \{\emptyset\}$, denoted P_i, and work with this ordinal preference relation directly, rather than a utility function. To be consistent with this literature, in this section, we follow this notation and write $a \; P_i \; b$ to denote that school $a \in S$ is strictly preferred to $b \in S$ by student i. Any s such that $\emptyset \; P_i \; s$ is said to be an **unacceptable** school for student i. Also, we let R_i denote the corresponding weak preference relation,[12] and write $P = (P_i)_{i \in I}$ to denote a profile of preference relations, one for each student. The schools are not strategic agents, but rather are simply objects to be consumed, and the school priorities are publicly known to all of the students.

We use $\phi(P)$ to denote the matching prodcued by mechanism ϕ at preference profile P, and write $\phi_i(P)$ for i's assigned school at matching $\phi(P)$. Given a mechanism ϕ, let

$$W_i^{\phi}(P_i') = \min_{P_{-i}} \phi_i(P_i', P_{-i}),$$

where the minimum is understood to be taken with respect to the true preferences P_i. In other words, $W_i^{\phi}(P_i')$ is the worst possible school for i in mech-

[12] That is, aR_ib if either aP_ib or $a = b$.

anism ϕ when she has true preferences P_i and reports preference P_i'. It is of course possible to set $P_i' = P_i$ and determine the worst-case outcome when i reports her true preferences. We define the best possible outcome analogously:

$$B_i^\phi(P_i') = \max_{P_{-i}} \phi_i(P_i', P_{-i}),$$

Using this notation, a manipulation P_i' is an obvious manipulation of mechanism ϕ (in the sense of Definition 2) if (i) $W_i^\phi(P_i')\ P_i\ W_i^\phi(P_i)$ or (ii) $B_i^\phi(P_i')\ P_i\ B_i^\phi(P_i)$ or (iii) $\phi_i(P_i', P_{-i})\ R_i\ \phi_i(P)$ for all P_{-i}. If none of these hold for any P_i', then ϕ is not obviously manipulable.

We illustrate this definition with two mechanisms that are well-known to be manipulable: the Boston mechanism and the school-proposing Deferred Acceptance algorithm.[13] Since neither mechanism is strategy-proof, there are situations for each mechanism where a student may benefit from strategizing. However, the types of manipulations are very different for the two mechanisms: the manipulations in the Boston mechanism are obvious, while those for school-proposing DA are not.

Example 1 (Boston Mechanism)**.** *Suppose there are three students, $I = \{i, j, k\}$ and three schools $S = \{a, b, c\}$. Each school has a capacity $q_s = 1$ for all $s \in S$. The preferences of the students and the priorities are as follows:*

P_i	P_j	P_k
a	b	a
b	$\vdots$	$\vdots$
c		

$\succ_a$	$\succ_b$	$\succ_c$
k	i	$\vdots$
i	j	
j	k	

Let $\phi = BM$ denote the Boston mechanism, and $BM_i(P)$ be student i's assigned school under preference profile P. If all students report their true preferences (those in the table), then $BM_i(P) = c$. However, if i reports $P_i' : b, a, c$, then $BM_i(P_i', P_{-i}) = b$, which she strictly prefers to c. Thus, P_i' is a profitable manipulation, and the Boston mechanism is manipulable. Further, note that if i reports P_i', then she is guaranteed to receive b for sure, no matter what the other students report, and so this is the worst case from reporting P_i': $W_i^{BM}(P_i') = b$. It is clear that the worst case from the truth is $W_i^{BM}(P_i) = c$, and so $W_i^{BM}(P_i') P_i W_i^{BM}(P_i)$. Therefore, P_i' is an obvious manipulation.

[13] Formal definitions of these and other standard school assignment mechanisms can be found in Appendix A.

Example 1 can easily be generalized to markets of any size, and so we have the following result.

Proposition 1. *The Boston mechanism is obviously manipulable.*

One easily recognized shortcoming of a "naive" implementation of the Boston mechanism is that in some rounds, students may end up applying to a school in round k even if it was filled to capacity in some round $k' < k$, thereby "wasting" their round k application. Several recent papers have considered a simple and intuitive modification of the Boston mechanism that adapts the students' preferences to prevent them from applying to a school in a given round if there is no capacity remaining. Dur (2018) shows that in every problem where this Modified Boston Mechanism (also sometimes referred to as the Adaptive Boston Mechanism) can be manipulated, the original Boston Mechanism can also be manipulated but that the converse is not true, and so the Modified Boston Mechanism is "less manipulable" than the original Boston Mechanism in the formal sense introduced by Pathak and Sönmez (2013).[14] Note that Example 1 is the same for the Boston or the Modified Boston mechanism, and therefore, although the Modified Boston Mechanism is less manipulable than the Boston Mechanism in a profile-counting sense, it is still obviously manipulable.

Next, we turn to the school-proposing DA mechanism. School-proposing DA is also a manipulable mechanism, but the form of the manipulations are much different from those of the Boston mechanism. This is highlighted by the following example.

Example 2 (School-Proposing Deferred Acceptance)**.** *We let $\phi = schDA$ denote the school-proposing DA algorithm. Suppose there are 3 students $I = \{i, j, k\}$ and three schools $S = \{a, b, c\}$. Each school has a capacity $q_s = 1$ for all $s \in S$. The preferences and priorities are as follows:*

P_i	P_j	P_k
a	b	c
b	c	a
c	a	b

$\succ_a$	$\succ_b$	$\succ_c$
j	k	i
k	i	j
i	j	k

[14] Various aspects of this mechanism are also considered by Miralles (2009), Mennle and Seuken (2014), and Harless (2016).

If all students report their true preferences (those in the table), then $schDA_i(P) = c$. If i reports $P'_i : a, \emptyset$, then $schDA_i(P'_i, P_{-i}) = a$, which she strictly prefers to c, and so P'_i is a profitable manipulation. However, reporting P'_i exposes i to worse outcomes than reporting her true preferences does. If i submits P_i, then c is her worst possible assignment, while if i submits P'_i and j ranks a first, then i will be unassigned, i.e.,

$$\min_{P_{-i}} schDA_i(P_i) = c \; P_i \; \emptyset = \min_{P_{-i}} schDA_i(P'_i).$$

Therefore, although P'_i is a profitable manipulation for i, it is not an obvious manipulation. (While this is only one example of a manipulation that is non-obvious, Theorem 1 below will imply that schDA is not obviously manipulable in general.)

Examples 1 and 2 provide an illustration of the different types of manipulations that we will distinguish. Under the Boston mechanism, when a student ranks her 'neighborhood school' first,[15] she is guaranteed to be assigned to it. It is very salient to students who participate in this mechanism that such a manipulation may be beneficial. On the other hand, to identify the truncation strategy in Example 2 as a manipulation is much more involved. It is far more difficult to identify the precise states when such a deviation will be profitable, yet seems intuitively obvious that listing a truly acceptable school as unacceptable may result in a worse possible outcome than if the agent were to submit her true preferences.

The truncation strategy in Example 2 is just one possible deviation, but we show that this intuition holds more broadly: no profitable manipulation of schDA is an obvious manipulation. In fact, we show this not only for schDA, but for a much larger class of mechanisms. To introduce this class, first define a matching μ as **stable** if there do not exist any **blocking pairs**, which are any (i, s) such that $sP_i\mu_i$ and either (i) $|\mu_s| < q_s$ or (ii) there exists some $j \in \mu_s$ such that $i \succ_s j$. Further, say that matching μ **Pareto dominates** matching μ' if $\mu_i R_i \mu'_i$ for all $i \in I$. Then, following Alva and Manjunath (2017), we define a matching μ to be **stable-dominating** if it

[15] More generally, if a student ranks first a school s where she has one of the q_s highest priorities. This is sometimes called a neighborhood school in the literature for convenience, though priorities need not be determined geographically in general.

is stable or Pareto dominates some stable assignment.[16] A mechanism ϕ is said to be stable if $\phi(P)$ is stable for all preferences profiles P; similarly, ϕ is a stable-dominating mechanism if $\phi(P)$ is a stable-dominating assignment for all P. Alva and Manjunath (2017) have shown that the only strategy-proof and stable-dominating mechanism is the (student-proposing) Deferred Acceptance mechanism (this is not to be confused with the school-proposing DA mechanism from Example 2; in this paper, unless otherwise specified, DA will always refer to the student-proposing version, while we use schDA to refer to the school-proposing versions). However, there are many other manipulable, stable-dominating mechanisms, including:[17]

- the school-proposing Deferred Acceptance mechanism (Gale and Shapley, 1962)
- the Efficiency Adjusted Deferred Acceptance mechanism (Kesten, 2010)
- Deferred Acceptance with Compensation Chains (Dworczak, 2016)
- the Deferred Acceptance plus Top Trading Cycles mechanism (Alcalde and Romero-Medina, 2017).

Two important properties of DA are that it is strategy-proof and fair; the main drawback is that it is not efficient. Because it is not efficient, it is important to understand whether we can improve upon DA from a welfare perspective. Abdulkadiroğlu et al. (2009) and Kesten (2010) show that any mechanism that Pareto dominates DA is necessarily manipulable (a result further extended by Alva and Manjunath (2017), discussed above). However, as our main result shows, while such mechanisms may be manipulable, they are not obviously manipulable.

