Vanderbilt University Department of Economics Working Paper Series
Nashville, TN 37235-1819

Vanderbilt University Department of Economics Working Papers 16-00019

Deferred acceptance is minimally manipulable

Martin Van der linden
Department of Economics, Vanderbilt University

Abstract

This paper shows that the deferred acceptance mechanism (DA) cannot be improved upon in terms of manipulability in the sense of either Pathak and Sönmez (2013) or or Arribillaga and Massó (2015) without compromising with stability.A conflict between manipulability and fairness is also identified. It is shown that miniworst stable mechanisms that make the set of individuals who match with their worst achievable mate minimal are maximally manipulable among the stable mechanisms. Miniworst mechanisms are also more manipulable than DA in the sense of Arribillaga and Massó (2015). A similar conflict between fairness and manipulability is identified in the case of the median stable mechanism (Teo and Sethuraman, 1998).

I am grateful to Tommy Andersson, Benoit Decerf, Greg Leo, Jordi Massó and John Weymark for helpful discussions and comments. I also want to thank Paul Edelman, Eun Jeong Heo, and Myrna Wooders for useful remarks. Special thanks go to Patrick Prosser and his java code for finding stable matchings (url{http://www.dcs.gla.ac.uk/~pat/roommates/distribution/}), which has been helpful in constructing some of the examples in this paper. This work is partially supported by National Science Foundation grant IIS-1526860.
Citation: Martin Van der linden, (2016) "Deferred acceptance is minimally manipulable", *Vanderbilt University Department of Economics Working Papers*, VUECON-16-00019.
Contact: Martin Van der linden - martin.van.der.linden@vanderbilt.edu.
Submitted: September 19, 2016. **Published:** September 21, 2016.
URL:http://www.accessecon.com/Pubs/VUECON/VUECON-16-00019.pdf

1 Introduction

In matching problems, the celebrated deferred-acceptance mechanism (DA) has been found to have many interesting properties. Importantly, the outcome of DA is *stable* (Gale and Shapley, 1962) meaning that no two individuals prefer each other to the individual they match with under DA and that DA is individually rational. The main issue with DA is that it does not give all individuals the incentives to reveal their preferences truthfully. In fact, this is a general feature of stable mechanisms : no stable mechanism makes it a dominant strategy for all individuals to reveal their preferences truthfully (Roth, 1982).

However, the question remains: does any mechanism perform better than DA in terms of incentives without compromising on stability? In other words, is any stable mechanism less manipulable than DA in some meaningful sense? In this paper, I answer negatively by showing that, in one-to-one matching DA is minimally manipulable in the senses of both Pathak and Sönmez (2013) (PS) and Arribillaga and Massó (2015) (AM) (extensions to many-to-one matching are discussed in Section 7).[1]

In the last two decades, much effort has been dedicated to the development of criteria that enable comparing the manipulability of mechanisms that, like the stable matching mechanisms, fail to have a truthful dominant strategy.[2] In this paper, I use two criteria from PS and AM to compare the manipulability of DA with that of other stable mechanisms. Informally, these criteria compare mechanisms based on their sets of manipulable profiles (PS) and preferences (AM). If for every profile (preference), the set of individuals who can manipulate in mechanism A is a subset of the set of individuals who can manipulate in mechanism B and that subset is strict for some profiles (preferences), then A is said to be *less manipulable than* B.

I show that no stable mechanism is less manipulable than DA in the sense of either PS or AM (Proposition 4.(i)).[3] In this sense, DA is *minimally* manipulable among the stable mechanisms. This property is rare among stable mechanisms: contrary to DA, most stable mechanisms are dominated by another stable mechanism in terms of manipulability (Proposition 5).

[1] Pathak and Sönmez (2013) introduce several comparison criteria. The criterion I use is the criterion introduced in Section III in Pathak and Sönmez (2013).

[2] Beside the two papers already cited, see Aleskerov and Kurbanov (1999), Maus et al. (2007), Andersson et al. (2014), Fujinaka and Wakayama (2012), Gerber and Barberà (2016) and Decerf and Van der Linden (2016), among others.

[3] In the process, an elegant characterization of the PS's comparison criterion is obtained for the one-to-one environment (Proposition 2). See also Section 7 for an extension of this characterization to a many-to-one environment.

DA is not only minimally manipulable, it is also less manipulable than other stable mechanisms (Proposition 4.(ii)). Although the mechanisms that DA dominates are rare when comparisons are in the sense of PS, they are abundant when comparisons are in the sense of AM (Proposition 6).

The above results follow mainly from the fact that, in a stable mechanism, if an individual cannot manipulate given some profile or preferences, then the individual must match with her or his most preferred achievable mate (Proposition 1, where an achievable mate is a mate that the individual matches with under *some* stable matching). As is well-known, when one individual matches with her or his most preferred achievable mate, this individual is also the least preferred achievable match for her or his mate (Gale and Shapley, 1962). This point toward a tension between fairness and manipulability: to reduce manipulability one must give some individuals the best match they could hope for, which implies others receive their worst possible match.

To achieve minimal manipulability, DA pushes this logic to the extreme and always matches one side of the market with their most preferred achievable mate. This is sometimes viewed as a downside of DA and fairer stable mechanisms have been devised that select "intermediate" stable matchings in which fewer individuals match with their least preferred achievable mate (see e.g. Irving et al. (1987) and Teo and Sethuraman (1998)).

I show that among stable mechanisms, such improvements in fairness come at the cost of an increase in manipulability. A minimal fairness criterion that I call *miniworst* requires that the set of individuals who match with their worst achievable mate be minimal (with respect to inclusion) As it turns out, if a stable mechanism is miniworst, then it is maximally manipulable, i.e., no other stable mechanism is strictly more manipulable in the sense of either PS or AM (Proposition 8). In fact, all miniworst mechanisms are dominated by DA in the sense of AM (Proposition 9), although this is not true in the sense of PS.

A similar trade-off between manipulability and fairness is identified in the case of the median stable mechanisms (Teo and Sethuraman, 1998): the median stable mechanism (i) fails to be minimally manipulable in the sense of either PS or AM (Proposition 10) and in the sense of AM, median stable mechanisms are (ii) more manipulable than DA and (ii) maximally manipulable (Propositions 12 and 11); although the same is again not true in the sense of PS.

Related literature

The results in this paper complement previous results of PS. Of the results

in this paper, only the fact that no stable mechanism is less manipulable than DA in the sense of PS can be obtained as a direct corollary of a result of PS. In particular, the manipulability comparisons in the sense of AM are new.

This paper contrasts with Chen et al. (2016) who show that no two stable mechanisms can be compared in the sense of a stronger comparison criterion also proposed by PS. Differently, I show that many stable mechanisms can be compared in the sense of both AM and the weaker comparison criterion of PS.

This paper also complements the literature on fair stable matchings (Knuth, 1997; Irving et al., 1987; Teo and Sethuraman, 1998; Klaus and Klijn, 2006). To my knowledge, this paper is the first to clarify the costs in terms of manipulability of selecting a fair stable matching instead of the extreme stable matchings that DA selects.

The paper is organized as follows. Section 2 gives a general definition of the comparison criteria of PS and AM. Section 3 defines the one-to-one matching environment. Section 4 restates some famous results about the one-to-one environment and derives preliminary results. Section 5 compares the manipulability of DA with respect to the whole class of stable mechanisms. Section 6 focuses on comparing the manipulability of DA with two classes of fair stable mechanisms: the miniworst and the median stable mechanisms. Section 7 discusses extensions of some results from Section 5 to many-to-one matching. I conclude with open questions.

2 Two criteria for manipulability comparisons

Let $N \coloneqq \{1, \dots, n\}$ be the set of individuals and T the set of outcomes. An individual $i \in N$ has a preference R_i over the outcomes in T. For any $s, t \in T$, $s\ R_i\ t$ indicates a weak preference for s over t and $s\ P_i\ t$ a strict preference ($s\ R_i\ t$ but not $t\ R_i\ s$). For any $i \in N$, the domain of i's possible preferences is $\mathcal{D}_i$.

A preference profile $R \coloneqq (R_1, \dots, R_n)$ is a list of the preferences of all the individuals in N. The set of preference profiles is $\mathcal{D} \coloneqq \times_{i \in N} \mathcal{D}_i$. The list of preferences in R for everyone but i is $R_{-i} \in \mathcal{D}_{-i} \coloneqq \times_{i \in N \setminus \{i\}} \mathcal{D}_i$. A pair $(T, \mathcal{D})$ is called an **environment**.

A **mechanism** A is a function that associates every preference profile $R \in \mathcal{D}$ with an outcome $A(R) \in T$. When participating in a mechanism, individual i will naturally wonder whether reporting her true preferences is a good strategy, or whether she would be better-off manipulating her report. To do so, i will have to form some belief about the preferences R_{-i} that other individuals will report.

Suppose that i forms a *point* belief about R_{-i}. Then the comparison criterion from PS (PS-criterion) says that mechanism A is no more manipulable than mechanism B if for any $i \in N$, when i's belief about R_{-i} is the same in A and i does not find it profitable to manipulate in B given this belief, i does not find it profitable to manipulate in A either. Formally, mechanism A is **PS-manipulable for i given $R \in \mathcal{D}$** if

$$A(R_i', R_{-i}) \; P_i \; A(R_i, R_{-i}) \qquad \text{for some } R_i' \in \mathcal{D}_i. \tag{1}$$

That is, i can benefit from manipulating her reported preferences when other individuals report R_{-i}.

