Obvious Manipulations

Peter Troyan
Department of Economics
University of Virginia

Thayer Morrill*
Department of Economics
North Carolina State University

July 9, 2019

Abstract

A mechanism is *strategy-proof* if agents can never profitably manipulate it, 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.

*Authors' names appear in random order. Emails: troyan@virginia.edu and thayer_morrill@ncsu.edu. We thank an Associate Editor and two anonymous referees, whose comments and suggestions have greatly improved the paper, as well as audiences at the 2018 Ottawa Microeconomic Theory Conference, the 2019 Conference on Economic Design in Budapest, and the 20th ACM Conference on Economics and Computation (EC19) in Phoenix.

1 Introduction

When designing mechanisms for allocating resources, such as in auctions, matching, or other assignment problems, there is a long and rich literature studying strategy-proof direct mechanisms, which are often seen as desirable because an agent need not need forecast what they expect others to do in order to determine their own optimal strategy.[1] Indeed, strategy-proof direct mechanisms have played a large role in many practical market design applications, including auctions, school choice (Abdulkadiroğlu and Sönmez, 2003), medical residency matching (Roth and Peranson, 1999), and kidney exchange (Roth et al., 2004), among others. At the same time, imposing strategy-proofness can be costly, and allowing for non-strategy-proof (or, manipulable) mechanisms widens the space of possibilities. While some agents may benefit by lying in a manipulable mechanism, the ease of recognizing and enacting such manipulations may vary across mechanisms. By using mechanisms that, while not strategy-proof, are not easy to manipulate, designers may be able to improve outcomes on other important dimensions, such as efficiency. The goal of this paper is to provide a simple and tractable method for determining when a mechanism is easy to manipulate.

To motivate our project, consider two widely-used, manipulable mechanisms. The first is the Boston mechanism for school choice. Under this mechanism, a student loses her priority at a school unless she ranks it first. Therefore, 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.

[1] This literature goes back to at least Vickrey (1961), who writes that in a second-price auction "Each bidder can confine his efforts and attention to an appraisal of the value the article would have in his own hands, at a considerable saving in mental strain and possibly in out-of-pocket expense".

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, one of the primary objections to the Boston mechanism is the ability of strategic students, who recognize the potential manipulations, to profit at the expense of non-strategic students, who just report truthfully (Pathak and Sönmez, 2008). This has been a leading factor in the abandonment of the mechanism in some jurisdictions.[2]

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 may backfire: the manipulating hospital may not be assigned a doctor it would be happy to employ. This is in stark contrast to the Boston mechanism, where a student can guarantee a spot at her second-choice school, and thereby surely avoid a potentially worse outcome from reporting truthfully.

These examples 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" 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*.

For a given agent, a report θ' is a *manipulation* if the agent ever does strictly better reporting θ' over reporting her true type, θ. In this case, truthful reporting cannot be a dominant strategy. We define θ' to be an *obvious manipulation* if either the best possible outcome under θ' is strictly better than

[2]Though the use of the Boston mechanism has been abandoned in some places (including its namesake city and a total legislative ban in England), it still remains one of the most popular assignment mechanisms overall. Since so many school districts use an "obviously" manipulable mechanism, one might wonder whether the degree of manipulability is an important consideration for school districts. Pathak and Sönmez (2013) provide an extensive discussion on this issue, as well as a comprehensive list of authorities that have used (and abandoned) such mechanisms, past and present.

the best possible outcome under θ, or the worst possible outcome under θ' is strictly better than the worst possible outcome under θ. Clearly an obvious manipulation is also a manipulation; however, we argue that an obvious manipulation is identifiable to agents in a way that non-obvious manipulations are not.[3]

To formalize the idea that obvious manipulations are easier to identify, we consider an agent who is not fully informed (or does not fully understand) how a mechanism ϕ is defined, but instead is only able to determine the set of possible outcomes from any given strategy; mathematically, she knows the range of ϕ conditional on her own report, but not the full function itself, state-by-state. For example, in the context of school assignment, this could be a neighborhood parent group that does not fully understand (or has not been told) the assignment algorithm being run but has kept track of what preferences parents have submitted and what the resulting assignments were. Theorem 1 demonstrates that obvious manipulations are exactly the manipulations that can be identified by such an agent. This is our theoretical foundation of the term "obvious": even an agent who does not fully know how the mechanism is defined can deduce that the mechanism can be manipulated.

Both our formal definition and our behavioral characterization are inspired by the influential paper of Li (2017) on *obvious strategy-proofness (OSP)*. Li (2017) starts from the observation that real-world agents are often unable to engage in the intricate, contingent reasoning necessary to fully understand the implications of a given course of action on a state-by-state basis (mathematically, in our context this would be equivalent to knowing the entire function ϕ).[4] Formally, Li (2017) also considers agents who know only the set of possible outcomes from any given strategy, which can be understood as either a lack of ability to contingently reason, or equivalently as agents who are given only a partial description of the mechanism. Obviously dominant strategies are then those that are recognizable as dominant by such agents. While robust when they exist, very few mechanisms will have obviously dominant strategies; indeed, almost no normal-form games will be obviously strategy-proof.