[16] In one-sided matching problems such as school choice, where one side of the market (e.g., the schools) is viewed as objects to be consumed, rather than actual agents, stability is often interpreted as an important fairness criterion (see, e.g., Balinski and Sönmez (1999) and Abdulkadiroğlu and Sönmez (2003)). For expositional purposes and consistency with Alva and Manjunath (2017), we stick to the word stability. Additionally, in the next section we will discuss some two-sided matching applications where stability is given a positive interpretation.

[17] The Stable Improvement Cycles mechanism introduced in Erdil and Ergin (2008) is also manipulable and stable-dominating. However, our formal model does not have indifferences in priorities. In our setting, their algorithm is equivalent to DA.

Theorem 1. *Any stable-dominating, IIR mechanism is not obviously manipulable.*[18]

We prove Theorem 1 using a series of lemmas that may be of independent interest in their own right. Lemmas 3 and 4 focus on two particular classes of reports that have garnered much attention in the literature as focal classes of manipulations, and show no such report is an obvious manipulation under a stable-dominating, IIR mechanism. These results themselves, as well as Theorem 1, rely crucially on Lemmas 1 and 2, which we prove first. These lemmas provide a tight characterization of the worst possible assignment under a stable-dominating mechanism.

Given a mechanism ϕ, we define a school s to be a **safety school** for a student i with preferences P_i if, for every P_{-i}, we have $\phi_i(P)\ R_i\ s$. By definition, a student's worst possible assignment will be her favorite safety school. We call a school s an **aspirational school** if there exists a profile P_{-i} such that $s\ P_i\ \phi_i(P)$ (i.e., if s is not a safety school). We first note that all stable-dominating mechanisms have the same worst-case assignment.

Lemma 1. *If ϕ and ψ are both stable-dominating mechanisms, then $W_i^{\phi}(P_i) = W_i^{\psi}(P_i)$ for all i and all P_i.*

Proof. Our proof strategy will be to first find the worst-case outcome under a particular stable mechanism, namely, school-proposing DA. We label this school $\bar{w}$. Then, we will show that if ϕ is a stable-dominating mechanism, the worst-case under ϕ is also $\bar{w}$. Since ϕ is an arbitrary stable-dominating mechanism, this will establish the result.

Formally, for a student i with preferences P_i, define:

$$\bar{w} = \max_{P_i} \{s : \text{for every } P_{-i},\ schDA_i(P)\ R_i\ s\}. \tag{1}$$

Note that $\bar{w}$ is a safety school under schDA, and in fact, is i's most-preferred safety school. Therefore, $\bar{w}$ is a lower bound on i's worst possible assignment

[18] IIR is a technical condition to rule out degenerate mechanisms. For every problem P, we call a mechanism **independent of irrelevant rankings (IIR)** if for every profile P_i' that is identical to P_i up to $\phi_i(P)$, $\phi(P) = \phi(P_i, P_{-i})$. In words, if a student had changed her ranking of schools below the school she got into, this would not affect her or any other student's assignment. Every real world mechanism that we know of satisfies IIR. An example of a non-IIR mechanism would be a sequential dictatorship where the next dictator is determined by how the current dictator ranks objects below the one she chooses.

under schDA. To establish that $\bar{w}$ is, in fact, the worst possible assignment, we just need to find one profile P_{-i} such that $schDA_i(P) = \bar{w}$. This is trivial if $\bar{w}$ is i's favorite school.[19] Otherwise, let s be the school i ranks just above $\bar{w}$. Since s is not a safety school, there exists a P_{-i} such that $s \; P_i \; schDA_i(P)$. However, since $\bar{w}$ is a safety school for schDA, $schDA_i(P) \; R_i \; \bar{w}$. Therefore, $schDA_i(P) = \bar{w}$ (since s was chosen so that there is no s' such that $s \; P_i \; s' \; P_i \; \bar{w}$). This establishes that $\bar{w}$ is the worst possible assignment under schDA.

Now, define a matching $\lambda = schDA(P)$. Note that since ϕ is stable-dominating, for any P'_{-i}, $\phi_i(P_i, P'_{-i}) \; R_i \; schDA_i(P_i, P'_{-i}) \; R_i \; \bar{w}$;[20] therefore, $\bar{w}$ is a lower bound for i under ϕ. If we can find one profile P'_{-i} such that $\phi_i(P_i, P'_{-i}) = \bar{w}$, this will establish that $\bar{w}$ is in fact the worst possible assignment for ϕ. If λ is not a Pareto efficient matching, then for each $j \neq i$, define $P'_j := \lambda_j, \emptyset$ (where $\lambda_j = schDA_j(P)$ and it is understood that $\emptyset, \emptyset$ is replaced by $\emptyset$). It is straightforward to verify that $schDA_i(P_i, P'_{-i}) = \bar{w}$ and $schDA_i(P_i, P'_{-i})$ is Pareto efficient. Therefore, $\phi_i(P_i, P'_{-i}) = schDA_i(P_i, P'_{-i}) = \bar{w}$.

□

For a stable-dominating mechanism, the aspirational schools are determined by Hall's Theorem (Hall, 1935), which gives a necessary and sufficient condition for finding a matching that covers a bipartite graph. Intuitively, consider a student i whose favorite school is a. She is only guaranteed a if she has one of the q_a highest priorities; otherwise, if these students all rank a first, she will receive a worse assignment (under a stable or stable-dominating assignment). Suppose b is i's second favorite school. The key observation is that i may be guaranteed to be assigned to a or b, even if she does not have one of the q_a highest priorities at a nor one of the q_b highest priorities at b. This occurs when there are sufficiently many students ranked higher than her at both a and b (as these students can only be assigned to one school).

Lemma 2. *Let ϕ be a stable-dominating mechanism and for each school s', let $B_i(s') = \{j \in I : j \succ_{s'} i\}$. Consider a student i with preferences P_i. School s is a safety school for student i if and only if there exists a set of schools $S' \subseteq S$*

[19] In fact, in this case, for every P_{-i}, $schDA_i(P) = \bar{w}$.

[20] It is well known that schDA produces the student-pessimal stable assignment. That is to say all students weakly prefer any alternative stable assignment to the schDA assignment. See Roth and Sotomayor (1990) for a complete discussion.

such that $s' \; R_i \; s$ for all $s' \in S'$ and

$$\sum_{s' \in S'} q_{s'} > | \cup_{s' \in S'} B_i(s')|. \tag{2}$$

Proof. We first show the if direction. Fix a school s, and suppose there exists a set $S' \subseteq S$ such that for each $s' \in S'$, $s' \; R_i \; s$ and Equation 2 holds. Fix a profile P_{-i}, and let $\mu = schDA(P)$; specifically, μ_i is i's worst possible stable assignment. Suppose for contradiction that $sP_i\mu_i$. Note that each school $s' \in S'$ is assigned to it's capacity (or else μ is not stable). Therefore, by Equation 2, there must exist a school $s' \in S'$ and a student $j \notin B_i(s')$ such that $\mu_j = s'$. But $i \succ_{s'} j$ (by the definition of $B_i(s')$) and $s' \; P_i \; \mu_i$; therefore, i and s' block μ, contradicting the stability of μ. Therefore, $\mu_i \; R_i \; s$. Since ϕ is stable-dominating, $\phi_i(P) \; R_i \; \mu_i$. Therefore, $\phi_i(P) \; R_i \; s$. The same argument can be made for any profile P_{-i}, and so s is a safety school.

For the other direction, fix a school s, and assume that for every $S' \subseteq S$ such that $s' \; R_i \; s$ for all $s \in S'$, Equation 2 fails, i.e., for all such S', the following is true:

$$\sum_{s' \in S'} q_{s'} \leq | \cup_{s' \in S'} B_i(s')|. \tag{3}$$

In words, for every collection of schools weakly preferred to s, there are more students ranked higher at one of these schools than the total capacity of all of these schools. We will show that if Equation 3 holds for any possible set of schools i weakly prefers to s, then we can fill all of the seats at the preferred schools with students ranked higher than i. When these students rank their respective assignments first, it is not possible for i to be placed in a school weakly preferred to s in any stable assignment (or any Pareto improvement of one).