Mechanism A is **no more PS-manipulable than** mechanism B if

$$\begin{aligned} &\{i \in N \mid A \text{ is PS-manipulable for } i \text{ given } R\} \\ &\subseteq \{i \in N \mid B \text{ is PS-manipulable for } i \text{ given } R\} \quad \text{for } \textit{all } R \in \mathcal{D}. \end{aligned} \tag{2}$$

Similarly, mechanism A is **less PS-manipulable than** mechanism B if A is no more PS-manipulable than B but the converse is not true, i.e., (2) holds and

$$\begin{aligned} &\{i \in N \mid A \text{ is PS-manipulable for } i \text{ given } R^*\} \\ &\subset \{i \in N \mid B \text{ is PS-manipulable for } i \text{ given } R^*\} \quad \text{for } \textit{some } R^* \in \mathcal{D}. \end{aligned} \tag{3}$$

The PS-criterion has a clear appeal when beliefs do not vary between the two mechanisms that are being compared. The PS-criterion is harder to make sense of when i's belief about R_{-i} varies between mechanisms A and B, which is likely whenever A and B differ sufficiently from one another.[4]

The criterion proposed by AM (AM-criterion) overcomes this limitation by comparing *preferences* at which individuals cannot manipulate *whatever the preferences reported by other individuals*. Formally, for any *preference* $R_* \in \cup_{i \in N} \mathcal{D}_i$ and any $i \in N$ with $R_* \in \mathcal{D}_i$, mechanisms A is **AM-manipulable for i given R_*** if

$$A(R_i', R_{-i}) \; P_* \; A(R_*, R_{-i}) \qquad \text{for some } R_i' \in \mathcal{D}_i \text{ and } \textit{some } R_{-i} \in \mathcal{D}_{-i}. \tag{4}$$

That is, given the preference R_*, i could benefit from manipulating her reported preference for *some* R_{-i}. In other words, i does not have truthful dominant strategy given preference R_*.

Mechanism A is **no more AM-manipulable than** mechanism B if for all $i \in N$ and all $R_i \in \mathcal{D}_i$, if A is manipulable for i given R_i, then B is also manipulable for i given R_i. To stress the parallel with (2), observe that this is equivalent to saying that A is no more AM-manipulable than mechanism

[4] For example, individuals are unlikely to form the same belief about the values reported by other individuals in a first-price and in a second-price auction.

B if

$$\begin{aligned}&\{i \in N \text{ with } R_* \in \mathcal{D}_i \mid A \text{ is AM-manipulable for } i \text{ given } R_*\}\\&\subseteq \{i \in N \text{ with } R_* \in \mathcal{D}_i \mid B \text{ is AM-manipulable for } i \text{ given } R_*\}\\&\text{for } \textit{all } R_* \in \cup_{i\in N}\mathcal{D}_i.\end{aligned} \tag{5}$$

In other words, every time i fails to have a truthful dominant strategy given R_* in A, i also fails to have a truthful dominant strategy given R_* in B. In the context of a two-sided matching model in which the individual domains do not intersect, the above sets are either singletons or empty. Similarly, mechanism A is **less AM-manipulable than** mechanism B if A is no more AM-manipulable than B but the converse is not true, i.e., (5) holds and

$$\begin{aligned}&\{i \in N \text{ with } R_{**} \in \mathcal{D}_i \mid A \text{ is AM-manipulable for } i \text{ given } R_{**}\}\\&\subset \{i \in N \text{ with } R_{**} \in \mathcal{D}_i \mid B \text{ is AM-manipulable for } i \text{ given } R_{**}\}\\&\text{for } \textit{some } R_{**} \in \cup_{i\in N}\mathcal{D}_i.\end{aligned} \tag{6}$$

Mechanism A is **less (no more) manipulable than** a mechanism B if A is *both* less (no more) PS-manipulable and less (no more) AM-manipulable than B. The "***no less* (PS-, AM-) manipulable than**" and "***more* (PS-, AM-) manipulable than**" partial orders are defined symmetrically.

Of fundamental importance for this paper are the concepts of minimal and maximal manipulability. Informally, a mechanism A is minimally manipulable if it is impossible to improve upon A in terms of manipulability. Specifically, for any class of mechanisms $\mathcal{A}$, mechanism $A \in \mathcal{A}$ is ***minimally* (PS-, AM-) manipulable in $\mathcal{A}$** if there is no mechanism $B \in \mathcal{A}$ such that B is *less* (PS-, AM-) manipulable than A. Conversely, mechanism $A \in \mathcal{A}$ is ***maximally* (PS-, AM-) manipulable** in $\mathcal{A}$ if there is no mechanism $B \in \mathcal{A}$ such that B is *more* (PS-, AM-) manipulable than A.

As AM explain, the "no more PS-manipulable than" partial order is a subset of the "no more AM-manipulable than" partial order: if A is no more PS-manipulable than B, then A is no more AM-manipulable than B too. However, the converse is not true. In particular, A can be less AM-manipulable than B although A fails to be less PS-manipulable than B (see, e.g., Proposition 9 below).

Both criteria have advantages and disadvantages over one another. As illustrated in Examples 1 and 2, there are situations in which the AM-criterion yields counter-intuitive manipulability comparisons while the PS-criterion does not, and situations in which the PS-criterion yields counter-intuitive manipulability comparisons while the AM-criterion does not. However, both criteria capture interesting aspect of the realtive manipulability of mechanisms, and so I consider them both in what follows.

Example 1. Suppose that no individual $i \in N$ with $R_* \in \mathcal{D}_i$ can AM-manipulate mechanism A given preference R_*. On the other hand, suppose that for every preference $R_{**} \neq R_*$ and every individual $i \in N$ with $R_{**} \in \mathcal{D}_i$, there exists an $R_{-i}(R_{**})$ such that i can PS-manipulate given $(R_{**}, R_{-i}(R_{**}))$, but with A, i cannot PS-manipulate given any profile $(R_{**}, R_{-i}) \neq (R_{**}, R_{-i}(R_{**}))$. Further suppose that, unlike mechanism A, mechanism B is AM-manipulable for *every* preference $R_\circ$, but for every $R_\circ$ and for every $i \in N$ with $R_\circ \in \mathcal{D}_i$, i can PS-manipulate given *one and only one* profile $(R_\circ, R_{-i}(R_\circ))$ when i's preferences are $R_\circ$ (i.e., B is not PS-manipulable for i given any profile $(R_\circ, R_{-i}) \neq (R_\circ, R_{-i}(R_\circ))$.

By construction, A is less AM-manipulable than B. This seems counter-intuitive because A improves upon B in terms of AM-manipulability for a single preference relation R_* and does much worse in terms of PS-manipulability for every other preference relation. In this sense, the judgment of the AM-criterion in the case of A and B is a "false positive". The PS-criterion is more intuitively appealing because it refrains from concluding that A is less manipulable than B (although the PS-criterion does not conclude that B is less manipulable than A).

Example 2. For some $j \in N$, suppose that the set of sub-profiles $\mathcal{D}_{-j}$ can be partitioned into two sets $\mathcal{D}^1_{-j}$ and $\mathcal{D}^2_{-j}$ of equal size ($\#\mathcal{D}^1_{-j} = \#\mathcal{D}^2_{-j}$).[5]

Suppose also that for every $i \in N\backslash\{j\}$ and every $R_i \in \mathcal{D}_i$, mechanism A is *not* AM-manipulable for i given R. Also, suppose that for every $R_j \in \mathcal{D}_j$, A is PS-manipulable for i given (R_i, R^1_{-j}) whenever $R^1_{-j} \in \mathcal{D}^1_{-j}$, but A is *not* PS-manipulable for i given (R_i, R^2_{-j}) whenever $R^2_{-j} \in \mathcal{D}^2_{-j}$. Further suppose that, unlike mechanism A, mechanism B is AM-manipulable for every $i \in N\backslash\{j\}$ given any preference $R_i \in \mathcal{D}_i$. Also, suppose that for every $R_j \in \mathcal{D}_j$, B is *not* PS-manipulable for i given (R_i, R^1_{-j}) whenever $R^1_{-j} \in \mathcal{D}^1_{-j}$, but A is PS-manipulable for i given (R_i, R^2_{-j}) whenever $R^2_{-j} \in \mathcal{D}^2_{-j}$.

For every $R \in \mathcal{D}$ with $R_{-j} \in \mathcal{D}^1_{-j}$, j can PS-manipulate in A but j cannot PS-manipulate in B. Also, for every $R \in \mathcal{D}$ with $R_{-j} \in \mathcal{D}^2_{-j}$, j can PS-manipulate in B but j cannot PS-manipulate in A. Hence, A and B are not comparable using the PS-criterion. This is counter-intuitive because A does much better than B in terms of AM-manipulability for every individual in $N\backslash j$ and A performs similarly to B in terms of PS-manipulability for j. In this sense, the judgment of the PS-criterion in the case of A and B is a "false negative" (or a "false incomparability"). The AM-criterion is intuitively more appealing because A is less AM-manipulable than B.

[5] This is feasible, for example, if for every $i \in N$ and every $R_i \in \mathcal{D}_i$, the preference R_i^{-1} such that $a\ R_i\ b$ if and only if $b\ R_i\ a$ is also an element of $\mathcal{D}_i$.

3 The one-to-one matching environment

All the formal results in this paper are for the one-to-one matching environment. In Section 7, I discuss extensions to many-to-one matching environments.

In the one-to-one matching environment (henceforth, the *one-to-one environment*), the set N is partitioned into a set W of women and a set M of men. Throughout, these sets have cardinalities $\#W, \#M \geq 3$. A woman $w \in W$ has a preference R_w on the set of men and herself ($M \cup \{w\}$) and a man $m \in M$ has a preference R_m on the set of women and himself ($W \cup \{m\}$). For any $i \in N$, the set of individuals for which i has a preference are i's potential **mates**. Henceforth, let the domain $\mathcal{D}$ consist of all possible profiles of strict preferences, i.e., for any $R \in \mathcal{D}$ and any $i \in N$, no two different mates are ever indifferent to one another according to R_i.