[3]Implicit in our construction is the assumption that truthfully reporting your type is a focal strategy for an agent. Focal strategies trace back to Schelling (1980); and there is both experimental (e.g., Featherstone and Niederle, 2016; Pais and Pintér, 2008) and theoretical (e.g., Pathak and Sönmez, 2008; Bochet and Tumennasan, 2017; Dutta and Sen, 2012; Baillon, 2017) support for truth-telling as a focal strategy. For example, in interpreting the results of their school choice experiment, Featherstone and Niederle (2016) write "A plausible explanation is that truth-telling holds special sway as a focal strategy".

[4]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.

Many real-world applications like those we are concerned with (school choice, NRMP) have tens of thousands of agents, making it impractical to run an extensive-form (OSP) mechanism,[5] which motivates our restriction to direct mechanisms. Even in this context, strategy-proofness itself is limiting, and so, rather than strengthen it, our approach is to relax strategy-proofness and instead look for mechanisms that are not *obviously* manipulable.

After our behavioral characterization, we apply our definition to several canonical market design environments, starting with school choice. We first formalize the above discussion regarding the Boston mechanism and show it is indeed obviously manipulable (Proposition 1). The main alternative to the Boston mechanism, the (student-proposing) DA mechanism, is strategy-proof for the students, but may produce Pareto inefficient assignments. To correct this, many new mechanisms that Pareto improve on DA have been proposed. While it is known that any such mechanism 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 2). This has particularly important implications for the efficiency-adjusted deferred acceptance (EADA) mechanism of Kesten (2010), which has received renewed attention, as several recent 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: it is a manipulable mechanism. However, Theorem 2 implies that EADA is not obviously manipulable, and thus this may be less likely to be an issue in practice.

After presenting our results for school choice, we discuss several other canonical market design applications. For two-sided matching, we show that while DA is manipulable for the receiving side, it is not obviously so (Theorem 3). For multi-unit auctions, we show that first-price/pay-as-bid multi-unit auctions are obviously manipulable (Corollary 2), while the $(K+1)$-price auction is not (Theorem 4).[6] Finally, 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 de-

[5] Ashlagi and Gonczarowski (2018), Troyan (2019), 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] 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 $K > 1$.

sirable properties. Our last 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 stress 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 difficulty by classifying this mechanism as obviously manipulable.[7]

A common alternative approach to relaxing strategy-proofness in market design (without moving all of the way to Bayesian incentive compatibility) relies on large markets. 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. Azevedo and Budish (2018) define a related concept of strategy-proofness in the large (SPL). While similar in motivation, our approach is distinct in several respects. Most notably, we require no assumptions on how preferences are drawn or agent beliefs; further, our results hold for markets of any size, and not just in the limit.[8] 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 for some agent, the former is as well. We do not require any assumptions on cardinal preferences, nor do we attempt

[7]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) claim that ambiguity can be used to "solve" the impossibility of Myerson and Satterthwaite (1983) (see also Wolitzky, 2016), whereas our Theorem 5 reinforces Myerson and Satterthwaite's negative result.

[8]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 explain truthful reporting under DA.

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 require less information to recognize, 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, $I = \{i_1, \ldots, i_N\}$, and a finite set of outcomes, X. Agents have preferences over outcomes which we index by **types** $\theta_i \in \Theta_i$, where Θ_i is the set of possible types for agent i. The function $u_i(x; \theta_i)$ denotes agent i's utility for outcome x when his type is θ_i (note that values are private).[9] We focus on direct mechanisms. Letting $\Theta_I = \times_{i \in I} \Theta_i$, a (direct) **mechanism** is a function $\phi : \Theta_I \to X$ that maps type profiles to outcomes.[10] When convenient, we will use the notation x_i and $\phi_i(\theta)$ to denote i's individual allocation (e.g., in school choice, $\phi(\theta) = x$ is the entire assignment of all students to schools when the type profile is θ, while $\phi_i(\theta) = x_i$ is i's school and, with slight abuse of notation, we sometimes write $u_i(\phi_i(\theta); \theta_i) \equiv u_i(\phi(\theta); \theta_i)$ and $u_i(x_i; \theta_i) \equiv u_i(x; \theta_i)$ as i's utility for school $\phi_i(\theta) = x_i$ when of type θ_i).

[9] While we use utility function notation $u_i(\cdot; \cdot)$, this is only for presentation and readability. For all of our results (including the applications with transfers below), only ordinal preferences over outcomes are relevant, and the utility functions should not be interpreted as von Neumann-Morgenstern utilities (indeed, we think the lack of reliance on probability distributions is one of the advantages of our approach). Further, since each θ_i can be identified with an ordinal ranking over X and X is finite, Θ_i is finite as well.