The result is an application of Hall's Theorem. Let $U = \{s' : s' \; R_i \; s\}$. We define a bipartite graph as follows. For each $s' \in U$ create $q_{s'}$ vertices $\left\{v^1_{s'}, \ldots, v^{q_{s'}}_{s'}\right\}$ and define X to be the set of these vertices. Create a vertex for each student, and label the set of all such vertices Y. We create a graph by drawing an edge between student j and vertex $v^k_{s'}$ (the k^{th} copy of school s') if and only if $j \succ_{s'} i$. In this graph, the *neighborhood* of any vertex v, denoted $N(v)$, is the set of vertices it shares an edge with. For a set of vertices $W \subseteq X$, $N(W)$ is defined as $\cup_{w \in W} N(w)$. Note that by definition, there are no edges between student i and any $v \in X$, and so, $N(i) = \{\emptyset\}$. Hall's Theorem says the following:

Theorem (Hall 1935). *If $|W| \leq |N(W)|$ for every subset $W \subseteq X$, then there exists a matching that entirely covers X.*

We will show that in the graph we have constructed, the conditions for Hall's Theorem are satisfied. Take some $W \subseteq X$. Let T be the schools that have at least one copy in W. Note that for every $t \in T$, $tR_i s$. Therefore, Equation 3 applies, i.e.,

$$\sum_{t \in T} q_t \leq | \cup_{t \in T} B_i(t)|. \tag{4}$$

By construction, $N(W) = \{j : \exists t \in T \text{ s.t. } j \succ_t i\}$. Written differently,

$$N(W) = \cup_{t \in T} B_i(t). \tag{5}$$

For each school $t \in T$, there are at most q_t copies of t in W, so $|W| \leq \sum_{t \in T} q_t$. This implies

$$|W| \leq \sum_{t \in T} q_t \leq | \cup_{t \in T} B_i(t)| = |N(W)|,$$

where the second inequality follows from Equation 4 and the last inequality follows from Equation 5. Therefore, by Hall's Theorem, for each school that i weakly prefers to s, we can assign every copy of that school to a student ranked higher than her. Given this vertex cover, we induce a matching λ, defined as follows: if student j was assigned to a copy of school s', then we set $\lambda_j = s'$; if student j was not matched, we set $\lambda_j = \emptyset$. We then define a preference profile P_{-i} such that, for every $j \neq i$, we set $P_j := \lambda_j, \emptyset$ (where it is understood that $\emptyset, \emptyset$ is replaced by $\emptyset$). It should be clear from our construction that under P, there is only one stable assignment: each student $j \neq i$ is assigned to λ_j, while i is assigned to the school she ranks just below s. It is also clear that this assignment is Pareto efficient; therefore, any stable-dominating mechanism must make the same assignment. In particular, $s \; P_i \; \phi_i(P)$, and consequently, s is not a safety school for i, which is a contradiction.

□

The following corollary is immediate from the proof of Lemma 2 and will be helpful in the proof of the main theorem.

Corollary 1. *Let ϕ be a stable-dominating mechanism, and consider a student i with preferences P_i. If s is an aspirational school, then there exists a preference profile P_{-i} such that $\phi_i(P) = s$.*

Recall our main goal is to show that stable-dominating mechanisms have no obvious manipulations. However, there are actually two special classes of manipulations that have been widely studied in the literature, and thus deserve particular attention.

The first is a class of strategies called truncations. Formally, P_i' is a **truncation** of a preference list P_i containing k acceptable schools if P_i' contains $k' < k$ acceptable schools and both P_i and P_i' rank the first k' schools in an identical manner. Many papers in the literature have focused on truncation strategies as an interesting and focal class of deviations. For instance, in searching for advice for participants in hospital-resident matching markets, Roth and Rothblum (1999) show that in low-information environments, any profitable deviation of the hospital-proposing DA algorithm is a truncation.[21]

Lemma 3. *Let ϕ be a stable-dominating, IIR mechanism. For any student i, no truncation strategy is an obvious manipulation of ϕ.*

Proof. Let P_i' be any truncation strategy. It is straightforward to show that $B_i^{\phi}(P_i)$ is i's favorite school. Therefore, the best-case outcome cannot be better under any alternative strategy. Let $\bar{w}$ be as defined in Lemma 1 (i's worst case assignment under any stable-dominating mechanism). First, suppose P_i' truncates i's preferences before $\bar{w}$. Let P_{-i} be a preference profile such that $DA_i(P) = \bar{w}$ ($\bar{w}$ is the worst possible assignment under DA, so such a profile exists). Under DA, when the other students submit preferences P_{-i}, i runs out of acceptable schools to apply to under preferences P_{-i}'; therefore, $DA_i(P_i', P_{-i}) = \emptyset$. In particular, under P_i', the worst-case assignment under DA is being unassigned. Since ϕ has the same worst-case assignment as DA, the worst-case under P_i ($\bar{w}$) is better than the worst case under P_i' ($\emptyset$). This also immediately implies part (iii) of the definition, and therefore, P_i' is not an obvious manipulation.

[21] Other papers that have analyzed truncation strategies include Roth and Vande Vate (1991), Roth and Peranson (1999), and Ehlers (2008). Kojima and Pathak (2009) consider a generalization of truncation strategies they call dropping strategies and show that dropping strategies are exhaustive when searching for manipulations for agents with a capacity greater than 1 (in the school choice model here, only the students are strategic, and they have unit capacity, i.e., they will only be matched to at most one school).

Finally, suppose instead that P_i' truncates i's preferences after $\bar{w}$. Under DA, i never proposes to a school worse than $\bar{w}$ ($\bar{w}$ is i's worst possible assignment under DA). Therefore, the ranking of schools below $\bar{w}$ has no impact on the DA assignment. In particular, for DA, i's worst case under P_i' is the same as i's worst case under P_i. Since ϕ has the same worst case as DA, ϕ's worst case under P_i' is also $\bar{w}$. Therefore, for every possible profile P_{-i}, $\phi_i(P_i', P_{-i}) R_i \bar{w}$ and by IIR, $\phi_i(P_i', P_{-i}) = \phi_i(P)$. Therefore, P_i' is not an obvious manipulation.[22]

□

Note that truncations do not alter the ordering of any schools above the truncation point. The second main class of manipulations that we rule out before completing the proof of Theorem 1 are those that do alter the relative ordering of some schools. Following Maskin (1999), we say that P_i' is a **non-Maskin-monotonic transformation of P_i at** s if there exists some s' such that $s \; P_i \; s'$, but $s' \; P_i' \; s$; in other words, in moving from P_i to P_i', there is some school s' that "jumps" over s in i's ranking. The next lemma shows that under a stable-dominating mechanism ϕ, it is never an obvious manipulation for a student to submit a non-Maskin monotonic transformation relative to $\bar{w}$, her worst possible assignment under ϕ.

Lemma 4. *Consider any stable-dominating, IIR mechanism ϕ. Let $\bar{w}$ be i's worst possible assignment under preferences P_i. Any non-Maskin monotonic transformation at $\bar{w}$ is not an obvious manipulation.*

Proof. It is straightforward to show that $B_i^{\phi}(P_i)$ is i's favorite school. Therefore, the best-case outcome cannot be better under any alternative strategy. Let $\bar{w}$ be as defined in Lemma 1 (i's worst case assignment under a stable-dominating mechanism). Consider a non-Maskin monotonic manipulation P_i', i.e. a P_i' such that there exists some $s \in S$ such that $sP_i'\bar{w}$, but $\bar{w}P_i s$. Intuitively, this will not be an obvious manipulation because it is now possible for i to be assigned to s, whereas under her true preferences, she is always assigned to a school she strictly prefers to s. We show this formally. In particular, fix s as i's favorite such school, i.e.:

$$s := \max_{P_i} \{s' | s' \; P_i' \; \bar{w} \text{ and } \bar{w} \; P_i \; s'\}.$$

[22] Indeed, P_i' is not even a (profitable) manipulation according to Definition 1.

Each school s' such that $s' \; P_i \; \bar{w}$ satisfies Hall's matching condition, which is to say it is possible to fill all of their seats with students ranked higher than i according to $\succ_{s'}$. By the nature of Hall's condition, this is also true for any subset of these schools. In particular, $\{s' | s' \; P_i' \; s\} \subseteq \{s' | s' \; P_i \; \bar{w}\}$ and consequently, each of the schools in $\{s' | s' \; P_i' \; s\}$ is an aspirational school under P_i'.

Therefore, for i's worst possible assignment under P_i', which we label $\bar{w}'$, it must be true that $s \; R_i' \; \bar{w}'$. Therefore, by Corollary 1, there exists a profile P'_{-i} such that $\phi_i(P') = s$. Since $\bar{w} P_i s$ and i is never assigned to a school worse than $\bar{w}$ under P_i, P_i' is not an obvious manipulation.

□

We are now ready to complete our proof of Theorem 1.

Proof of Theorem 1. Let ϕ be a stable-dominating, IIR mechanism, and consider a student i of type P_i. Let $\bar{w}$ be as defined in Lemma 1. We classify manipulations into two possible types: "monotonic" or "non-monotonic" (where monotonicity is relative to $\bar{w}$).

1. **Monotonic manipulation:** For all $a \in S$ such that $a P_i' \bar{w}$, we have $a P_i \bar{w}$.

2. **Non-monotonic manipulation:** There exists some $a \in S$ such that $a P_i' \bar{w}$, but $\bar{w} P_i' a$.

We have already proven in Lemma 4 that no non-monotonic manipulation is an obvious manipulation. Thus, consider a monotonic manipulation P_i'. Condition (ii) can be dispensed with immediately for any manipulation P_i', as it is easy to see that the best case from truth-telling is that agent i gets her (true) top choice. Next, consider condition (i). If $\bar{w}$ is ranked first under P_i', then if all students rank all schools as unacceptable, i is assigned to $\bar{w}$. Therefore, the worst possible case under P_i' cannot be strictly better than under P_i. Now suppose $\bar{w}$ is not ranked first under P_i'. For Hall's condition in Lemma 2 to be satisfied, every possible subset of schools preferred to $\bar{w}$ must have sufficient total capacity. Under a monotonic transformation, there are fewer possible subsets of schools preferred to $\bar{w}$; therefore, Hall's condition continues to hold. In particular, if $s \; P_i' \; \bar{w}$, then s is an aspirational school. Therefore, if $\bar{w}$ is a safety school, it is the most preferred safety school, and by the argument in Lemma 2, i's worst possible assignment. Alternatively,

$\bar{w}$ could be an aspirational school, but in either case, from Corollary 1, there exists a P'_{-i} such that $\phi_i(P') = \bar{w}$. From this, we can conclude that the worst case for i (under true preferences) from submitting P'_i is weakly worse than submitting her true preferences.