A **matching** is a function $\mu\colon W \cup M \to W \cup M$ that matches every individual $i \in N$ with a mate and for which nobody has more than one mate. Formally, (i) $\mu(w) \in M \cup \{w\}$ for all $w \in W$, (ii) $\mu(m) \in W \cup \{m\}$ for all $m \in M$, (iii) $\mu(w) \neq \mu(w')$ for all $w, w' \in W$ with $w \neq w'$, and (iv) $\mu(m) = w$ if and only if $\mu(w) = m$ for all $w \in W$ and $m \in M$. The set of outcomes T in the one-to-one environment is the set of matchings.

A **one-to-one matching mechanism** A (henceforth, a *mechanism*) associates every profile of preferences $R \in \mathcal{D}$ with a matching. To simplify the notation, let $A_i(R)$ and μ_i be i's mate with the matchings $A(R)$ and μ.

Given matching μ and profile R, a **blocking pair** consists of a man and a woman who prefer being matched together to their match in μ. Formally a blocking pair in μ is any (w, m) with $w \in W$ and $m \in M$ such that $m \; P_w \; \mu_w$ and $w \; P_m \; \mu_m$. For any $i \in N$, if $i \; P_i \; j$ for some $j \in N$, then mate j is **unacceptable** to i. A matching is **individually rational** if no individual matches with an unacceptable mate.

A **matching** is **stable** if it does not contain any blocking pairs *and* it is individually rational. A ***mechanism*** is **stable** if it selects a stable matching for every preference profile in $\mathcal{D}$.

Henceforth, the focus is on stable mechanisms. Therefore, and to keep the terminology simple, I suppress the reference to the class of stable mechanisms throughout. For example, when some mechanism A is said to be minimally manipulable, it should be understood that A is minimally manipulable *in the class of stable mechanisms*.

As is well-known, the *deferred acceptance* mechanism (DA) comes in two variants: women-proposing (DA^W) and men-proposing (DA^M), both of which are stable (Gale and Shapley, 1962). For any $i \in N$, the variant of DA in which i's side proposes is denoted DA^i. When a property applies

irrespective of the proposing side, a deferred acceptance mechanism is simply referred to as DA. In this case, an individual $i \in N$ is a **proposer** if i is on the side that proposes in DA (e.g., W in DA^W) and an **acceptor** if i is on the side that does not propose (e.g, M in DA^W). A typical proposer is denoted by $p \in N$, while a typical acceptor is denoted by $a \in N$.

For any $i \in N$, any individual $j \in N$ is an **achievable** mate given R if j matches with i under some stable matching (where stability is with respect to R). Individual i is **single** under μ if $\mu_i = i$. Individual i is **married** in μ if i is not single in μ ($\mu_i \neq i$).

For all $i \in N$ and all $R \in \mathcal{D}$, let f_i^R be i's most preferred achievable mate given R. Similarly, let ℓ_i^R be i's least preferred achievable mate given R. Observe that, because preferences are strict, f_i^R and ℓ_i^R are unique for all $i \in N$ and all $R \in \mathcal{D}$. Finally, for any $i \in N$, any $R_i \in \mathcal{D}_i$, and any acceptable mate x, $R_i|_x$ is the **truncation of** R_i **after** x, that is, $R_i|_x$ is the preference constructed from R_i by moving i up in the ranking to the point where i is ranked right after x, but not changing any other rankings.[6]

4 Preliminary results

In this section, I restate some classical properties of DA and of stable matchings in the one-to-one environment. These properties are then used to prove Proposition 1, which is central to the results in this paper.

First, DA always selects a stable matching in which proposers match with their most preferred achievable mate, while acceptors match with their least preferred achievable mate.

Lemma 1 (Gale and Shapley, 1962)**.** *For any $R \in \mathcal{D}$, (i) $DA(R)$ is stable with respect to R. (ii) $DA_p(R) = f_p^R$ for every proposer $p \in N$ and $DA_a(R) = \ell_a^R$ for every acceptor $a \in N$.*

Also, with DA it is a dominant strategy for proposers to report their true preferences.

Lemma 2 (Dubins and Freedman, 1981)**.** *For any proposer $p \in N$ and any $R \in \mathcal{D}$,*

$$DA_p(R_p, R_{-p}) \; R_p \; DA_p(R'_p, R_{-p}) \qquad \text{for all } R'_p \in \mathcal{D}_p.$$

These two lemmas imply that i's best achievable mate with the profile R is preferred according to R_i to i's best achievable mate when i reports a preference $R'_i \neq R_i$.

[6] Formally, (i) $x \; R_i \; i$, (ii) for all $y, z \neq i$, $y \; R_i|_x \; z$ if and only if $y \; R_i \; z$, and (iii) for all $z \neq i$, $z \; R_i|_x \; i$ if and only if $z \; R_i \; x$ and $i \; R_i|_x \; z$ if and only if $x \; R_i \; z$. .

Lemma 3. *For any $i \in N$, any $R \in \mathcal{D}$, and any $R'_i \in \mathcal{D}_i$, $f_i^R \; R_i \; f_i^{(R'_i, R_{-i})}$.*

Proof. By Lemma 2, $DA_i^i(R_i, R_{-i}) \; R_i \; DA_i^i(R'_i, R_{-i})$ for all $R'_i \in \mathcal{D}_i$ By Lemma 1, this is equivalent to $f_i^R \; R_i \; f_i^{(R'_i, R_{-i})}$ for all $R'_i \in \mathcal{D}_i$. ■

The following is another well-known result about the set of stable matchings.

Lemma 4 (Roth and Sotomayor, 1992)**.** *For a given $R \in \mathcal{D}$, the set of single individuals is the same in every stable matching.*

Together, the above lemmas can be used to show that, in a stable mechanism, i can manipulate given that other individuals report R_{-i} if and only if i is not matched with f_i^R.

Proposition 1. *For any stable mechanism A, any $i \in N$, and any $R \in \mathcal{D}$,*

$$A_i(R_i, R_{-i}) \; R_i \; A_i(R'_i, R_{-i}) \qquad \text{for all } R'_i \in \mathcal{D}_i \tag{7}$$

if and only if

$$A_i(R) = f_i^R. \tag{8}$$

Proof. **Sufficiency**. Because A is stable, by Lemma 3 and the definition of a most preferred achievable mate,

$$f_i^R \; R_i \; f_i^{(R'_i, R_{-i})} \; R_i \; A_i(R'_i, R_{-i}) \qquad \text{for all } R'_i \in \mathcal{D}_i. \tag{9}$$

Thus, (7) follows directly from (8).

Necessity. If $f_i^R = i$, then $A_i(R) = f_i^R$ because stable matching are individually rational. Thus, suppose that $f_i^R \; P_i \; i$ (this is the only other case to consider because $i \; P_i \; f_i^R$ is inconsistent with individual rationality). In order to derive a contradiction, suppose that $A_i(R) \neq f_i^R$. Because A is stable, this implies $f_i^R \; P_i \; A_i(R)$. There are two cases.

Case 1: $A_i(R_i|_{f_i^R}, R_{-i}) \; R_i \; f_i^R$. Then we have $A_i(R_i|_{f_i^R}, R_{-i}) \; R_i \; f_i^R \; P_i \; A_i(R)$, contradicting (7).

Case 2: $f_i^R \; P_i \; A_i(R_i|_{f_i^R}, R_{-i})$. Because A is individually rational and by the construction of $R_i|_{f_i^R}$, it follows that $A_i(R_i|_{f_i^R}, R_{-i}) = i$. Also, $DA_i^i(R) = f_i^R \; P_i \; A_i(R_i|_{f_i^R}, R_{-i})$ by Lemma 1). By the construction of $R_i|_{f_i^R}$, $DA^i(R)$ is therefore stable given $(R_i|_{f_i^R}, R_{-i})$. Indeed, because $DA^i(R)$ is stable with respect to R, i is not part of a blocking pair with any mate that i ranks above $DA_i^i(R)$ given R_i. But because $R_i|_{f_i^R}$ and R_i have the same ranking of mates up to $DA_i^i(R)$, this is also true given $R_i|_{f_i^R}$.

Thus, there exists a stable matching given $(R_i|_{f_i^R}, R_{-i})$ in which i is married $(DA^i(R))$ and another in which i is single $(A_i(R_i|_{f_i^R}, R_{-i}))$, contradicting Lemma 4. ■

Proposition 1 plays a pivotal role in most of the results in this paper.[7] The next section uses Proposition 1 and Lemma 1 to compare the manipulability properties of DA with those of other stable mechanisms.

5 Maximal and minimal manipulability among stable mechanisms

As in apparent from (2), a useful feature of the PS-criterion is that it can be decomposed into a separate comparison for each profile. This decomposability enables the use of Proposition 1 to obtain the following characterization of the "no more PS-manipulable than" partial order.

Proposition 2. *Stable mechanism A is no more PS-manipulable than stable mechanism B if and only if*

$$\{i \in N \mid A_i(R) = f_i^R\} \supseteq \{i \in N \mid B_i(R) = f_i^R\} \quad \text{for all } R \in \mathcal{D}. \tag{10}$$

Proof. Proposition 2 is a direct consequence of Proposition 1 and the contrapositive of the definition of the "no more PS-manipulable than" partial order in (2). ■

Proposition 2 is used to establish the following characterization of minimally and maximally PS-manipulable mechanisms. (Recall that every minimal and maximal manipulability property is defined implicitly with respect to the class of stable mechanisms.)