[10] While our main ideas can also be applied more generally, the restriction to private values and direct mechanisms is an important class of problems motivated by the real-world applications we consider in the following sections where such mechanisms are commonly used, such as school choice (Abdulkadiroğlu and Sönmez, 2003; Pathak and Sönmez, 2013), hospital-resident matching (Roth and Peranson, 1999), and centralized college admissions (Balinski and Sönmez, 1999; Chen and Kesten, 2017), among others.

An important concern when choosing a mechanism is the incentives given to the agents to report their preferences truthfully. 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, 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. One approach is to assume, in addition to her payoff type, each agent also has a belief type, and uses this to evaluate her different options and choose the one that maximizes her (expected) utility.[11] However, extensive calculations of this type may be difficult for real word agents. In defining obvious dominance, for example, Li (2017) considers 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 in the original), and looks for mechanisms where all possibilities from one strategy are weakly better than all possibilities from any other. At the same time, calculating worst (or best) possible outcomes is typically much simpler than calculating all possible outcomes; further, 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 require the agents to engage in extensive contingent reasoning or to calculate expectations (and therefore is not sensitive to assumptions on the agents' beliefs).

[11] While not strictly necessary (see, e.g., Bergemann and Morris, 2005) applied game theory models often impose the further restriction that each agent's payoff type is drawn from Θ according to some prior distribution that is common knowledge.

Rather, we focus on comparing the best/worst cases, which we think are both simpler and particularly salient. In the next section, we will motivate this definition more formally by providing a characterization of obvious manipulations as those that can be recognized by cognitively limited agents.

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

If either (i) or (ii) does 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., $\max_{\theta_{-i}} u_i(\phi(\theta_i', \theta_{-i}); \theta_i) > \max_{\theta_{-i}} u_i(\phi(\theta_i, \theta_{-i}); \theta_i)$). If either (i) or (ii) is 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.

3 Characterization

In this section, we provide a characterization of obvious manipulations in the spirit of the characterization of obviously strategy-proof mechanisms presented in Li (2017).[12] Li (2017) considers an agent who is aware of the possible outcomes from her choices, but who is unable to engage in contingent reasoning. This agent is aware of the experiences that a mechanism will generate, and at each information set, knows the set of outcomes that can result from a strategy. If an agent cannot distinguish between two mechanisms when armed only with this information, Li (2017) defines the two mechanisms to be i-indistinguishable, and shows that an agent who cannot distinguish between i-indistinguishable mechanisms is able to determine that a strategy is weakly dominant if and only if the strategy is obviously dominant.

We consider the same cognitively limited agent as in Li (2017). As we work only with direct mechanisms, our definitions are correspondingly simpler. In

[12] We thank an anonymous referee for suggesting the analysis in this section.

particular, in a direct mechanism, the set of experiences the mechanism can generate is the range of possible outcomes. Recall that we use x_i to denote i's individual allocation under outcome x and $\phi_i(\theta)$ to be i's individual allocation under mechanism ϕ. Formally, given an agent i, strategy θ_i', and mechanism ϕ, we denote the range of possible outcomes from strategy θ_i' by $\pi_i^{\phi}(\theta_i') := \{x_i | \exists \theta_{-i} \text{ s.t. } \phi_i(\theta_i', \theta_{-i}) = x_i\}$. Mechanisms ϕ and ϕ' are **i-indistinguishable** if for every $\theta_i' \in \Theta_i$, $\pi_i^{\phi}(\theta_i') = \pi_i^{\phi'}(\theta_i')$.

Our first theorem shows that a mechanism ϕ has an obvious manipulation if and only if, for every mechanism ψ that is i-indistinguishable from ϕ, the corresponding manipulation is profitable. One interpretation of this theorem is that even cognitively limited agents who may not fully understand the mechanism they are playing will still be able to recognize manipulations if they are obvious. Another interpretation is that the set of obvious manipulations are exactly those that can be identified by agents who are only given partial information about the mechanism that will be run, in the sense that they know the range of possible outcomes from any given report. For instance, in a school choice context, parent groups may have historical data that keeps track of the preferences parents have submitted in previous years, and what their resulting assignments were. Such parents will be able to identify obvious manipulations, even without knowing (or fully understanding) exactly what mechanism is being used.

For our characterization, we impose a mild restriction on the model to avoid trivialities. The assumption we impose is a richness condition that the type spaces are "large enough". Formally, type space Θ is **rich** if for any agents i and j, $|\Theta_j| \geq |\{x_i | x \in X\}|$.[13] This condition is easily satisfied in the applications we consider in the next section. For example, in school assignment, each student has $(|S|+1)!$ possible types (each way of ranking each school and being unassigned is a distinct type), but for an individual agent, there are only $|S|+1$ possible allocations (i.e., schools she may be assigned).