Last, consider condition (iii). Let n be the first place where P_i and P'_i rank a school differently and suppose i ranks school s in the n^{th} place under P_i and school s' under P'_i (where s or s' could possibly be $\emptyset$). Note that $s \ P_i \ s'$, and suppose $sR_i\bar{w}$.[23] By Corollary 1, there is a preference profile P_{-i} where $DA_i(P) = s$.[24] We define a new preference profile $\tilde{P}_{-i}$ as follows. If $DA_j(P) \ P_i \ s$, then we set $\tilde{P}_j := DA_j(P), \emptyset$. Otherwise, $\tilde{P}_j := \emptyset$. The preferences are constructed this way so that the set of stable assignments is trivial. If i submits P_i, then i is assigned to s and each school i prefers to s is filled with students that rank the school first. Therefore, $\phi(P_i, \tilde{P}_{-i}) = s$. If i submits P'_i, then i is assigned to s' and each school i prefers to s' is assigned to students that rank the school first. Therefore, $\phi(P'_i, \tilde{P}_{-i}) = s'$. In particular, $\phi_i(P_i, \tilde{P}_{-i}) \ P_i \ \phi_i(P'_i, \tilde{P}_{-i})$, and so condition (iii) is verified.

□

4 Other applications

While Section 3 focused mainly on school choice, the formal definition of an obvious manipulation is much more general, and can be applied to arbitrary mechanisms. In this section, we consider several canonical mechanism design settings (two-sided matching, auctions, bilateral trade). We look at well-known manipulable mechanisms in these settings, and show that our classification of mechanisms as OM or NOM continues to perform in line with intuition and practical experience.

4.1 Two-sided matching

The first setting we consider is closely related to the school choice model considered above: two-sided matching. Everything is the same as in Section

[23] If $\bar{w}P_is$, then, by IIR, for every P_{-i}, $\phi_i(P) = \phi_i(P'_i, P_{-i})$, and so P'_i is not a profitable manipulation.

[24] $DA(P)$ here denotes outcome of the the student-proposing deferred acceptance under preferences P.

3, except both sides of the market are treated as strategic agents with (strict) preferences over the other side. One potential application first considered in the seminal paper of Gale and Shapley (1962) is to marriage markets, where the two sides consist of men and women. Such models are also be used to study labor markets, where the two sides are workers and firms. Here, we partition the set of agents as $I = M \cup W$, where for convenience we refer to M as set of "men" and W as a set of "women". Each man $m \in M$ has a strict preference relation P_m over $W \cup \{\emptyset\}$, where $\emptyset$ is interpreted as remaining unmatched. Similarly, each woman $w \in W$ has a preference relation P_w over $M \cup \{\emptyset\}$. A **matching** here is a function $\mu : M \cup W \to M \cup W \cup \{\emptyset\}$ where $\mu_m = w$ denotes that man m is matched with woman w (and thus $\mu_w = m$); for any $i \in I$, $\mu_i = \emptyset$ means that agent i is unmatched. Stability is also defined equivalently as in Section 3. We additionally say that matching μ is **individually rational** if $\mu_i R_i \emptyset$. A mechanism ϕ is individually rational if $\phi_i(P)\ R_i\ \emptyset$ for all P, i.e., if it always produces an individually rational matching.

The key difference between two-sided matching and school choice is that both sides are strategic agents and are included in welfare considerations. Thus, while there is a strategy-proof and stable mechanism in the school choice model (student-proposing DA), this no longer holds when both sides are strategic, a result first shown by Roth (1982).

Theorem (Roth 1982)**.** *There exists no mechanism that is both stable and strategy-proof.*

Sönmez (1999) considers a far more general environment than just two-sided matching. His main result is much stronger than what we present, but in the context of two-sided matching (with strict preferences), it can be stated succinctly.

Theorem (Sönmez 1999)**.** *Given a matching problem (M, W, P_M, P_W), a mechanism ϕ is individually rational, Pareto efficient, and strategy-proof if and only if there is a unique stable assignment and ϕ chooses the stable assignment.*

Neither of these results continue to hold when we replace strategy-proofness with NOM. In particular, our next result shows that any stable mechanism is individually rational, Pareto efficient, and NOM. This has implications for markets such as the NRMP, which matches residents to hospitals using the doctor-proposing DA mechanism. While this mechanism (as well as any other stable mechanism) is technically manipulable by the hospitals, it is not obviously manipulable, and thus hospitals may find it difficult to execute profitable manipulations in practice. Formally, the result below follows readily from Theorem 1.

Theorem 2. *Any stable, IIR mechanism is individually rational, Pareto efficient, and not obviously manipulable.*

Proof. It is clear that any stable mechanism is individually rational and Pareto efficient. That a stable mechanism is NOM follows from Theorem 1. In particular, if a woman (man) had an obvious deviation, then she would also have an obvious deviation when the men (women) are treated as objects, which would contradict Theorem 1.

□

4.2 Single-Unit Auctions

Our remaining applications depart from what we have considered so far in that we allow for transfers. We also return to the notation of Section 2, where types are denoted by θ_i, outcomes by x, and utility functions $u_i(x;\theta_i)$. We begin by considering a simple first-price auction for a single good, and show that it is obviously manipulable.

An outcome is now denoted $x = (y,t)$, where $y \in \{0,1\}^{|I|}$ is an allocation vector such that $\sum_i y_i \leq 1$ and $t \in \mathbb{R}^{|I|}$ is a vector of transfers. Agent i's type space is $\Theta_i \subset \mathbb{R}_+$, and i's utility function when his type is $\theta_i \in \Theta_i$ is $u_i((y,t);\theta_i) = \mathbb{1}_{\{y_i=1\}}\theta_i - t_i$.[25] In a first-price auction, each agent submits a bid (which we take as equivalent to reporting his type), the highest bid wins and pays his bid, and all other bidders pay 0. Let $\phi^{FP}(\theta) = (y^{FP}(\theta), t^{FP}(\theta))$ denote the first-price auction mechanism, where $y_i^{FP}(\theta) = 1$ and $t_i^{FP}(\theta) = \theta_i$ if and only if $\theta_i > \theta_j$ for all $j \neq i$, and $y_i^{FP}(\theta) = t_i^{FP}(\theta) = 0$ otherwise.[26]

Proposition 2. *The first-price auction is obviously manipulable.*

This proposition follows straightforwardly from the definition. To see this, consider an agent of type θ_i, and an alternative report $0 < \theta_i' < \theta_i$. Under θ_i, both the worst and best cases are 0: $\min_{\theta_{-i}} u_i(\phi^{FP}(\theta_i,\theta_{-i});\theta_i) =$

[25] While quasilinear utility is commonly assumed in the auctions literature, it is not necessary for our results.

[26] In the event of a tie, the winner is chosen randomly among those who submitted the highest bid.

$\max_{\theta_{-i}} u_i(\phi^{FP}(\theta_i, \theta_{-i}); \theta_i) = 0$. Under θ_i', the worst-case is still 0 (when i loses), but the best case is strictly better: $\max_{\theta_{-i}} u_i(\phi^{FP}(\theta_i', \theta_{-i}); \theta_i) = \theta_i' > 0 = \max_{\theta_{-i}} u_i(\phi^{FP}(\theta_i', \theta_{-i}); \theta_i)$, and thus, according to Definition 2, reporting θ_i' is an obvious manipulation.

That the first-price auction is obviously manipulable is perhaps not surprising: it is "obviously" a bad strategy for a bidder to bid her true valuation, because she is sure to receive a payoff of 0, even in the best case scenario when she wins; bid shading at least gives the potential for some profit if she wins. However, this does not fully capture how "difficult" the first-price auction can be. While Proposition 2 gives a clear negative prescription of what not to do (bid truthfully), it does not positively answer the question of what one should do. Indeed, we can actually show a stronger negative result for first-price auctions, which is that not only is truth-telling susceptible to obvious manipulations, but in fact, for any possible strategy an agent may consider, there is an obvious deviation.

To state this formally, we first generalize the notion of an obvious manipulation beyond just direct revelation mechanisms to arbitrary mechanisms. A general **mechanism** (M, g) consists of a message space M_i for each agent i and an outcome function $g : M \to X$, where $M = \times_{i \in I} M_i$. Let $u_i(g(m_i, m_{-i}); \theta_i)$ be agent i's utility function when of type θ_i and the agents send the message profile (m_i, m_{-i}), resulting in outcome $g(m_i, m_{-i})$. Similar to Definition 1, we say that m_i' is a (profitable) **deviation from** m_i for type θ_i if there exists some m_{-i} such that $u_i(g(m_i', m_{-i}); \theta_i) > u_i(g(m_i, m_{-i}); \theta_i)$.