Proposition 3. *A mechanism A is minimally (resp. maximally) PS-manipulable if and only if, for every $R \in \mathcal{D}$, there does* not *exist a stable matching μ such that*

$$\{i \in N \mid A_i(R) = f_i^R\} \subset \{i \in N \mid \mu_i = f_i^R\} \tag{11}$$

$$(\textit{resp. } \{i \in N \mid A_i(R) = f_i^R\} \supset \{i \in N \mid \mu_i = f_i^R\}). \tag{12}$$

Next, I show that DA is minimally manipulable. (Recall that manipulability properties that do not refer to either PS of AM hold for both criteria). In the case of minimal PS-manipulability, this follows straightforwardly from Proposition 3 and Lemma 1: enlarging (with respect to inclusion) the set of

[7] The proof of Proposition 1 is inspired by the fact that every report of a preference is dominated by the report of a truncation, which was first proven by Roth and Vande Vate (1991, Theorem 2). The proof of Proposition 1 follows the same proof strategy as the proof of Roth and Vande Vate's theorem. This proof strategy is also used in the proof of Pathak and Sönmez (2013, Lemma 1). A similar result appears in Coles and Shorrer (2014).

acceptors who are matched with their most preferred achievable mate implies that the set of proposers who are matched with their most preferred achievable mates is shrunk.

An informative characterization of AM-minimally and maximally manipulable mechanisms is harder to obtain because the "no more AM-manipulable than" partial order is not decomposable: enlarging (with respect to inclusion) the set of individuals who cannot AM-manipulate given some preference R_* may have an impact on the set of individuals who cannot AM-manipulate given some other preference R_{**}.[8] Therefore, the proof for minimal AM-manipulability does not rely on a characterization similar to Proposition 3. Instead, the minimal AM-manipulability of DA is proven directly from Proposition 1 and Lemma 1.

Proposition 4. *(i) DA is minimally manipulable. (ii) There exists stable mechanisms that are more manipulable than DA.*

Proof of (ii).[9] Consider the following profile from Klaus and Klijn (2006):

$$\begin{array}{llll} R_{w_1}\colon & m_3 & m_2 & m_1 \\ R_{w_2}\colon & m_2 & m_1 & m_3 \\ R_{w_3}\colon & m_1 & m_3 & m_2 \end{array} \qquad \begin{array}{llll} R_{m_1}\colon & w_1 & w_2 & w_3 \\ R_{m_2}\colon & w_3 & w_1 & w_2 \\ R_{m_3}\colon & w_2 & w_3 & w_1 \end{array} \quad . \tag{13}$$

When being self-matched is omitted as in (13), being self-matched is implicitly the worst outcome. Given R, the stable matchings are

$$\begin{array}{llll} \mu^1\colon & m_1 & m_3 & m_2 \\ \mu^2\colon & m_2 & m_1 & m_3 \\ \mu^3\colon & m_3 & m_2 & m_1 \end{array} ,$$

where in μ^1, for example, w_1 matches with m_1, w_2 matches with m_3, and w_3 matches with m_2. Observe that $DA^W(R) = \mu^3$ (resp. $DA^M(R) = \mu^1$) and all the women (men) are matched with their most preferred mates in $DA^W(R)$ ($DA^M(R)$). The mechanism constructed from DA by only changing the stable matching selected for R to μ^2 is more manipulable than DA because in this case, nobody is matched with her or his most preferred achievable mate.

To see that more than one mechanism is more manipulable than DA, repeat the above argument for the variant of (13) in which m_1 appears on the *downward* diagonal of the womens' profile and w_1 appears on the *upward* diagonal of the mens' profile. If $\#M > 3$ or $\#W > 3$, simply consider the extension of (13) in which all individuals except for $w_1, w_2, w_3, m_1, m_2,$

[8] However, see the proof of Proposition 5.(iii) in the Appendix for hints at a characterization of AM-minimal and AM-maximal manipulability.

[9] The proof of part (i) may be found in the Appendix.

and m_3 prefer to be self-matched to being matched with any other potential partner. ■

Proposition 4 shows that DA cannot be improved upon in terms of manipulability without compromising on stability.[10]

As it turns out, minimal manipulability is a relatively uncommon property in the class of stable mechanisms. To illustrate, I focus on the case in which $\#W = \#M$ and the domain of preferences $\bar{\mathcal{D}} \subset \mathcal{D}$ for which each individual ranks being self-matched last. Let $h \coloneqq \frac{n}{2}$. For any h, any individual $i \in N$, and any preference $R_i \in \mathcal{D}_i$, it is possible to construct a subprofile $R_{-i}^{R_i}$ such that the profile $(R_i, R_{-i}^{R_i})$ mimics the "Latin Square" pattern of profile (13). Any of these Latin Square profiles admits h stable matchings. As in (13), out of these h stable matchings, minimally PS-manipulable mechanisms can only select either the men optimal or the women optimal matching. For all h, an upper bound on the proportion of minimally PS-manipulable mechanisms can therefore be obtained by considering the proportion of stable mechanisms that select one of these two stable matching in every Latin Square profile (see Proposition 5 below).

As for an upper bound on minimal AM-manipulability, if a mechanism A is minimally AM-manipulable, then DA cannot be less AM-manipulable than A. By Proposition 1 and Lemma 1, this implies that there exists an acceptor $a \in N$, a proposer $p \in N$, and a pair of preferences $R_a \in \bar{\mathcal{D}}_a$ and $R_p \in \bar{\mathcal{D}}_p$ such that A always matches a and p with their most preferred achievable mate when they report R_a or R_p, respectively.[11] In particular, a and p must match with their most preferred achievable mate given the Latin Square profiles $(R_a, R_{-a}^{R_a})$ and $(R_p, R_{-p}^{R_p})$. Because this only needs to be true for a single acceptor-proposer pair and for a single pair of profiles, this fact alone is not sufficient to prove that the proportion of mechanisms that are more AM-manipulable than DA is large. For every preference R_i and every $i \in N$, it is however possible to construct sufficiently many variants of the Latin Square profiles $(R_i, R_{-i}^{R_i})$ to show that this proportion is, in fact, large.

Proposition 5. *Suppose that $\#W = \#M = h$ and the domain of preferences is $\bar{\mathcal{D}}$. (i) The proportion of minimally PS-manipulable mechanisms is at most $\left(\frac{2}{h}\right)^{h!}$. (ii) The proportion of minimally AM-manipulable mechanisms is at most $\left(\frac{h}{(h-1)!} + \frac{1}{h((h-1)!)^2}\right)$. (iii) There exist minimally AM-manipulable*

[10] For the case of minimal PS-manipulability, Proposition 4.(i) can be viewed as a consequence of Pathak and Sönmez (2013, Theorem 2).

[11] To be precise, only such mechanisms *and* DA can possibly be minimally AM-manipulable. The addition of DA is reflected by the second term in the bound of Propositions 5.(ii) and 6.(ii).

	Upper bounds on the proportion of			Lower bound on the proportion of
h	minimally AM-manipulable mechanisms	minimally PS-manipulable mechanisms	mechanisms more PS-manipulable than DA	mechanisms more AM-manipulable than DA
4	.674	.000	.001	.326
5	.209	.000	.000	.891
6	.050	.000	.000	.950
7	.001	.000	.000	.999

Table 1: Numerical values for the upper and lower-bounds of Propositions 5 and 6.

mechanisms different from DA and minimally PS-manipulable mechanisms different from DA.

Although the bounds in Proposition 5 are loose, they converge rapidly to zero as h increases.[12] The values of the bounds in Proposition 5 are given in Table 1 for some values of h.

By the above argument, $\left(\frac{h}{(h-1)!} + \frac{1}{h((h-1)!)^2}\right)$ is also an upper bound on the proportion of stable mechanisms that are *not* more AM-manipulable than DA. Thus, the proportion of stable mechanisms that *are* more AM-manipulable than DA tends to one as h tends to infinity.

A similar result does not hold for PS-manipulability. By the definition of the PS-criterion, DA fails to be less PS-manipulable than any stable mechanism A that, for *at least one* profile R^*, selects a stable matching for which some acceptor a is matched with f_a^R. Such mechanisms abound. For example, consider profile (13) and DA^W. A third of the stable mechanisms select matching μ^1 for this profile. Hence, only $(1 - \frac{1}{3})$ of the stable mechanisms select a matching given (13) that allows them to be more PS-manipulable than DA^W. For $h = 3$, there are 3! such Latin Square profiles, one for each possible preference of w_1. Thus, only $(1 - \frac{1}{3})^{3!}$ of the stable mechanisms select matchings given the 3! Latin Square profiles in a way that allows them to be more PS-manipulable than DA. In general, there are $h!$ Latin Square profiles, and only $(1 - \frac{1}{h})^{h!}$ of the stable mechanisms select matchings given these $h!$ profiles in a way that allows them to be more PS-manipulable than DA.

The next result summarizes the two last arguments.

[12] The bounds in Propositions 5 and 6 are established by analyzing the behavior of the DA and the minimally and maximally manipulable mechanisms on a small subset of profiles. Considering additional profiles would tighten the bounds.

Proposition 6. *Suppose that* $\#W = \#M$ *and the domain of preferences is* $\bar{\mathcal{D}}$. *(i) DA is less PS-manipulable than* at most $\left(1 - \frac{1}{h}\right)^{h!}$ *of the stable mechanisms.*[13] *(ii) DA is less AM-manipulable than* at least $1 - \left(\frac{h}{(h-1)!} + \frac{1}{h((h-1)!)^2}\right)$ *of the stable mechanisms.*

Again the bounds in Proposition 6 are loose (see footnote 12) but they converge rapidly to 0 in the case of (i) and to 1 in the case of (ii) as h increases (see Table 1).