Theorem 1. *Suppose there are at least three agents and the type space is rich. For any i, θ_i, θ_i', it holds that θ_i' is an obvious manipulation for θ_i under ϕ if and only if for every ψ that is i-indistinguishable from ϕ, θ_i' is a profitable manipulation for θ_i.*

The intuition behind the equivalence between an obvious manipulation and the manipulations that an agent with limited information can recognize is

[13] The RHS is the number of possible individual allocations for agent i. For instance, in school assignment, each agent may be assigned to one of $|S|$ possible schools or remain unassigned, and so $|\{x_i | x \in X\}| = |S|+1$ for all i.

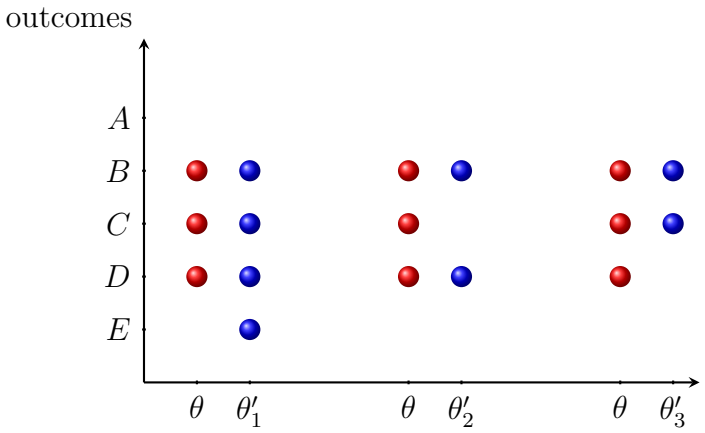

Figure 1: When, by observing just the range of outcomes, can we determine if one strategy weakly dominates another? Consider an agent who ranks outcomes $A \succ B \succ C \succ D \succ E$. For this agent, it is not possible to know if θ weakly dominates θ'_1; or if θ weakly dominates θ'_2 (the inputs that resulted in assignment C could now result in either B or D); however, we are certain that θ does not weakly dominate θ'_3.

provided in Figure 1. Here, an agent ranks the outcomes $A \succ B \succ C \succ D \succ E$, and the range of possible outcomes are given for various reports. In the first comparison, between θ and θ'_1, the agent knows that θ'_1 is sometimes worse than θ, but she cannot tell if it is sometimes better. When there is a "hole" in the range, such as when comparing θ to θ'_2, the agent cannot determine if that outcome has been replaced with something better or worse. Only in the last comparison, θ versus θ'_3, can the agent be certain that the alternative, θ'_3, is sometimes strictly better than θ. There is some state of the world where she receives D after submitting θ, and whenever this occurs, she does strictly better submitting θ'_3 instead of θ.

Proof of Theorem 1. Suppose θ'_i is an obvious manipulation for θ_i under mechanism ϕ. Denote i's best possible outcome under any mechanism ϕ' when reporting θ'_i by $B_i^{\phi'}(\theta'_i)$. We demonstrate the result for the case where $u_i(B_i^{\phi}(\theta'_i);\theta_i) > u_i(B_i^{\phi}(\theta_i);\theta_i)$ (in words, when her best possible outcome from submitting θ'_i is strictly preferred to her best possible outcome from submitting θ_i). Consider any i-indistinguishable mechanism ϕ', which is to say that $\pi_i^{\phi}(\theta'_i) = \pi_i^{\phi'}(\theta'_i)$ for all θ'_i (including θ_i). Note that in particular, this implies that $B_i^{\phi}(\theta_i) = B_i^{\phi'}(\theta_i)$ and $B_i^{\phi}(\theta'_i) = B_i^{\phi'}(\theta'_i)$. Since $\pi_i^{\phi}(\theta'_i) = \pi_i^{\phi'}(\theta'_i)$, there exists a θ_{-i} such that $\phi'_i(\theta'_i,\theta_{-i}) = B_i^{\phi}(\theta'_i) = B_i^{\phi'}(\theta'_i)$.

By definition, $\phi_i'(\theta) \in \pi_i^{\phi'}(\theta_i)$ and $u_i(B_i^{\phi'}(\theta_i);\theta_i) \geq u_i(\phi_i'(\theta);\theta_i)$. Therefore, $u_i(B_i^{\phi}(\theta_i');\theta_i) > u_i(B_i^{\phi}(\theta_i);\theta_i)$ (since θ_i' is an obvious manipulation under ϕ) and $B_i^{\phi}(\theta_i) = B_i^{\phi'}(\theta_i)$ and $B_i^{\phi}(\theta_i') = B_i^{\phi'}(\theta_i')$ (since ϕ and ϕ are i-indistinguishable), and so we have $u_i(B_i^{\phi'}(\theta_i');\theta_i) > u_i(B_i^{\phi'}(\theta_i);\theta_i) \geq u_i(\phi_i'(\theta);\theta_i)$, and it follows that $u_i(\phi_i'(\theta_i',\theta_{-i});\theta_i) > u_i(\phi_i'(\theta);\theta_i)$, i.e., θ_i' is a profitable manipulation for θ_i under ϕ'. A symmetric argument establishes that when the worst outcome (under ϕ) from θ_i' is strictly better than the worst outcome under θ_i, then θ_i' is a profitable manipulation of θ_i for every i-indistinguishable mechanism.