Definition 3. Consider a mechanism (M, g) and a deviation m_i' from m_i for type θ_i. Message m_i' is an **obvious deviation from** m_i for type θ_i if any of the following hold:

(i) $\min_{m_{-i}} u_i(g(m_i', m_{-i}); \theta_i) > \min_{m_{-i}} u_i(g(m_i, m_{-i}); \theta_i)$

(ii) $\max_{m_{-i}} u_i(g(m_i', m_{-i}); \theta_i) > \max_{m_{-i}} u_i(g(m_i, m_{-i}); \theta_i)$

(iii) $u_i(g(m_i', m_{-i}); \theta_i) \geq u_i(g(m_i, m_{-i}); \theta_i)$ for all m_{-i}.

If no such message m_i' exists, then we say that type θ_i has **no obvious deviations from** m_i.

Note that if we set $M_i = \Theta_i$, $g = \phi$, and $m_i = \theta_i$ (i.e., truthful reporting) in Definition 3, we recover the definition of an obvious manipulation from Definition 2. While we have seen several examples of manipulable mechanisms that admit message profiles with no obvious deviations (e.g., any direct mechanism that is not obviously manipulable), this need not hold in general; that is, for some mechanisms, it may be the case that starting from any strategy m_i, there is an alternative m_i' that is an obvious deviation.

Our next result shows that this is indeed the case for first-price auctions. We refer to the messages as *bids* b_i, and define each agent's message space as $M_i = B_i = \mathbb{R}_+$. For any bid vector $b \in B$, agent i's utility function is as $u_i(g(b_i, b_{-i}); \theta_i) = \theta_i - b_i$ if $b_i > b_j$ for all $j \neq i$ and 0 otherwise.[27]

Theorem 3. *In the first price auction, for any type $\theta_i > 0$ and any possible bid b_i, there is an alternative bid b_i' that is an obvious deviation from b_i.*

Proof. Fix an agent i of type $\theta_i > 0$, and consider a bid b_i. First, it is immediately clear that any $b_i > \theta_i$ has an obvious deviation to $b_i' = \theta_i$. Thus, consider some strategy b_i such that $\theta_i > b_i > 0$. Let $b_i' = b_i - \epsilon$ be an alternative bid for some small $\epsilon \in (0, b_i)$. In the table below, we calculate the worst and best cases from b_i and b_i'.

Bid	$\min_{b_{-i}} u_i(g(\cdot, b_{-i}); \theta_i)$	$\max_{b_{-i}} u_i(g(\cdot, b_{-i}); \theta_i)$
b_i	0	$\theta_i - b_i$
b_i'	0	$\theta_i - b_i - \epsilon$

So, the worst cases are equivalent, but $\max_{b_{-i}} u_i(g(b_i', b_{-i}); \theta_i) > \max_{b_{-i}} u_i(g(b_i, b_{-i}); \theta_i)$, and thus b_i' is an obvious deviation from b_i.

We have eliminated all $b_i > 0$. Last, consider $b_i = 0$ and some bid $b_i' = \epsilon$ for some small $\epsilon > 0$. We calculate the same table as above for this case (recall that if there is a tie, the winner is chosen randomly from among those who submitted the highest bid).

Bid	$\min_{b_{-i}} u_i(g(\cdot, b_{-i}); \theta_i)$	$\max_{b_{-i}} u_i(g(\cdot, b_{-i}); \theta_i)$
$b_i = 0$	0	θ_i/N
$b_i' = \epsilon$	0	$\theta_i - \epsilon$

Note that if i bids $b_i = 0$, then the only chance for her to win the object is if all other bidders also bid 0, in which case her payoff is θ_i/N, and so $\max_{b_{-i}} u_i(g(0, b_{-i}); \theta_i) = \frac{\theta_i}{N}$. If she bids ϵ, the best case is $\max_{b_{-i}}(g(\epsilon, b_{-i}); \theta_i) =$

[27] In the event of a tie, the winner is determined randomly, and so $u_i(g(b_i, b_{-i}); \theta_i) = (\theta_i - b_i)/r$, where r is the number of bidders who submitted the highest bid.

$\theta_i - \epsilon$. For small enough ϵ, $\theta_i - \epsilon > \frac{\theta_i}{N}$ and so $b_i' = \epsilon$ is an obvious deviation from $b_i = 0$.

□

The upshot of Theorem 3 is that agents for whom mins and maxes are salient outcomes will have a very difficult time deciding on an optimal strategy, in the sense that starting from any given strategy, it is very easy to recognize scenarios in which there is an alternative strategy that will perform better. Note also that just as Theorem 3 generalizes Proposition 2, we can similarly use Definition 3 to generalize Proposition 1: in particular, in the Boston mechanism, for any student who can guarantee a seat at some school that is not her first choice, every possible strategy (ranking of schools) has an obvious deviation.

4.3 Multi-Unit Auctions

In single-unit auctions, the first-price auction is (obviously) manipulable, while the second-price auction is famously strategy-proof (Vickrey, 1961). For our purposes, multi-unit auctions are actually more interesting, because while the analogue of the first-price auction, the pay-as-bid auction, is still (obviously) manipulable, the analogue of the second-price auction is no longer formally strategy-proof. In this section, we show that while this auction is manipulable, it is not obviously so.

The auctioneer now has K identical objects to be sold. Let $y_i \in \{0, 1, \ldots, K\}$ denote the number of units assigned to agent i, and $t_i \in \mathbb{R}$ be the payment of agent i. Defining $y = (y_1, \ldots, y_N)$ and $t = (t_1, \ldots, t_N)$, an outcome is a vector $x = (y, t)$ such that $\sum_i y_i \leq K$. Bidder i's type is a K-dimensional vector $\theta_i = (\theta_i^1, \ldots, \theta_i^K)$. Because the objects are identical, it is without loss of generality to assume that $\theta_i^1 \geq \theta_i^2 \geq \cdots \geq \theta_i^K$ for all $\theta_i \in \Theta_i$. The utility of a bidder of type θ_i is $u_i((y, t); \theta_i) = \sum_{\ell=0}^{y_i} \theta_i^\ell - t_i$.

The natural counterpart of the first-price auction is the **pay-as-bid auction** (sometimes also called the *discriminatory price auction*): each bidder submits a vector of bids for each of the K units (which we take as reporting her type, and which may be 0 for some units), and pays the sum of her winning bids. Indeed, the first-price auction introduced above is a special case of a pay-as-bid auction, and the same arguments can be used to prove the following.

Corollary 2. *The pay-as-bid auction is obviously manipulable.*

A natural extension of Theorem 3 also continues to hold for pay-as-bid auctions. The arguments are analogous, and so rather than repeating them, we instead move directly to analyzing the $(K+1)-$price auction. The $(K+1)-$price auction works as follows. Each agent again submits a bid for each of the K units (some of which may be 0). All of the bids are ordered from highest to lowest. The K units are awarded to the K highest submitted bids, with the price of each unit equal to the $(K+1)^{th}$ highest bid. Note that when $K = 1$, we recover the second-price auction, which is strategy-proof. While for any $K > 1$ the $(K+1)-$price auction is not strategy-proof, it is intuitively much less susceptible to manipulation than the pay-as-bid auction. Our next result formalizes this intuition.[28]

Theorem 4. *The* $(K+1)-$*price auction is not obviously manipulable.*

Proof. Let ϕ^{K+1} denote the $(K+1)$-price auction mechanism. Consider an agent of type θ_i, and first consider reporting truthfully. It is simple to calculate that $\min_{\theta_{-i}} u_i(\phi^{K+1}(\theta_i, \theta_{-i}); \theta_i) = 0$ and $\max_{\theta_{-i}} u_i(\phi^{K+1}(\theta_i, \theta_{-i}); \theta_i) = \sum_{k=1}^{K} \theta_i^k$. We must show that for any manipulation $\tilde{\theta}_i \neq \theta_i$, parts (i)-(iii) of Definition 2 all hold. First, it should be clear again that for any $\tilde{\theta}_i$, we have $\min_{\theta_{-i}} u_i(\phi^{K+1}(\tilde{\theta}_i, \theta_{-i}); \theta_i) = 0$ and $\max_{\theta_{-i}} u_i(\phi^{K+1}(\tilde{\theta}_i, \theta_{-i}); \theta_i) = \sum_{k=1}^{K} \theta_i^k$. Therefore, we have $\min_{\theta_{-i}} u_i(\phi^{K+1}(\tilde{\theta}_i, \theta_{-i}); \theta_i) = \min_{\theta_{-i}} u_i(\phi^{K+1}(\theta_i, \theta_{-i}); \theta_i)$ and $\max_{\theta_{-i}} u_i(\phi^{K+1}(\tilde{\theta}_i, \theta_{-i}), \theta_{-i}) = \max_{\theta_{-i}} u_i(\phi^{K+1}(\theta_i, \theta_{-i}), \theta_{-i})$, and so parts (i) and (ii) of Definition 2 are satisfied.