Propositions 5 and 6 look at the relative abundance *among the set of all stable mechanisms* of (a) minimally manipulable mechanisms and (b) mechanisms that are more manipulable than DA. The set of stable mechanisms contains a number of mechanisms that are "exotic" in the sense that they associate profiles with stable matchings in a very unsystematic way. Rather than comparing DA with the whole class of stable mechanisms, it may be useful to compare DA with *salient* subsets of this class. The next section provides such comparisons when saliency is understood as the satisfaction of some fairness properties.

6 A conflict between fairness and manipulability

When only ordinal information on preferences is available, it is not easy to define a comprehensive concept of fairness. Some natural reference points can, however, be used to devise minimal fairness requirements. One such reference point is the situation in which an individual receives her or his worst possible outcome out of the set of admissible outcomes.

6.1 A conflict between miniworst and manipulability

In the spirit of the *minimum regret* criterion (Knuth, 1997), a minimal fairness requirement is that the set of individuals who receive their worst outcome in the set of admissible outcomes be minimal (with respect to inclusion).[14] In a matching problem, a natural set of acceptable outcomes would be the set of stable matchings. Formally, suppose that $C(R) \subseteq T$ is the subset of

[13] It can be shown that $\lim_{h\to\infty}\left(1 - \frac{1}{h}\right)^{h!} = 0$.

[14] Although the minimum regret and the miniworst criteria are similar in spirit, they differ in many ways. For example, the miniworst criterion does not ascribe a cardinal meaning to the rank of a mate. The two criteria are not logically related; neither criterion implies the other.

admissible outcomes given profile R. Mechanism A is **miniworst on** C if for all $R \in \mathcal{D}$ there exists no outcome $t \in C(R)$ such that

$$\big\{i \in N \mid t' \; R_i \; A(R) \text{ for all } t' \in C(R)\big\} \supset \big\{i \in N \mid t' \; R_i \; t \text{ for all } t' \in C(R)\big\}. \tag{14}$$

Henceforth, I focus on mechanisms that are miniworst *on the set of stable matchings* and the reference to the set stable matching is suppressed.

It is often argued that DA is unfair because proposers are matched with their most preferred achievable mate, whereas acceptors are matched with their least preferred achievable mate. In the one-to-one environment, the miniworst criterion captures similar fairness concerns. Indeed, observe that a stable mechanism A is miniworst if and only if, for all $R \in \mathcal{D}$, there exists no stable matching μ such that

$$\big\{i \in N \mid A_i(R) = \ell_i^R\big\} \supset \big\{i \in N \mid \mu_i = \ell_i^R\big\}. \tag{15}$$

Any Latin Square profile R^{LS} then shows that DA is not miniworst because a stable matching in which no individual i matches with $\ell_i^{R^{LS}}$ could be selected instead of the extreme matching selected by DA. On the contrary, miniworst mechanisms do select a stable matching in which no individual i matches with $\ell_i^{R^{LS}}$ for any Latin Square profile R^{LS}. This observation lead to next proposition.

Let $\mathcal{D}^3 \subset \mathcal{D}$ be the domain all profiles in $\mathcal{D}$ in which individuals have at least three acceptable mates. When no reference to a subdomain of $\mathcal{D}$ is made, the result holds for the domain $\mathcal{D}$.

Proposition 7. *(i) No miniworst mechanism is minimally PS-manipulable. (ii) When the domain is $\mathcal{D}^3$, no miniworst mechanism is minimally AM-manipulable.*

Even outside of the one-to-one environment, the miniworst criterion is likely to conflict with manipulability. Suppose that $A_i(R)$ is not i's least preferred outcome according to R_i. If i can report $R_i|_{A_i(R)}$ instead of R_i, there are chances that i benefits from this misreport. Indeed, if the outcome does not change and $A(R_i|_t, R_{-i}) = A(R)$, then i becomes one of the individuals who, in $A(R_i|_t, R_{-i})$, get their worst outcome among the outcomes in $C(R_i|_{A(R)}, R_{-i})$ (according to $R_i|_t$). If A satisfies the miniworst criterion, this can only occur if there is no way to select another outcome that i prefers to $A_i(R)$ according to $R_i|_{A(R)}$ – and hence according to R_i – without making more individuals get their worst possible outcome in $C(R_i|_{A(R)}, R_{-i})$. In any other cases, the mechanism has to select an outcome $A(R_i|_t, R_{-i})$ that i prefers to $A(R)$ according to R_i.

In the one-to-one environment, the next proposition shows just how deep the conflict between the miniworst criterion and manipulability can get.

Proposition 8. *(i) Mechanism A is miniworst if and only if A is maximally PS-manipulable. (ii) When the domain is $\mathcal{D}^3$, any miniworst mechanism A is maximally A-manipulable. (iii) The converse is not true: when $\#W, \#M \geq 8$ and the domain is $\mathcal{D}^3$ there exists maximally AM-manipulable mechanisms that are not miniworst.*

As was observed earlier, given any Latin Square profile R^{LS}, miniworst mechanisms select matchings in which no individual matches with $\ell_i^{R^{LS}}$. This yields the following proposition.[15]

Proposition 9. *When the domain is $\mathcal{D}^3$, DA is less AM-manipulable than any miniworst mechanism.*

Proposition 9 indicates that, when fairness is understood as the satisfaction of the miniworst criterion, any fairness improvement over DA that does not compromises on stability comes at the cost of an increase in AM-manipulability.

A similar result does not hold for PS-manipulability for the same reason as in Proposition 6: a mechanism A only needs to match an acceptor with her or his most preferred achievable mate *on a single profile* to guarantee that DA is not less PS-manipulable than A. For example, consider the following profile:

$$\begin{array}{llll} R_{w_1}\colon & m_1 & m_2 & m_3 \\ R_{w_2}\colon & m_2 & m_3 & m_1 \\ R_{w_3}\colon & m_3 & m_1 & m_2 \end{array} \qquad \begin{array}{llll} R_{m_1}\colon & w_3 & w_1 & w_2 \\ R_{m_2}\colon & w_1 & w_2 & w_3 \\ R_{m_3}\colon & w_2 & w_3 & w_1 \end{array} . \tag{16}$$

Given R, there are only two stable matchings that correspond to the men and the women optimal matchings:

$$\begin{array}{llll} \mu^1\colon & m_1 & m_2 & m_3 \\ \mu^2\colon & m_2 & m_3 & m_1 \end{array} .$$

Both stable matchings can therefore be selected by a miniworst mechanism and DA^W (DA^M) will not be less PS-manipulable than the miniworst mechanism that selects the men (women) optimal matching given the above profile.

[15] Proposition 9 does *not* follow directly from Proposition 8. In general, it is possible for a mechanism to be minimally manipulable but to fail to be less manipulable than a maximally manipulable mechanism. See the example for PS-manipulability after the next proposition.

6.2 A conflict between median stable mechanisms and manipulability

Although the miniworst criterion excludes the selection of *some* stable matching, it does not provide a systematic procedure to select a unique intermediate stable matching for every R. One clever approach to do so was proposed by Teo and Sethuraman (1998). For any profile R, let k be the number of stable matchings given R. For every individual $i \in N$, the k stable matching can be (weakly) ordered according to R_i. Surprisingly, Teo and Sethuraman (1998) show that for any $\ell \in \{1, \ldots, k\}$, matching :

(i) every woman with the man she would match with under the stable matching she ranks ℓ-th, and

(ii) matching every man with the woman he would match with under the stable matching he ranks $(k - \ell + 1)$-th

results in a well-defined stable matching.

Teo and Sethuraman (1998) then suggest the following procedure to select a compromise stable matching: for every R, select the stable matchings obtained from the above procedure with ℓ equal to (one of the) median(s) of $\{1, \ldots, k\}$. This procedure defines the **median stable mechanisms** (MSM).

Like mechanisms that satisfy the miniworst criterion, MSM select a compromise stable matching at the cost of an increase in manipulability. It is easy to find profiles for which MSM select a stable matching in which *not a single individual* matches with her or his best achievable mate. Examples include the Latin Square profiles in $\mathcal{D}^3$. As a consequence, we have the following proposition.

Proposition 10. *(i) MSM are not minimally PS-manipulable and (ii) when the domain is $\mathcal{D}^3$, MSM are not minimally AM-manipulable either.*

In fact, MSM select a stable matching in which no individual matches with her or his best achievable mate in *every* Latin Square profile. By an argument already used above, DA is therefore less AM-manipulable than MSM. Again, this is not the case for PS-manipulability because MSM sometimes selects stable matchings in which both a woman *and* a man match with their best achievable mates (despite both individuals having multiple achievable mates).