For the reverse direction, we prove the contrapositive. Fix a mechanism ϕ, an agent i, a type θ_i, and an alternative report θ_i' such that θ_i' is not an obvious manipulation of θ_i. We will construct a mechanism ϕ' that is i-indistinguishable from ϕ such that θ_i' is not a profitable manipulation of θ_i. First, for any $\theta_i'' \notin \{\theta_i, \theta_i'\}$ and any θ_{-i}, let $\phi'(\theta_i'';\theta_{-i}) = \phi(\theta_i'';\theta_{-i})$. Denote i's worst possible outcome under mechanism ϕ' from submitting θ_i' by $W_i^{\phi'}(\theta_i')$ and let $\bar{a} = B_i^{\phi}(\theta_i)$, $\bar{a}' = B_i^{\phi}(\theta_i')$, $\underline{a} = W_i^{\phi}(\theta_i)$, and $\underline{a}' = W_i^{\phi}(\theta_i')$. Since we have assumed that θ_i' is not an obvious manipulation of θ_i, $u_i(\bar{a};\theta_i) \geq u_i(\bar{a}';\theta_i)$ and $u_i(\underline{a};\theta_i) \geq u_i(\underline{a}';\theta_i)$.

For every $a \in \pi_i^{\phi}(\theta_i)$ fix a (distinct) θ_{-i}^a. For these values, define $\phi_i'(\theta_i, \theta_{-i}^a) = a$ and $\phi_i'(\theta_i', \theta_{-i}^a) = \underline{a}'$. Similarly, for every $a' \in \pi_i^{\phi}(\theta_i')$ fix a (distinct) $\theta_{-i}^{a'}$, such that each $\theta_{-i}^{a'}$ is distinct from all $\theta_{-i}^{a''}$ previously chosen (including those chosen for the set $\pi_i^{\phi}(\theta_i)$), and for these values, define $\phi_i'(\theta_i, \theta_{-i}^{a'}) = \bar{a}$ and $\phi_i'(\theta_i', \theta_{-i}^{a'}) = a'$. Our assumptions that there are at least three agents and a rich type space ensures that there are sufficiently many distinct profiles $\theta_{-i} \in \Theta_{-i}$ so that this procedure is well-defined. For every other profile θ_{-i}, define $\phi_i'(\theta_i, \theta_{-i}) = \overline{a}$ and $\phi_i'(\theta_i', \theta_{-i}) = \overline{a}'$.

We have constructed ϕ' so that $\pi_i^{\phi}(\theta_i) = \pi_i^{\phi'}(\theta_i)$ and $\pi_i^{\phi}(\theta_i') = \pi_i^{\phi'}(\theta_i')$. It is trivially true that for all other types θ_i'', $\pi_i^{\phi}(\theta_i'') = \pi_i^{\phi'}(\theta_i'')$. Therefore, ϕ and ϕ' are i-indistinguishable. Further, by construction, for every profile θ_{-i}, $u_i(\phi_i'(\theta_i, \theta_{-i});\theta_i) \geq u_i(\phi_i'(\theta_i', \theta_{-i});\theta_i)$. To see this, note that for any θ_{-i} such that $\theta_{-i} = \theta_{-i}^a$ for some $a \in \pi_i^{\phi}(\theta_i)$, we have $u_i(\phi_i'(\theta_i, \theta_{-i});\theta_i) = u_i(a;\theta_i) \geq u_i(\underline{a};\theta_i) \geq u_i(\underline{a}';\theta_i) = u_i(\phi_i'(\theta_i', \theta_{-i});\theta_i)$, where the inequalities come from the definition of $\underline{a}$ and $\underline{a}'$ and the fact that θ_i' is not an obvious manipulation of θ_i. Similarly, for any $\theta_{-i} = \theta_{-i}^{a'}$ for $a' \in \pi_i^{\phi}(\theta_i')$, we have $u_i(\phi_i'(\theta_i, \theta_{-i});\theta_i) = u_i(\overline{a};\theta_i) \geq u_i(\overline{a}';\theta_i) \geq u_i(a';\theta_i) = u_i(\phi_i'(\theta_i', \theta_{-i});\theta_i)$, where the inequalities again come from the definition of $\overline{a}$ and $\overline{a}'$ and the fact that θ_i' is not an obivous manipulation of θ_i. For any other θ_{-i}, $\phi_i'(\theta_i;\theta_{-i}) = \overline{a}$ and $\phi_i'(\theta_i';\theta_{-i}) = \overline{a}'$, and so, by definition of $\overline{a}$ and $\overline{a}'$ and the fact that θ_i' is not an obvious manipulation of θ_i, we have $u_i(\phi_i'(\theta_i;\theta_{-i});\theta_i) \geq u_i(\phi_i'(\theta_i';\theta_{-i});\theta_i)$. Therefore, $u_i(\phi_i'(\theta_i, \theta_{-i});\theta_i) \geq u_i(\phi_i'(\theta_i', \theta_{-i});\theta_i)$ for all θ_{-i}, or, in other words, θ_i' is not a

profitable manipulation of θ_i in mechanism ϕ'. $\square$

4 Applications

In this section, we apply the ideas introduced above to several important market design environments, including school choice, two-sided matching, auctions, and bilateral trade. In these environments, we use the definition of an obvious manipulation to classify commonly-used, non-strategy-proof mechanisms as either obviously manipulable (e.g., the Boston mechanism, pay-as-bid auctions) or not obviously manipulable (e.g., school-proposing DA, uniform price auctions).