Last, we show part (iii) by exhibiting a report of the others θ_{-i} such that $u_i(\phi^{K+1}(\theta_i, \theta_{-i}); \theta_i) > u_i(\phi^{K+1}(\tilde{\theta}_i, \theta_{-i}), \theta_{-i})$. It is without loss of generality to assume that $\theta_i^{k'} \geq \theta_i^k$ for all $k' \leq k$, and similarly $\tilde{\theta}_i^{k'} \geq \tilde{\theta}_i^k$ for all $k' \leq k$ (i.e., the valuations/bids are ordered from highest to lowest). First, consider the case where there exists some k such that $\tilde{\theta}_i^k > \theta_i^k$, and let k be the lowest index for which this is true. Consider a profile θ_{-i} such that $\theta_j^k = \frac{\tilde{\theta}_i^k + \theta_i^k}{2}$ for all $j \neq i$ and all $k = 1, \ldots K$. Under profile $(\tilde{\theta}_i, \theta_{-i})$, the $(K+1)^{th}$-highest bid, and

[28] Indeed, in the specific context of Treasury auctions, Friedman (1991) proposed a switch from a pay-as-bid auction format to $(K+1)$-price auction format (sometimes called a *uniform-price auction*), precisely as a way to reduce strategizing and bid shading. His proposal was eventually adopted, and is still used today. Our results provide a formal theoretical justification for this intuition (see also Pathak and Sönmez (2013) and Azevedo and Budish (2018) for complementary analyses).

thus the final price, is $p = \frac{\tilde{\theta}_i^k + \theta_i^k}{2}$. So, at $(\tilde{\theta}_i, \theta_{-i})$, i wins k units and pays a total price of kp, giving her utility

$$u_i \left(\phi^{K+1}(\tilde{\theta}_i, \theta_{-i}); \theta_i \right) = \sum_{\ell=1}^{k} \theta_i^\ell - kp = \sum_{\ell=1}^{k} \left(\theta_i^\ell - \frac{\tilde{\theta}_i^k + \theta_i^k}{2} \right) \tag{6}$$

On the other hand, if i had reported truthfully, the price remains the same, but she only wins $k-1$ units, and so

$$u_i \left(\phi^{K+1}(\theta_i, \theta_{-i}); \theta_i \right) = \sum_{\ell=1}^{k-1} \theta_i^\ell - (k-1)p = \sum_{\ell=1}^{k-1} \left(\theta_i^\ell - \frac{\tilde{\theta}_i^k + \theta_i^k}{2} \right) \tag{7}$$

Subtracting (6) from (7), we get

$$u_i \left(\phi^{K+1}(\theta_i, \theta_{-i}); \theta_i \right) - u_i \left(\phi^{K+1}(\tilde{\theta}_i, \theta_{-i}); \theta_i \right) = \frac{\tilde{\theta}_i^k + \theta_i^k}{2} - \theta_i^k > 0,$$

where the last inequality follows because $\tilde{\theta}_i^k > \theta_i^k$.

The remaining case to consider is $\theta_i^k \geq \tilde{\theta}_i^k$ for all $k = 1, \ldots, K$. Let k be the lowest index for which this inequality is strict (if there is no such k, then $\tilde{\theta}_i^k$ is equivalent to truth-telling). Again, consider a profile θ_{-i} such that $\theta_j^k = \frac{\tilde{\theta}_i^k + \theta_i^k}{2}$. Under profile $(\tilde{\theta}_i, \theta_{-i})$, the price will be $p = \frac{\tilde{\theta}_i^k + \theta_i^k}{2}$, and agent i will win $k-1$ objects at this price, for a utility of

$$u_i \left(\phi^{K+1}(\tilde{\theta}_i, \theta_{-i}); \theta_i \right) = \sum_{\ell=1}^{k-1} \theta_i^\ell - (k-1)p = \sum_{\ell=1}^{k-1} \left(\theta_i^\ell - \frac{\tilde{\theta}_i^k + \theta_i^k}{2} \right) \tag{8}$$

If instead i reported θ_i truthfully, the price will again be p, and he will win k objects at total price kp, for a utility of

$$u_i \left(\phi^{K+1}(\theta_i, \theta_{-i}); \theta_i \right) = \sum_{\ell=1}^{k} \theta_i^\ell - kp = \sum_{\ell=1}^{k} \left(\theta_i^\ell - \frac{\tilde{\theta}_i^k + \theta_i^k}{2} \right) \tag{9}$$

Subtracting (8) from (9), we have

$$u_i \left(\phi^{K+1}(\theta_i, \theta_{-i}); \theta_i \right) - u_i \left(\phi^{K+1}(\tilde{\theta}_i, \theta_{-i}); \theta_i \right) = \theta_i^k - \frac{\tilde{\theta}_i^k + \theta_i^k}{2} > 0,$$

where the final inequality follows because in this case, $\theta_i^k > \tilde{\theta}_i^k$. Therefore, condition (iii) of Definition 2 holds, which completes the proof.

□

4.4 Bilateral Trade

As a final application, we consider the classic bilateral trade setting. The set of agents is $I = \{B, S\}$, where B is a potential buyer and S a seller of a single object. We normalize the type spaces for both the buyer and the seller to $\Theta_S = \Theta_B = [0, 1]$, where $\theta_S \in \Theta_S$ is the seller's cost to produce the object, and $\theta_B \in \Theta_B$ is the buyer's value for the object. Each agent knows their own type, but not the type of the other agent.

A mechanism here is written $\phi(\theta) = (y(\theta), t_B(\theta), t_S(\theta))$, where for any $\theta = (\theta_B, \theta_S)$, $y(\theta) \in \{0, 1\}$ denotes whether or not trade occurs, $t_B(\theta)$ is the transfer from the buyer, and $t_S(\theta)$ is the transfer to the seller. Given a mechanism ϕ and reported types $(\hat{\theta}_B, \hat{\theta}_S)$, utilities are thus written

$$U_B(\phi(\hat{\theta}_B, \hat{\theta}_S); \theta_B) = \theta_B y(\hat{\theta}_B, \hat{\theta}_S) - t_B(\hat{\theta}_B, \hat{\theta}_S)$$
$$U_S(\phi(\hat{\theta}_B, \hat{\theta}_S); \theta_S) = -\theta_S y(\hat{\theta}_B, \hat{\theta}_S) + t_S(\hat{\theta}_B, \hat{\theta}_S)$$

We first consider one of the simplest and most well-known mechanisms for this setting, the double auction mechanism analyzed by Chatterjee and Samuelson (1983). In this mechanism, each agent reports her type. If $\theta_B \geq \theta_S$, then trade occurs at a price $p = \frac{\theta_B + \theta_S}{2}$; otherwise, no trade occurs, and no transfers are made. Formally:

$$y(\theta) = \begin{cases} 1, & \theta_B \geq \theta_S \\ 0, & \theta_B < \theta_S \end{cases} \qquad t_S(\theta) = t_B(\theta) = \begin{cases} \frac{\theta_B + \theta_S}{2}, & \theta_B \geq \theta_S \\ 0, & \theta_B < \theta_S \end{cases}$$

Proposition 3. *The double auction mechanism is obviously manipulable.*

To see this, consider a buyer of type θ_B, and let $\theta'_B = \theta_B - \epsilon$ (a completely analogous argument can be made for the seller). Then, it is simple to calculate that $\max_{\theta_S} U_B(\phi(\theta'_B, \theta_S); \theta_B) = \theta_B/2 + \epsilon/2$, while $\max_{\theta_S} U_B(\phi(\theta_B, \theta_S); \theta_B) = \theta_B/2$. Therefore, θ'_B is an obvious manipulation.

Myerson and Satterthwaite (1983) prove a general theorem that in this setting, there is no efficient, individually rational, and Bayesian incentive compatible mechanism (without the infusion of an outside subsidy). One common interpretation of this negative result is that two-sided private information introduces "transaction costs" that preclude efficient bargaining (a la Coase, 1960); in other words, in the presence of asymmetric information, there is a fundamental conflict between incentives and efficiency.

More recent work on mechanism design under ambiguity has re-evaluated these claims by considering agents who may not be classical expected utility maximizers, but instead are ambiguity averse. For instance, De Castro and Yannelis (2018) argue that ambiguity "solves" the conflict between incentives and efficiency. In particular, they show that if agents have maximin preferences, then an efficient, incentive compatible, individually rational, and budget-balanced mechanism exists, and further, one such mechanism is the double auction mechanism described above.[29] The intuition is that the worst case from any report is that trade does not occur, and so when agents evaluate outcomes using maximin preferences, all reports are equivalent, and everyone is willing to report truthfully. While this requires an arguably quite strong assumption that agents are completely pessimistic and certain trade will not occur, Wolitzky (2016) considers a more general model of ambiguity averse agents and shows that there are still conditions under which the conclusion of the Myerson-Satterthwaite theorem is "reversed".

The agents in our model also compare worst (and best) case outcomes, but in a different way, and in particular one that reinforces Myerson and Satterthwaite's original insight. To see what we mean, first, note that Proposition 3 shows that double auctions are obviously manipulable (an extreme form of *non*-incentive compatibility), which is in contrast to results that show such a mechanism is incentive compatible when agents are ambiguity averse. Second, we can extend this beyond double auctions and further prove an analogue to Myerson and Satterthwaite's impossibility theorem for general mechanisms. Following this literature, we consider mechanisms $\phi(\theta) = (y(\theta), t_B(\theta), t_S(\theta))$ that satisfy the following properties:[30]

[29] Borgers and Li (2018) define a relaxation of strategy-proofness called strategic simplicity. They show that double auctions are not strategically simple.

[30] Myerson and Satterthwaite (1983) assume an interim version of individual rationality; however, one of the goals of our project is to move away from a reliance on prior distributions, and so an ex-post formulation of individual rationality is more appropriate for our setting.