Proposition 11. *(i) When the domain is $\mathcal{D}^3$, DA is less AM-manipulable than MSM but (ii) DA is* not *less PS-manipulable than MSM (even on $\mathcal{D}^3$).*

Proof of (ii). Consider the following profile:[16]

$$\begin{array}{lllll} R_{w_1}: & m_2 & m_1 & m_3 & m_4 \\ R_{w_2}: & m_4 & m_2 & m_1 & m_3 \\ R_{w_3}: & m_3 & m_1 & m_2 & m_4 \\ R_{w_4}: & m_2 & m_3 & m_1 & m_4 \end{array} \qquad \begin{array}{lllll} R_{m_1}: & w_4 & w_3 & w_2 & w_1 \\ R_{m_2}: & w_3 & w_2 & w_1 & w_4 \\ R_{m_3}: & w_2 & w_1 & w_4 & w_3 \\ R_{m_4}: & w_1 & w_4 & w_3 & w_2 \end{array} .$$

Given R, the stable matchings are

$$\begin{array}{llllll} \mu^1: & m_2 & m_4 & m_1 & m_3 \\ \mu^2: & m_3 & m_4 & m_2 & m_1 \\ \mu^3: & m_4 & m_3 & m_2 & m_1 \end{array} .$$

MSM select stable matching μ^2. Note that $\mu^2_{w_2} = m_4 = f^R_{w_2}$ and $\mu^2_{m_1} = w_4 = f^R_{m_1}$. Thus, the set of individuals who cannot PS-manipulate at R contains $\{w_2, m_1\}$. This set is contained in neither W nor M, which are the sets of individuals who cannot PS-manipulate at R in the two variants of DA. Hence, neither variant of DA is less PS-manipulable than MSM. ■

Because of the behavior of MSM on Latin Square profiles, MSM are also maximally AM-manipulable on $\mathcal{D}^3$. Again, the same is not true for PS-manipuability. As much as MSM strive to select compromise stable matchings, there exists profiles for which MSM select a stable matching in which the set of individuals who match with their worst achievable mate is not minimal. That is, MSM does *not* satisfy the miniworst criterion on the set of stable matchings. By Proposition 8, this leads to the following proposition.

Proposition 12. *(i) When the domain is $\mathcal{D}^3$, MSM is maximally AM-manipulable. (ii) For $\#W, \#M \geq 8$, MSM are* not *maximally PS-manipulable (even on $\mathcal{D}^3$).*

7 Extensions: many-to-one matching

The results in this paper are for the one-to-one environment. Some of the results in Sections 4 and 5 extend, however, to more general matching environments.

Consider a many-to-one matching environments (also known as "college admission environment") where each student $s \in S$ matches with at most one college $c \in C$, but colleges can admit up to $q_c \geq 1$ students. Suppose that colleges $c \in C$ have responsive preferences over subsets of students (Roth, 1985).

[16] The example below is for $\#W, \#M \geq 4$. For $4 \geq \#W, \#M \geq 3$, each variant of DA fails to be less PS-manipulable than one variant of the MSM mechanisms given profile (16).

In this environment, the *student-proposing* DA (a) assign students to their most preferred achievable college (Roth, 1985) and (b) makes it a dominant strategy for students to report their preferences truthfully. Also (c) the set of students who match with a college is the same in every stable matching (Roth, 1984).

Properties (a), (b), and (c) can be used in place of Lemmas 1, 2 and 4 to prove an equivalent to Proposition 1 for students: for any stable mechanism A, *student* $s \in S$ has a truthful dominant strategy in A given that other students and colleges report R_{-s} (i.e., (7) holds) if and only if s matches with her or his most preferred achievable college (i.e., (8) holds).

Let mechanism A be no more PS-manipulable than mechanism B *for students* if 2 holds with "$i \in N$" replaced by "$s \in S$". Propositions 2 and 3 then follow for students: (I) stable mechanism A is no more PS-manipulable than stable mechanism B for students if and only if (10) holds with "$i \in N$" replaced by "$s \in S$"; and (II) mechanism A is PS-minimally (resp. maximally) manipulable *for students* if and only if (11) (resp. (12)) holds with "$i \in N$" replaced by "$s \in S$". By the extension of Proposition 1 to students and (a), Proposition 4 also generalizes to the student-proposing DA.[17]

Things are more complicated for colleges and the college-proposing DA. The college-proposing DA does not provide colleges with a truthful dominant strategy : in fact, no stable mechanism does (Roth, 1985). As a consequence, there is no equivalent of Lemmas 2 and 3 for colleges. In particular, colleges may be able to manipulate even when they match with their most preferred achievable set of students because they prefer some achievable set of students under different preferences (i.e., $f_c^{(R'_c,R_{-c})} \; P_c \; f_c^R$, see Roth (1985)).

However, the college-proposing DA does (a') assign colleges to their most preferred achievable set of students and (b') the number of students assigned to each colleges is the same in every stable matching (Roth, 1984). These two properties are sufficient to generalize the necessity part of Proposition 1: for any stable mechanism A, if *college* $c \in C$ has a truthful dominant strategy in A given that other students and colleges report R_{-c} (i.e., (7) holds), then c matches with its most preferred achievable set of students (i.e., (8) holds), although the converse needs not be true.[18]

[17] Proposition 1.(ii) follows for the student-proposing DA because the domain of one-to-one profiles is a subset of the domain of many-to-one profiles. The proof of Proposition 1.(ii) can easily be extended to domains of many-to-one profiles in which all colleges have multiple seats, by a replication argument.

[18] In the corresponding proof for colleges in the many-to-one environment, $R_c|_{f_c^R}$ is replaced by the preference in which only the students in f_c^R are acceptable. Then $f_c^R \; P_c \; A_c(R_c|_{f_c^R}, R_{-c})$ implies that $A_c((R_c|_{f_c^R}, R_{-c})$ is some *strict subset* of f_c^R. This

The extension of necessity part of Proposition 1 enables generalizing the sufficiency parts of Proposition 2 and 3: (I) if (10) holds, then stable mechanism A is no more PS-manipulable than stable mechanism B; and (II) if (11) (resp. (12)) holds, then mechanism A is PS-minimally (resp. maximally) manipulable (although the converse needs not be true for both (I) and (II)). Propositions 4 then follows for the college-proposing DA.[19]

8 Concluding remarks

Whether the other results from Section 5 and the results from Section 6 extend to more general matching environments is left as an open question. For example, a natural is question is whether the results in Section 6.2 extend to median stable mechanisms in a many-to-one matching environment (see Klaus and Klijn (2006)).

Another question is whether DA is less PS-manipulable than miniworst mechanisms and the MSM *in the large*. As Section 6 illustrates, this is not true in general because some profiles do not admit a stable matching in which no individual matches with her or his least preferred achievable mate (or do not admit sufficiently many of these matchings in the case of MSM). Based on the examples in Section 6, one might hope that these profiles are rare and become arbitrarily unlikely as the number of individuals grows. In this sense, DA would be less PS-manipulable than any miniworst mechanisms or MSM *in the large*.

Formally, consider a sequence of domains $\{\mathcal{D}_k\}_{k=1}^{\infty}$ and any mechanism that associates an outcome with every profile in every domain of the sequence. Mechanism A is **less PS-manipulable than** mechanism B **in the large on** $\boldsymbol{\{\mathcal{D}_k\}_{k=1}^{\infty}}$ if the proportion of profiles $R \in \mathcal{R}^k$ for which

$$\begin{aligned}&\{i \in N \mid A \text{ is PS-manipulable for } i \text{ given } R\}\\ &\subset \{i \in N \mid B \text{ is PS-manipulable for } i \text{ given } R\}\end{aligned}$$

tends to one as $k \to \infty$.

Whether DA is less PS-manipulable than every miniworst mechanisms or every MSM in the large for non-trivial sequences $\{\mathcal{D}_k\}_{k=1}^{\infty}$ is less obvious than it may seems. Consider miniworst mechanisms and the sequence of domains $\{\bar{\mathcal{D}}_h\}_{h=1}^{\infty}$ where each $\bar{\mathcal{D}}_h \subset \mathcal{D}$ contains every profiles with no unacceptable

means that $\#A_c(R_c|_{f_c^R}, R_{-c}) \neq \#f_c^R$ and because $DA^C(R)$ is stable with respect to $(R_c|_{f_c^R}, R_{-c})$ and $DA_c^C(R) = f_c^R$, we have a contradiction of (b').

[19] Again, in the case of PS-manipulability, the generalizations of Proposition 4.(i) to the student and college-proposing DA can be viewed as consequences of Pathak and Sönmez (2013, Theorem 2).

mates when $\#W = \#M = h$. Then DA is *not* less PS-manipulable than every miniworst mechanism in the large *provided* that a conjecture by Pittel et al. (2008) holds.

Pittel et al. (2008) show that if a profile is drawn uniformly at random from $\bar{\mathcal{D}}_h$, the expected number of individuals with exactly two achievable mates tends to infinity as h tends to infinity. They conjecture that the exact distribution of this number is in fact *concentrated* around its expected value. If this is true, then as h tends to infinity, the probability that at least one $i \in N$ has exactly two achievable mates tends to one.[20] As illustrated in profile (16), when i has exactly two achievable mates in a given profile, a miniworst mechanism can select at least two sorts of stable matchings: one where i matches with i's most preferred achievable mate and another in which i matches with i's least preferred achievable mate. Thus, with probability one as h grows, a miniworst mechanism M^* that matches acceptors with their most preferred achievable mate when they have two achievable mates does so at least once in every profile. If the conjecture by Pittel et al. (2008) is true, then DA is clearly not less PS-manipulable in the large than any such miniworst mechanism M^*.

A Omitted proofs

Throughout, I provide proofs for DA^W and minimal manipulability. The proofs for DA^M and maximal manipulability are analogous.

Proof of Proposition 3.

Necessity. In order to derive a contradiction, suppose that some stable mechanism B is less manipulable than A. By Proposition 2, this implies that for some $R^* \in \mathcal{D}$,

$$\{i \in N \mid A_i(R^*) = f_i^{R^*}\} \subset \{i \in N \mid B_i(R^*) = f_i^{R^*}\}.$$

But because B is stable, $B_i(R^*)$ is stable with respect to R^*, which contradicts (11).