4.1 School Choice

We begin by considering a canonical school choice model, 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 $I \cup \{\emptyset\}$, where $\emptyset$ is interpreted as remaining unmatched (or taking some outside option). 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.

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, it is standard notation to denote an agent's type as P_i, where P_i is agent i's strict ordinal preferences over individual schools in the set $S \cup \{\emptyset\}$. To be consistent with this literature, in this section, rather than use utility functions indexed by types θ_i, we 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,[14] 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. The school priorities and capacities are public information and are known to all of the students.

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

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

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 mechanism 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) $B_i^{\phi}(P_i')\ P_i\ B_i^{\phi}(P_i)$. 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.[15] Since neither mechanism is strategy-proof, there are situations for each mechanism where a student may benefit from misreporting. 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,

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

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 $W_i^{BM}(P_i')P_iW_i^{BM}(P_i)$. Therefore, P_i' is an obvious manipulation.

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).[16] 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 the sense of Pathak and Sönmez (2013), 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

[16] 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 2 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,[17] 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 (see the Introduction). 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 in which such a deviation will be profitable, yet it 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 $i \succ_s j$.[18] Further, say that matching μ **Pareto dominates** matching μ' if $\mu_i R_i \mu_i'$ for all $i \in I$ and $\mu_i P_i \mu_i'$ for some $i \in I$. A matching μ is

[17] 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.

[18] 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 prior literature, 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.

said to be to be **stable-dominating** if it is stable or Pareto dominates some stable assignment. 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.

Why might one be interested in the class of stable-dominating mechanisms? In school choice settings, stability is usually interpreted as a fairness constraint: a priority is a "right" to a seat at a school, and if a student with lower priority is assigned to a school that i desires, then i has a right to protest the allocation, perhaps by taking legal action (see, e.g., Balinski and Sönmez (1999) and Abdulkadiroğlu and Sönmez (2003)). While desirable, a drawback of stability is that it is incompatible with Pareto efficiency; indeed, the student-proposing DA mechanism, which produces the student-optimal stable assignment (an assignment that Pareto dominates every other stable assignment), may still be Pareto inefficient. Because of this impossibility, there has been recent work looking at weakenings of stability that are normatively justified and also compatible with efficiency. They include partial fairness (Dur et al., 2015), legality (Ehlers and Morrill, 2017), essential stability (Troyan et al., 2018), and weak stability (Tang and Zhang, 2017).

Indeed, this is more than just a theoretical consideration. Using data from New York City, Abdulkadiroğlu et al. (2009) conduct an exercise in which they start from the student-optimal stable assignment and Pareto improve it using Gale's top trading cycles. They find that over 7% of eighth graders in their sample could be matched to schools they strictly prefer to their DA assignment without making anyone strictly worse off (though of course stability may be violated at the new assignment). The procedure Abdulkadiroğlu et al. (2009) use for their exercise is one particular example of a stable-dominating mechanism; there are of course many others, and because of the potential for significant efficiency gains, a growing literature has recently begun exploring the class of stable-dominating mechanisms more fully. This literature includes Kesten (2010), who introduces the efficiency-adjusted DA (EADA) mechanism; Dur et al. (2015), who introduce the top priority rule; Alcalde and Romero-Medina (2017), who analyze the deferred acceptance plus top trading cycles mechanism; and Ehlers and Morrill (2017), who generalize Kesten's EADA mechanism to allow for a larger class of choice functions on the school side.[19] While the allocations produced by these mechanisms satisfy many nice

[19]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, this algorithm is equivalent to DA. Other stable-dominating mechanisms include the school-proposing deferred acceptance mechanism (Gale and Shapley, 1962) and deferred acceptance with compensation chains (Dworczak, 2016).

properties that are explored in the aforementioned papers (most importantly, efficiency), they all suffer from the same shortcoming with regard to implementation: none are strategy-proof. This follows from a general impossibility result of Alva and Manjunath (2017), who show that the only strategy-proof and stable-dominating mechanism is the student-proposing Deferred Acceptance mechanism, which, as already discussed, is not efficient.[20] Our Theorem 2, which we state next, sheds a new light on this problem: while any stable-dominating mechanism will be manipulable, none of the manipulations will be obvious. This theorem covers all of the efficient mechanisms discussed in this paragraph (and others), since they are all stable-dominating.[21]

Theorem 2. *Any stable-dominating mechanism is not obviously manipulable.*

Proof of Theorem 2. We prove Theorem 2 using a series of lemmas. We present and prove these lemmas explicitly, as they may be of independent interest. 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 mechanism. These results themselves, as well as Theorem 2, 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,

[20] See Abdulkadiroğlu et al. (2009) and Kesten (2010) for related impossibility results on strategy-proof Pareto-improvements of student-proposing DA.