1. Efficiency: $y(\theta_B, \theta_S) = 1$ if and only if $\theta_B \geq \theta_S$.
2. Individual rationality: $U_B(\phi(\theta_B, \theta_S); \theta_B) \geq 0$ and $U_S(\phi(\theta_B, \theta_S); \theta_S) \geq 0$ for all (θ_B, θ_S).
3. (Weak) budget balance: $t_S(\theta) \leq t_B(\theta)$ for all θ.

We then have the following result.

Theorem 5. *Every efficient, individually rational, and weakly budget balanced mechanism is obviously manipulable.*

Proof. Assume that $\phi(\theta) = (y(\theta), t_B(\theta), t_S(\theta))$ is an efficient, individually rational, weakly budget-balanced mechanism that is not obviously manipulable. Define

$$\bar{p}_S = \max_{\theta \text{ s.t. } y(\theta)=1} t_S(\theta)$$

$$\underline{p}_B = \min_{\theta \text{ s.t. } y(\theta)=1} t_B(\theta).$$

In words, $\bar{p}_S$ is the highest possible price the seller may receive, conditional on selling the object and $\underline{p}_B$ is the lowest possible price the buyer may pay, conditional on buying the object.

Now, note that efficiency combined with individual rationality imply the following about t_S and t_B:

$$t_S(\theta_B, \theta_S) \geq \theta_S \text{ for all } \theta_B \geq \theta_S \tag{10}$$

$$t_B(\theta_B, \theta_S) \leq \theta_B \text{ for all } (\theta_B, \theta_S). \tag{11}$$

(for the first line, we must have $y(\theta_B, \theta_S) = 1$ for all $\theta_B \geq \theta_S$, by efficiency; IR then says $t_S(\theta_B, \theta_S) \geq \theta_S$. The second line is immediate from the buyer's IR constraint.) Now, equations (10) and (11) imply $\bar{p}_S \geq 1$ and $\underline{p}_B = 0$ (for the former, substitute $(\theta_B, \theta_S) = (1, 1)$, and for the latter, substitute $(\theta_B, \theta_S) = (0, 0)$). By weak budget-balance, $t_S(\theta_B, \theta_S) \leq t_B(\theta_B, \theta_S) \leq \theta_B \leq 1$ for all (θ_B, θ_S), and so the former inequality is actually an equality: $\bar{p}_S = 1$.

Consider some type of the seller $\theta_S < 1$. Note that $\bar{p}_S = 1$ implies that $\max_{\theta'_B} U_S(\phi(\theta'_B, 1); \theta_S) = 1 - \theta_S$. For ϕ to be not obviously manipulable then requires that $\max_{\theta'_B} U_S(\phi(\theta'_B, \theta_S); \theta_S) \geq 1 - \theta_S$ for all θ_S; in other words, we must have $\max_{\theta'_B} t_S(\theta'_B, \theta_S) = 1$ for all θ_S. Since $t_S(\theta'_B, \theta_S) \leq \theta'_B$, the only possibility is that $t_S(1, \theta_S) = 1$ for all θ_S. On the other hand, consider a buyer of type $\theta_B > 0$, and note that $\max_{\theta'_S} U_S(\phi(0, \theta'_S); \theta_B) = \theta_B$. Again, NOM implies that $\max_{\theta'_S} U_B(\phi(\theta_B, \theta'_S); \theta_B) \geq \theta_B$ for all θ_B; in other words, for all θ_B, there must exist some θ'_S such that $y(\theta_B, \theta'_S) = 1$ and $t_B(\theta_B, \theta'_S) = 0$. Budget balance and the seller's IR constraint imply that the only possibility is $\theta'_S = 0$, i.e., for all θ_B, we must have $y(\theta_B, 0) = 1$ and $t_B(\theta_B, 0) = 0$.

To summarize, we have shown that if ϕ is an efficient, individually rational, weakly budget balanced, and NOM mechanism, then the following must be true: (i) $y(1, \theta_S) = 1$ and $t_S(1, \theta_S) = 1$ for all θ_S, and (ii) $y(\theta_B, 0) = 1$ and $t_B(\theta_B, 0) = 0$ for all θ_B. In particular, setting $\theta_S = 0$ in (i) and $\theta_B = 1$ in (ii) gives $t_S(1, 0) = 1$ and $t_B(1, 0) = 0$, which contradicts weak budget balance.

□

5 Conclusion

Market design is fortunate in that there are known, strategy-proof mechanisms that achieve attractive market outcomes. In school choice, the Deferred Acceptance algorithm is strategy-proof for students and respects priorities. When there is a single unit for sale, the second price auction is strategy-proof and efficient without any sacrifice to revenue (in expectation). In more general auction environments, the celebrated Vickrey auction is a strategy-proof and efficient mechanism, and yet it is rarely seen in practice (Ausubel et al. (2006) refer to it as the "lovely but lonely Vickrey auction"); in this environment, the constraints on revenue imposed by strategy-proofness are too great. The school choice literature, on the other hand, has had difficultly moving past strategy-proof mechanisms, even though it may also come at a heavy cost in terms of efficiency (e.g., Abdulkadiroğlu et al., 2009).[31]

In markets where a planner attempts to achieve a more desirable outcome by using a non-strategy-proof mechanism, they must ask: to what extent are the gains undone by strategic behavior of the agents? For example, in school choice, both the Boston mechanism and the EADA mechanism are efficient if students submit their true preferences, but neither is strategy-proof. The key question then becomes: to what extent are real-world agents able to identify

[31] There is a strategy-proof and efficient mechanism that has been proposed for school choice, the top trading cycles mechanism (a generalization of the famous algorithm introduced by Shapley and Scarf, 1974), but it suffers from the drawback that it does a very poor job at respecting priorities, and thus is rarely used in practice.

and enact manipulations under a particular mechanism, and thereby negate theoretical gains in efficiency (with respect to true preferences) via untruthful reporting?

This paper provides an intuitive and tractable taxonomy for determining when it will be obvious to participants that a mechanism can be manipulated. If it is obvious to participants that a mechanism can be manipulated, then a policy maker should be skeptical that any properties relative to the agents' true preferences will be retained in practice; the Boston mechanism and pay-as-bid multi-unit auctions are examples of obviously manipulable mechanisms, and indeed have reputations of being easily manipulated in practice. Alternatively, if it is not obvious that a mechanism can be manipulated, then there is reason to be optimistic that improvements will be realized; the $(K+1)-$price auction and doctor-proposing DA mechanism (for two-sided matching markets) are examples of mechanisms that are manipulable, but are not obviously manipulable, and indeed seem to perform well in practice. The EADA mechanism is also manipulable, but not obviously so, and thus we would expect it to perform relatively well, especially given that it has many desirable features outlined in other work. We further hope that the ideas in this paper lead to the design of new mechanisms that, while not formally strategy-proof, are not obviously manipulable, and thus more likely achieve desired outcomes in practice.

Appendix A Definition of the Mechanisms

In this appendix, we give formal definitions of the mechanisms analyzed in Section 3 (these definitions are taken directly from Dur and Morrill, 2018).

Boston Mechanism:

For a given problem P, BM mechanism selects its outcome through the following mechanism:

Step 1**:** Each student applies to her most preferred school. Each school s accepts the best students according to its priority list, up to q_s, and rejects the rest.

Step $k > 1$**:** Each student rejected in Step $k-1$ applies to her k^{th} choice. Each school s accepts the best students among the new applicants, up to the number of remaining seats, and rejects the rest.

School-Proposing DA Mechanism:

For a given problem P, school-proposing DA mechanism selects its outcome through the following mechanism:

Step 1**:** Each school s proposes to top q_s students under $\succ_s$. Each student i accepts the best proposal it gets according to P_i, and rejects the rest.

Step $k > 1$**:** Each school s proposes to top q_s students under $\succ_s$ who have not rejected it yet. Each student i accepts the best proposal it gets according to P_i, and rejects the rest.

Top Trading Cycles Mechanism:

For a given problem P, TTC mechanism selects its outcome through the following mechanism:

Step 0**:** Assign a counter to each school and set it equal to the quota of each school.

Step 1**:** Each student points to her most preferred school among those remaining. Each remaining school points to the top-ranked student in its priority order. Due to the finiteness there is at least one cycle.[32] Assign each student in a cycle to the school she points to and remove her. The counter of each school in a cycle is reduced by one and if it reduces to zero, the school is removed.

Step $k > 1$**:** Each student points to her most preferred school among the remaining ones. Each remaining school points to the student with the highest priority among the remaining ones. There is at least one cycle. Assign each student in a cycle to the school she points to and remove her. The counter of each school in a cycle is reduced by one and if it reduces to zero, the school is also removed.

Deferred Acceptance-Top Trading Cycles Mechanism

For a given problem P, DA-TTC mechanism selects its outcome through the following mechanism:

Round DA**:** Run the DA mechanism. Update the priorities by giving the highest priorities for each school to the students assigned to it.

Round TTC: Run the TTC mechanism by using the preference profile and updated priorities.