Sufficiency. In order to derive a contradiction, assume that μ^* is a stable matching satisfying (11) for some profile R^*. Consider mechanism B constructed from A by setting $B(R) = A(R)$ for all $R \in \mathcal{D}$ with $R \neq R^*$ and $B(R^*) = \mu^*$. Clearly, for all $R \in \mathcal{D}$ with $R \neq R^*$,

$$\{i \in N \mid A_i(R) = f_i^R\} = \{i \in N \mid B_i(R) = f_i^R\}. \tag{17}$$

[20] Pittel (1992, Note 2) show that this is true for the probability that at least one individual has *one and only one* achievable mate.

Also, by (11) and because $B_i(R^*) = \mu_i^*$ by construction,

$$\{i \in N \mid A_i(R^*) = f_i^{R^*}\} \subset \{i \in N \mid B_i(R^*) = f_i^{R^*}\}. \tag{18}$$

By Proposition 2, (17) and (18) imply that B is no more PS-manipulable than A but the converse is not true. Hence, by definition, B is less PS-manipulable than A and so A is not minimally PS-manipulable, a contradiction. ■

Proof of Proposition 4.(i).

PS-criterion. By Proposition 3, (i) holds provided that there does not exist a profile R^* and a stable matching μ^* such that

$$\left\{i \in N \mid DA_i^W(R^*) = f_i^{R^*}\right\} \subset \left\{i \in N \mid \mu_i^* = f_i^{R^*}\right\}. \tag{19}$$

In order to derive a contradiction, suppose that there exists such a stable matching and preference profile. By Lemma 1,

$$W \subseteq \left\{i \in N \mid DA_i^W(R^*) = f_i^{R^*}\right\}. \tag{20}$$

Hence, (19) implies that

$$\left\{i \in M \mid DA_i^W(R^*) = f_i^{R^*}\right\} \subset \left\{i \in M \mid \mu_i^* = f_i^{R^*}\right\}. \tag{21}$$

By (21), there exists $m^* \in M$ such that $\mu_{m^*}^* \neq DA_{m^*}^W(R^*)$. But this implies that $\mu^* \neq DA^W(R^*)$. Hence, there exists a woman $w^* \in W$ for whom $\mu_{w^*}^* \neq DA_{w^*}^W(R^*) = f_{w^*}^{R^*}$ and $W \not\subset \{i \in N \mid \mu_i^* = f_i^{R^*}\}$, contradicting (21).

AM-criterion. In order to derive a contradiction, suppose that some stable mechanism A is less AM-manipulable than DA^W. By Lemma 1 and Proposition 1, for all $R_* \in \cup_{i \in W} \mathcal{D}_i$,

$$\{w \in W \text{ with } R_* \in \mathcal{D}_w \mid DA^W \text{ is manipulable for } w \text{ given } R_*\} = \emptyset. \tag{22}$$

Thus, because A is less AM-manipulable than DA^W, for all $R_* \in \cup_{i \in W} \mathcal{D}_i$,

$$\{w \in W \text{ with } R_* \in \mathcal{D}_w \mid A \text{ is manipulable for } w \text{ given } R_*\} = \emptyset. \tag{23}$$

But by Proposition 1 again, (22) and (23) imply that

$$A_w(R) = DA_w^W(R) = f_w^R \qquad \text{for all } w \in W. \tag{24}$$

Hence, $A = DA^W$ by the definition of a matching, contradicting the assumption that A is less manipulable than DA^W. ■

Proof of Proposition 5.

Consider any $i \in N$ and any $R_i \in \bar{\mathcal{D}}_i$. By symmetry, we can let $i = w_1$ without loss of generality. Without loss of generality again, let us label the individuals in M in such a way that

$$R_i = R_{w_1}\colon \quad m_h \quad m_{(h-1)} \quad \dots \quad m_2 \quad m_1 \tag{25}$$

I construct a Latin Square profile $(R_i, R_{-i}^{R_i})$ that generalizes profile (13). The

preferences of the women are as follows:

$$\begin{array}{lccccccc}
R_{w_1}: & m_h & m_{h-1} & & \dots & & m_2 & m_1 \\
R^{R_i}_{w_2}: & m_{h-1} & & & & & & m_h \\
& & & & m_2 & \iddots & & \\
\vdots & \vdots & & m_2 & m_1 & m_h & & \vdots \\
& & & \iddots & m_h & & & \\
R^{R_i}_{w_{h-1}}: & m_2 & & & & & & m_3 \\
R^{R_i}_{w_h}: & m_1 & m_h & & \dots & & m_3 & m_2
\end{array} \tag{26}$$

The preferences of the men in $(R_i, R^{R_i}_{-i})$ are constructed symmetrically to (26) with woman w_1 appearing on the downward diagonal as in (13).

(i). For each $R_i \in \mathcal{D}_i$, profile $(R_i, R^{R_i}_{-i})$ has h stable matchings, only two of which (the women and men optimal matchings) can be selected by a PS-minimally manipulable mechanism. Consider the construction of a stable mechanism A. Fix the selection of a stable matching in A given any profile different from $(R_i, R^{R_i}_{-i})$ for some $R_i \in \bar{\mathcal{D}}_i$. Because there are $h!$ preferences in $\bar{\mathcal{D}}_i$, there are $h!$ profiles $(R_i, R^{R_i}_{-i})$, one for each $R_i \in \bar{\mathcal{D}}_i$. Among the $h^{h!}$ possible choices of stable matchings for these $h!$ profiles, only the $2^{h!}$ that select one of the women or the men optimal matchings for each $(R_i, R^{R_i}_{-i})$ make it possible for A to be PS-minimally manipulable. Hence, the proportion of minimally manipulable mechanisms among the class of stable mechanisms is at most $\left(\frac{2}{h}\right)^{h!}$.

(ii). See the proof of Proposition 6.(ii) below.

(iii). **PS-minimal manipulability**. Consider any Latin Square profile R^{LS} as described at the beginning of the proof. The mechanism A constructed from DA by setting $A(R^{LS})$ equal to the optimal matching of the accepting side of DA and $A(R') = DA(R')$ for all $R \in \bar{\mathcal{D}}$ with $R' \neq R^{LS}$ is minimally PS-manipulable. Because there are $h!$ Latin Square profiles and $h \geq 3$, there exists at least $h!$ such minimally PS-manipulable mechanisms different from DA.

AM-minimal manipulability. For any pair (i, R_i) with $i \in N$ and $R_i \in \bar{\mathcal{D}}_i$, if i cannot AM-manipulate in some mechanism A given R_i, then by Proposition 1,

$$A_i(R_i, R_{-i}) = f^R_i \qquad \text{for all } R_{-i} \in \bar{\mathcal{D}}_{-i} \tag{26.(i, R_i)}$$

Choose an arbitrary proposer-acceptor pair (a, p). Also, choose an arbitrary pair of profiles $R^*_p \in \bar{\mathcal{D}}_p$ and $R^*_a \in \bar{\mathcal{D}}_a$ such that R^*_p ranks a as p's least acceptable mate and R^*_a ranks p as a's least acceptable mate, excluding self-matches. Given the choice of a, p, R^*_a, and R^*_p, consider the following procedure:

Step 0. DA is individually rational by Lemma 1. Hence, by Lemma 4 and the construction of R_p^* and R_a^*, $f_p^{(R_p^*,R_a^*,R_{-\{p,a\}})} = a$ implies that p matches with a in every stable matching given $(R_p^*, R_a^*, R_{-\{p,a\}})$. Hence, $f_p^{(R_p^*,R_a^*,R_{-\{p,a\}})} = a$ implies $f_a^{(R_a^*,R_p^*,R_{-\{a,p\}})} = p$ and (26.(p, R_p^*)), and (26.(a, R_a^*)) are jointly feasible. Let $\mathcal{A}_0$ be the class of stable mechanisms that satisfy both (26.(p, R_p^*)) and (26.(a, R_a^*)).

Step 1. If there exists a pair $(i_1, R_{i_1}^*)$ with $i_1 \in N$, $R_{i_1}^* \in \bar{\mathcal{D}}_{i_1}$ and $(i_1, R_{i_1}^*) \notin \{(a, R_a^*), (p, R_p^*)\}$ such that some mechanism A satisfies all of (26.$(i_1, R_{i_1}^*)$), (26.(p, R_p^*)), and (26.(a, R_a^*)), let $\mathcal{A}_1 \subseteq \mathcal{A}_0$ be the class of such mechanisms and proceed to the next step. Otherwise, terminate the procedure.

$\vdots$

Generic step for $r \geq 1$. If there exists a pair $(i_r, R_{i_r}^*)$ with $i_r \in N$, $R_{i_r}^* \in \bar{\mathcal{D}}_{i_r}$, and $(i_r, R_{i_r}^*) \notin \{(a, R_a^*), (p, R_p^*), (i_1, R_{i_1}^*), \ldots, (i_{(r-1)}, R_{i_{(r-1)}}^*)\}$ such that some mechanism A satisfies all of (26.(p, R_p^*)), (26.(a, R_a^*)), (26.$(i_1, R_{i_1}^*)$), $\ldots$, (26.$(i_r, R_{i_r}^*)$), let $\mathcal{A}_r \subseteq \mathcal{A}_{r-1}$ be the class of such mechanisms and proceed to the next step. Otherwise, terminate the procedure.