[21] Note that Kesten's EADA mechanism (and its generalization due to Ehlers and Morrill (2017)) in particular will be Pareto efficient, stable-dominating, and satisfies all of the weaker stability definitions cited in the previous paragraph, further strengthening the argument that its only drawback is lack of strategy-proofness. These results, combined with Theorem 2, give strong theoretical support for this 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 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.[22] Otherwise, let s be the school i ranks just above $\bar{w}$ (that is, there is no s' such that $s\ P_i\ s'\ P_i\ \bar{w}$). 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}, $\phi_i(P_i, P'_{-i})\ R_i\ schDA_i(P_i, P'_{-i})\ R_i\ \bar{w}$;[23] 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. Since $schDA(P_i, P'_{-i})$ is stable and Pareto-efficient, it is the student-optimal stable matching, so the lattice of stable matchings is a singleton. Thus, if $\phi(P_i, P'_{-i})$ Pareto dominates any stable matching, then it Pareto dominates $schDA(P_i, P'_{-i})$, which contradicts the efficiency of $schDA(P_i, P_{-i})$. Thus $\phi(P_i, P'_{-i})$ is stable, and equal to $schDA(P_i, P'_{-i})$, and in particular, $\phi_i(P_i, P'_{-i}) = schDA_i(P_i, P'_{-i}) = \bar{w}$.

□

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

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

[23] 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.

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 $D_i(s') = \{j \in I : j \succ_{s'} i\}$. 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 $S' \subseteq S$ such that $s'\ R_i\ s$ for all $s' \in S'$ and*

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

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 its capacity (or else μ is not stable). Therefore, by Equation 2, there must exist a school $s' \in S'$ and a student $j \notin D_i(s')$ such that $\mu_j = s'$. But $i \succ_{s'} j$ (by the definition of $D_i(s')$) 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:

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

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 $\left\{v_{s'}^1, \ldots, v_{s'}^{q_{s'}}\right\}$ 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 $v_{s'}^k$ (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.,

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

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

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

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

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

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.[24]

Lemma 3. *Let ϕ be a stable-dominating 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).[25] 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

[24]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).

[25]Note that $DA(\cdot)$ always refers to the student-proposing version of deferred acceptance (we use $schDA(\cdot)$ to refer to the school-proposing version). Also note, though, that $\bar{w}$ is the worst-case under both versions (as well as under any stable-dominating mechanism), by Lemma 1.

DA, the worst-case under P_i ($\bar{w}$) is better than the worst case under P_i' ($\emptyset$). Therefore, P_i' is not an obvious manipulation.

Finally, suppose instead that P_i' truncates i's preferences after $\bar{w}$. Let P_{-i} be a preference profile such that $DA_i(P) = \bar{w}$. Consider an alternative profile $\hat{P}_{-i}$ where each $j \neq i$ ranks $DA_j(P)$ first and the other schools arbitrarily. By construction, under profile $(P_i', \hat{P}_{-i})$, there is a unique stable assignment; this stable assignment is Pareto efficient; and under this assignment, i is assigned to $\bar{w}$. Since ϕ is a stable dominating assignment, $\phi_i(P_i', \hat{P}_{-i}) = \bar{w}$. Therefore, i's worst possible assignment from reporting P_i' is either $\bar{w}$ or else a worse school. Therefore, P_i' is not an obvious manipulation.

□

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 2 are those that do alter the relative ordering of some schools. Following Maskin (1999), we say that P_i' is a **non-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-monotonic transformation relative to $\bar{w}$, her worst possible assignment under ϕ.

Lemma 4. *Consider any stable-dominating mechanism ϕ. Let $\bar{w}$ be i's worst possible assignment under preferences P_i. Any non-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-monotonic manipulation P_i', i.e. a P_i' such that there exists some $s \in S$ such that $sP_i'\bar{w}$, but $\bar{w}P_is$. 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.:

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

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 2. Let ϕ be a stable-dominating 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 $aP_i'\bar{w}$, we have $aP_i\bar{w}$.

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

□

4.2 Two-sided matching

Two-sided matching is closely related to school choice, and the results in this section will be immediate corollaries from the results above. The first model of two-sided matching appeared in the seminal paper of Gale and Shapley (1962), where the two sides consist of men and women. For convenience, we follow this classic literature and partition the set of agents as $I = M \cup W$, where M is a set of "men" and W is a set of "women". Such models are also often used in other contexts, such as college admissions (also discussed in Gale and Shapley (1962)), where the two sides are relabeled students and colleges, or labor markets, where the two sides are relabeled as workers and firms.

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** 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 above. 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.