Efficiency-Adjusted Deferred Acceptance Mechanism:

In order to define the mechanism selecting the outcome of EADAM, we first present a notion that we use in the definition. If student i is tentatively

[32] A cycle is an ordered list of distinct schools and distinct students $(s_1, i_1, s_2, ..., s_k, i_k)$ where s_1 points to i_1 , i_1 points to s_2 , ... , s_k points to i_k , i_k points to s_1 .

accepted by school s at some step t and is rejected by s in a later step t' of DA and if there exists another student j who is rejected by s in step $t'' \in \{t, t+1, ..., t'-1\}$, then i is called an **interrupter** for s and (i, s) is called an **interrupting pair** of step t'. Under EADAM, each student decides to consent or not. For a given problem P and consent decisions, EADAM selects its outcome through the following algorithm:

Round 0**:** Run the DA mechanism.

Round $k > 0$: Find the last step of the DA run in Round $k-1$ in which a consenting interrupter is rejected from the school for which she is an interrupter. Identify all the interrupting pairs of that step with consenting interrupters. For each identified interrupting pair (i, s), remove s from the preferences of i without changing the relative order of the other schools. Rerun the DA algorithm with the updated preference profile. If there are no more consenting interrupters, stop.

References

Abdulkadiroğlu, A. and T. Sönmez (2003): "School choice: A mechanism design approach," *American economic review*, 729–747.

Abdulkadiroğlu, A., P. A. Pathak, and A. E. Roth (2009): "Strategy-proofness versus Efficiency in Matching with Indifferences: Redesigning the NYC High School Match," *American Economic Review*, 99, 1954–1978.

Abdulkadiroğlu, A., P. A. Pathak, and A. E. Roth (2009): "Strategy-proofness versus Efficiency in Matching with Indifferences: Redesigning the NYC High School Match," 99, 1954–1978.

Abdulkadiroğlu, A. and T. Sönmez (2003): "School Choice: A Mechanism Design Approach," *American Economic Review*, 93, 729–747.

Alcalde, J. and A. Romero-Medina (2017): "Fair student placement," *Theory and Decision*, 83, 293–307.

Alva, S. and V. Manjunath (2017): "Strategy-proof Pareto-improvement under voluntary participation," .

Arribillaga, R. P., J. Massó, and A. Neme (2017): "Not All Majority-based Social Choice Functions Are Obviously Strategy-proof," .

Ashlagi, I. and Y. A. Gonczarowski (2018): "Stable matching mechanisms are not obviously strategy-proof," *Journal of Economic Theory*, 177, 405–425.

Ausubel, L. M., P. Milgrom, et al. (2006): "The lovely but lonely Vickrey auction," *Combinatorial auctions*, 17, 22–26.

Azevedo, E. M. and E. Budish (2018): "Strategyproofness in the Large," *Review of Economic Studies*, forthcoming.

Bade, S. and Y. A. Gonczarowski (2016): "Gibbard-Satterthwaite Success Stories and Obvious Strategyproofness," *arXiv preprint arXiv:1610.04873*.

Balinski, M. and T. Sönmez (1999): "A Tale of Two Mechanisms: Student Placement," 84, 73–94.

Barberà, S. and B. Dutta (1995): "Protective behavior in matching models," *Games and Economic Behavior*, 8, 281–296.

Borgers, T. and J. Li (2018): "Strategically simple mechanisms," .

Budish, E. B. and J. B. Kessler (2017): "Can Agents "Report Their Types"? An Experiment that Changed the Course Allocation Mechanism at Wharton," *Chicago Booth Research Paper*.

Carroll, G. (2011): "A Quantitative Approach to Incentives: Application to Voting Rules," Working paper, MIT.

Charness, G. and D. Levin (2009): "The origin of the winner's curse: a laboratory study," *American Economic Journal: Microeconomics*, 1, 207–36.

Chatterjee, K. and W. Samuelson (1983): "Bargaining under incomplete information," *Operations research*, 31, 835–851.

Coase, R. H. (1960): "The problem of social cost," in *Classic papers in natural resource economics*, Springer, 87–137.

De Castro, L. I. and N. C. Yannelis (2018): "Uncertainty, efficiency and incentive compatibility: Ambiguity solves the conflict between efficiency and incentive compatibility," *Journal of Economic Theory*, 177, 678–707.

Dur, U. (2018): "The Modified Boston Mechanism," *Mathematical Social Sciences*.

Dur, U., A. Gitmez, and O. Yilmaz (2015): "School Choice Under Partial Fairness," Tech. rep., Working paper, North Carolina State University, 2015.[19].

Dur, U., R. G. Hammond, and T. Morrill (2018): "Identifying the harm of manipulable school-choice mechanisms," *American Economic Journal: Economic Policy*, 10, 187–213.

Dur, U. and T. Morrill (2018): "What You Don't Know Can Help You in School Assignment," *mimeo.*

Dworczak, P. (2016): "Deferred acceptance with compensation chains," in *Proceedings of the 2016 ACM Conference on Economics and Computation*, ACM, 65–66.

Ehlers, L. (2008): "Truncation Strategies in Matching Markets," *Mathematics of Operations Research*, 33, 327–335.

Ehlers, L. and T. Morrill (2017): "(Il) Legal Assignments in School Choice," .

Erdil, A. and H. Ergin (2008): "What's the matter with tie-breaking? Improving efficiency in school choice," *American Economic Review*, 98, 669–689.

Esponda, I. and E. Vespa (2014): "Hypothetical thinking and information extraction in the laboratory," *American Economic Journal: Microeconomics*, 6, 180–202.

Fernandez, M. A. (2018): "Deferred Acceptance and Regret-free Truthtelling: A Characterization Result," Ph.D. thesis, California Institute of Technology.

Friedman, M. (1991): "How to sell government securities," *Wall Street Journal*, A8.

Gale, D. and L. S. Shapley (1962): "College Admissions and the Stability of Marriage," 69, 9–15.

Gilboa, I. and D. Schmeidler (1989): "Maxmin expected utility with non-unique prior," *Journal of mathematical economics*, 18, 141–153.

Hall, P. (1935): "On representatives of subsets," *Journal of London Mathematical Society*, 10, 26ñ30.

Harless, P. (2016): "Immediate acceptance with or without skips: comparing school assignment procedures," Tech. rep., Mimeo.

Immorlica, N. and M. Mahdian (2005): "Marriage, Honesty, and Stability," in *Proceedings of the 16th ACM-SIAM Symposium on Discrete Algorithms*, 53–62.

Kesten, O. (2010): "School Choice with Consent," *Quarterly Journal of Economics*, 125, 1297–1348.

Kojima, F. and P. A. Pathak (2009): "Incentives and Stability in Large Two-Sided Matching Markets," *American Economic Review*, 99, 608–627.

Li, S. (2017): "Obviously Strategy-Proof Mechanisms," *American Economic Review*, 107, 3257–87.

MASKIN, E. (1999): "Nash Equilibrium and Welfare Optimality," 66, 23–38.

MENNLE, T. AND S. SEUKEN (2014): "The Naïve versus the Adaptive Boston Mechanism," *arXiv preprint arXiv:1406.3327*.

MIRALLES, A. (2009): "School choice: The case for the Boston mechanism," in *Auctions, Market Mechanisms and Their Applications*, Springer, 58–60.

MYERSON, R. B. AND M. A. SATTERTHWAITE (1983): "Efficient mechanisms for bilateral trading," *Journal of economic theory*, 29, 265–281.

PATHAK, P. A. AND T. SÖNMEZ (2008): "Leveling the playing field: Sincere and sophisticated players in the Boston mechanism," *The American Economic Review*, 98, 1636–1652.

——— (2013): "School admissions reform in Chicago and England: Comparing mechanisms by their vulnerability to manipulation," *The American Economic Review*, 103, 80–106.

PYCIA, M. AND P. TROYAN (2016): "Obvious Dominance and Random Priority," .

ROTH, A. E. (1982): "The Economics of Matching: Stability and Incentives," *Mathematics of Operations Research*, 7, 617–628.

ROTH, A. E. AND E. PERANSON (1999): "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design," *American Economic Review*, 89, 748–780.

ROTH, A. E. AND U. ROTHBLUM (1999): "Truncation Strategies in Matching Markets: In Search of Advice for Participants," *Econometrica*, 67, 21–43.

ROTH, A. E. AND M. SOTOMAYOR (1990): *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, Cambridge University Press.

ROTH, A. E. AND J. VANDE VATE (1991): "Incentives in two-sided matching with random stable mechanisms," *Economic Theory*, 1, 31–44.

SHAPLEY, L. AND H. SCARF (1974): "On Cores and Indivisibility," 1, 23–37.

SÖNMEZ, T. (1999): "Strategy-Proofness and Essentially Single-Valued Cores," *Econometrica*, 67, 677–690.

TANG, Q. AND Y. ZHANG (2017): "Weak Stability and Pareto Efficiency in School Choice," .

TROYAN, P. (2016): "Obviously Strategy-Proof Implemenation of Top Trading Cycles," *working paper, University of Virginia*.

TROYAN, P., D. DELACRETAZ, AND A. KLOOSTERMAN (2018): "Essentially Stable Matchings," *working paper*.

VICKREY, W. (1961): "Counterspeculation, Auctions and Competitive Sealed

Tenders," *Journal of Finance*, 16, 8–37.

Wilson, R. (1987): *Game-Theoretic Analyses of Trading Processes*, Cambridge University Press., chap. 2, 33–70.

Wolitzky, A. (2016): "Mechanism design with maxmin agents: Theory and an application to bilateral trade," *Theoretical Economics*, 11, 971–1004.