Theorem 3. *Any stable 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 2. 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 2.

□

4.3 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 $u_i(x;\theta_i)$.[26] 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 $u_i((y,t);\theta_i) = \mathbf{1}_{\{y_i=1\}}\theta_i - t_i$. 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.[27]

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: $\min_{\theta_{-i}} u_i(\phi^{FP}(\theta_i,\theta_{-i});\theta_i) =$

[26]Note that these utility functions need not be given cardinal interpretations, i.e., we only assume that agents have ordinal preferences over allocations that are increasing in money. Also, in this subsection and the next, to be consistent with much of the auctions literature, we allow for continuum outcome/type spaces. The fundamental analysis would be unchanged if we assumed only a finite space of possible transfers.

[27]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: $\max_{\theta_{-i}} u_i(\phi^{FP}(\theta_i', \theta_{-i}); \theta_i) = \theta_i - \theta_i' > 0 = \max_{\theta_{-i}} u_i(\phi^{FP}(\theta_i', \theta_{-i}); \theta_i)$, and thus, according to Definition 2, reporting θ_i' is an obvious manipulation.

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 $u_i((y, t); \theta_i) = \sum_{\ell=1}^{y_i} \theta_i^\ell - t_i$.

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.*

In a $(K+1)-$price auction, 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]

[28]In the specific context of Treasury auctions, Friedman (1960) 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

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)-(ii) 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.

□

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:

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

(2018) for complementary analyses). Ausubel et al. (2014) compares the efficiency and revenue properties of pay-as-bid and uniform price auctions in Bayes-Nash equilibrium.

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. 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:[29]

[29]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,

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:

$$t_S(\theta_B, \theta_S) \geq \theta_S \text{ for all } \theta_B \geq \theta_S \tag{6}$$
$$t_B(\theta_B, \theta_S) \leq \theta_B \text{ for all } (\theta_B, \theta_S). \tag{7}$$

(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 (6) and (7) 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

and so an ex-post formulation of individual rationality is more appropriate for our setting.

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. At the same time, strategy-proofness is a constraint that limits the choice of mechanisms, and so may hinder performance in some dimensions, such as efficiency or revenue.

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? 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. While it has not yet been used (to our knowledge) in practice, the many desirable features of this mechanism outlined in other work, combined with the fact that EADA is not obviously manipulable, suggests that it may be worthy of further investigation.

Our paper opens up several avenues for further investigation. For instance, we defined an obvious manipulation with respect to truthful reporting as the default focal strategy. However, it is possible to generalize the idea to compare any two strategies, and look for an equilibrium in 'no obvious deviations'.

Additionally, we restricted attention in this paper to direct revelation mechanisms. While this this an important class in its own right, with many market design applications, it would nevertheless be interesting to explore indirect mechanisms as well. For instance, which mechanisms in this broader class can be identified as manipulable by cognitively limited agents? If such a mechanism has no obvious deviations, is there a corresponding NOM direct mechanism? These are interesting questions for future work.

References

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," Mimeo, Universitat Autònoma de Barcelona.

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. Cramton, M. Pycia, M. Rostek, and M. Weretka (2014): "Demand reduction and inefficiency in multi-unit auctions," *The Review of Economic Studies*, 81, 1366–1400.

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*.

Baillon, A. (2017): "Bayesian markets to elicit private information," *Proceedings of the National Academy of Sciences*, 114, 7958–7962.

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.

Bergemann, D. and S. Morris (2005): "Robust Mechanism Design," *Econometrica*, 73, 1771–1813.

Bochet, O. and N. Tumennasan (2017): "One truth and a thousand lies: Focal points in mechanism design," *Available at SSRN 3002539*.

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.

Chen, Y. and O. Kesten (2017): "Chinese college admissions and school choice reforms: A theoretical analysis," *Journal of Political Economy*, 125, 99–139.

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.

Dutta, B. and A. Sen (2012): "Nash implementation with partially honest individuals," *Games and Economic Behavior*, 74, 154–169.

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.

Featherstone, C. R. and M. Niederle (2016): "Boston versus deferred acceptance in an interim setting: An experimental investigation," *Games and Economic Behavior*, 100, 353–375.

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

Friedman, M. (1960): *A program for monetary stability*, vol. 3, Fordham University Press New York.

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.

Pais, J. and A. Pintér (2008): "School Choice and Information: An Experimental Study on Matching Mechanisms," *Games and Economic Behavior*,

64, 303–328.

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., T. Sönmez, and M. U. Ünver (2004): "Kidney Exchange," *Quarterly Journal of Economics*, 119, 457–488.

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.

Schelling, T. C. (1980): *The Strategy of Conflict*, Harvard University Press.

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. (2019): "Obviously Strategy-Proof Implemenation of Top Trading Cycles," *International Economic Review*, 60.

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.

A Definition of the Mechanisms

In this appendix, we give formal definitions of the matching mechanisms analyzed in Section 3.

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.[30] 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

[30] 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 .

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 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.