Because the set of pairs (i, R_i) with $i \in N$ and $R_i \in \bar{\mathcal{D}}_i$ is finite, the procedure must terminate at some step $\tilde{r}$ sufficiently large. Because the procedure terminates at step $\tilde{r}$, for all $A_{\tilde{r}} \in \mathcal{A}_{\tilde{r}}$, it is impossible to find a mechanism $A_{\tilde{r}+1}$ such that the set of pairs (i, R_i) for which (26.(i, R_i)) holds for $A_{\tilde{r}+1}$ is a superset of the set of pairs for which (26.(i, R_i)) hold for $A_{\tilde{r}}$. Hence, the mechanisms in $\mathcal{A}_{\tilde{r}}$ are minimally AM-manipulable. Also, because the mechanisms in $\mathcal{A}_{\tilde{r}}$ satisfy (26.(p, R_p^*)), and (26.(a, R_a^*)), they are different from DA. Because the choice of a proposer-acceptor pair (a, p) is arbitrary, there exists at least h^2 such minimally AM-manipulable mechanisms different from DA. ■

Proof of Proposition 6.

(i). If DA is less PS-manipulable than stable mechanism A, then A can never select the optimal stable matching of the accepting side whenever any acceptor has more than one achievable mate. This implies that for any acceptor a and any of the $h!$ preferences $R_a \in \bar{\mathcal{D}}_a$, we have $A_a(R_a, R_{-a}^{R_a}) \neq f_a^{(R_a,R_{-a}^{R_a})}$, where the construction of $R_{-a}^{R_a}$ is described in the proof of Proposition 5. Because each $(R_a, R_{-a}^{R_a})$ has h stable matchings only one of which matches a with $f_a^{(R_a,R_{-a}^{R_a})}$, there are $h-1$ ways to select a stable $A_a(R_a, R_{-a}^{R_a}) \neq f_a^{(R_a,R_{-a}^{R_a})}$ for each $(R_a, R_{-a}^{R_a})$. Hence, of all the $h^{h!}$ possible selections of a stable matching with A for the $h!$ profiles $(R_a, R_{-a}^{R_a})$, only $(h-1)^{h!}$ make it possible for

DA to be less PS-manipulable than A. Therefore, at most $\left(\frac{h-1}{h}\right)^{h!}$ stable mechanisms A are more manipulable than DA.

(ii). For a stable mechanism A, if there exists no $R_* \in \cup_{i\in N}\bar{\mathcal{D}}_i$ and no acceptor $a \in N$ with $R_* \in \bar{\mathcal{D}}_a$ such that $A_a(R_*, R_{-a}) = f_a^{(R_*,R_{-a})}$ for all $R_{-a} \in \bar{\mathcal{D}}_{-a}$, then either $A = DA$ or A is more AM-manipulable than DA.[21] We are interested in the proportion of these mechanisms relative to the set of stable mechanisms.

Let $\mathbb{P}(X)$ denote the proportion of stable mechanisms A for which X is true. For every $i \in N$, let us label the preferences in $\bar{\mathcal{D}}_i$ following some arbitrary order $R_i^1, \ldots, R_i^{h!}$. The proportion we want to compute is equal to

$$1 - \mathbb{P}\Big(\vee_{\{i\in N | i \text{ is an acceptor}\}} \vee_{k\in\{1,\ldots,h!\}} A_i(R_i^k, R_{-i}) = f_i^{(R_i^k,R_{-i})} \text{ for all } R_{-i} \in \bar{\mathcal{D}}_{-i}\Big) \tag{27}$$

where $\vee$ stands for "or". The expression in (27) is at least

$$1 - \sum_{\{i\in N | i \text{ is an acceptor}\}} \sum_{k\in\{1,\ldots,h!\}} \mathbb{P}\big(A_i(R_i^k, R_{-i}) = f_i^{(R_i^k,R_{-i})} \text{ for all } R_{-i} \in \bar{\mathcal{D}}_{-i}\big). \tag{28}$$

For any profile $R \in \bar{\mathcal{D}}$, let $\sigma^R(X)$ denote the proportion of stable matchings μ for which X is true. Observe that

$$\begin{aligned}\mathbb{P}\big(A_i(R_i^k, R_{-i}) = f_i^{(R_i^k,R_{-i})} \text{ for all } R_{-i} \in \bar{\mathcal{D}}_{-i}\big) \\ = \prod_{R_{-i}\in\bar{\mathcal{D}}_{-i}} \sigma^{(R_i^k,R_{-i})}\big(\mu_i = f_i^{(R_i^k,R_{-i})}\big).\end{aligned}$$

For example, for the Latin Square profile $(R_i^k, R_{-i}^{R_i^k})$, we have $\sigma^{(R_i^k,R_{-i}^{R_i^k})}(\mu_i = f_i^{(R_i^k,R_{-i}^{R_i^k})}) = \frac{1}{h}$, which implies that

$$\mathbb{P}\big(A_i(R_i^k, R_{-i}) = f_i^{(R_i^k,R_{-i})} \text{ for all } R_{-i} \in \bar{\mathcal{D}}_{-i}\big) \leq \frac{1}{h}. \tag{29}$$

Substituting $\frac{1}{h}$ into (28) yields a bound that is looser than the bound in Proposition 6. A tighter bound for (28) can be obtained by tightening the bound in (29). This can be done by considering profiles different from the Latin Square profile. Specifically, I consider variations of the Latin Square profile for which (a) the number of stable matchings and (b) the proportion of stable matchings that match i with her or his most preferred achievable mate are easy to compute.

In what follows, I use the relabeling introduced at the beginning of the proof of Proposition 5, with $R_i^k = R_{w_1}$. The first variation of the Latin

[21] Hence, in the second case, A is not minimally AM-manipulable.

Square profile that is considered has $(h-1)$ stable matchings and is denoted by $(R_{w_1}, R_{-w_1}^{R_{w_1}}(h-1,1))$. In $(R_{w_1}, R_{-w_1}^{R_{w_1}}(h-1,1))$, the preferences of the women and of man m_h are as follows:

$$
\begin{array}{rccccccccc}
R_{w_1}\colon & m_h & m_{h-1} & m_{h-2} & & \cdots & & m_2 & m_1 \\
R_{w_2}^{R_{w_1}}(h-1,1)\colon & m_h & m_{h-2} & & & & & & m_{h-1} \\
 & & & & & m_2 & \iddots & & \\
\vdots & \vdots & & & m_2 & m_1 & m_{h-1} & \vdots & \vdots \\
 & & & & \iddots & m_{h-1} & & & \\
R_{w_{h-2}}^{R_{w_1}}(h-1,1)\colon & m_h & m_2 & & & & & & m_3 \\
R_{w_{h-1}}^{R_{w_1}}(h-1,1)\colon & m_h & m_1 & m_{h-1} & & \cdots & & m_3 & m_2 \\
R_{w_h}^{R_{w_1}}(h-1,1)\colon & m_h & & & & & & & \\
R_{m_h}^{R_{w_1}}(h-1,1)\colon & w_h & & & & & & &
\end{array}
\tag{30}
$$

In (30), every women ranks m_h first. Among the first $h-1$ women, the sub-profile excluding m_h has a Latin Square structure of dimension $h-1$ similar to (26). For w_h and m_h, only the most preferred mate is specified.

In $(R_{w_1}, R_{-w_1}^{R_{w_1}}(h-1,1))$, the preferences of men other than m_h are constructed symmetrically to the preferences of the women other than w_h in (30) with w_h ranked last and woman w_1 appearing on the downward diagonal as in (13).

Observe that m_h and w_h are matched together in every stable matching given $(R_{w_1}, R_{-w_1}^{R_{w_1}}(h-1,1))$ and m_h and w_h are therefore not achievable for other men and women. By analogy with (26), there are $(h-1)$ stable matching among the remaining individuals $\{m_1, \ldots, m_{h-1}, w_1, \ldots, w_{h-1}\}$ due to the Latin Square structure of the profile once m_h and w_h are removed. There are therefore $(h-1)$ stable matchings given $(R_{w_1}, R_{-w_1}^{R_{w_1}}(h-1,1))$, only one of which matches w_1 with her most preferred achievable mate.

A natural variant of (30), denoted $(R_{w_1}, R_{-w_1}^{R_{w_1}}(h-1,2))$, also has $(h-1)$ stable matchings. In $(R_{w_1}, R_{-w_1}^{R_{w_1}}(h-1,2))$, the preferences of the women and

of man m_1 are as follows:

$$
\begin{array}{lccccccccc}
R_{w_1}: & m_h & m_{h-1} & & \dots & & m_3 & m_2 & m_1 \\
R_{w_2}^{R_{w_1}}(h-1,2): & m_{h-1} & & & & & & m_h & m_1 \\
 & & & & m_3 & \iddots & & & \\
\vdots & \vdots & & m_3 & m_2 & m_h & & \vdots & \vdots \\
 & & & \iddots & m_h & & & & \\
R_{w_{h-2}}^{R_{w_1}}(h-1,2): & m_3 & & & & & & m_4 & m_1 \\
R_{w_{h-1}}^{R_{w_1}}(h-1,2): & m_2 & m_h & & \dots & & m_4 & m_3 & m_1 \\
R_{w_h}^{R_{w_1}}(h-1,2): & m_1 & & & & & & & \\
R_{m_1}^{R_{w_1}}(h-1,2): & w_h & & & & & & &
\end{array}
\tag{31}
$$

In (31), the first $h-1$ women rank m_1 last. Among the first $h-1$ women, the sub-profile excluding m_1 has a Latin Square structure of dimension $h-1$ similar to (26). For w_h and m_1, only the most preferred mate is specified.

In $(R_{w_1}, R_{-w_1}^{R_{w_1}}(h-1,2))$, the preferences of men other than m_1 are constructed symmetrically to the preferences of the women other than w_h in (30) with w_h ranked last and woman w_1 appearing on the downward diagonal as in (13).

Similarly to $(R_{w_1}, R_{-w_1}^{R_{w_1}}(h-1,1))$, there are $(h-1)$ stable matchings given $(R_{w_1}, R_{-w_1}^{R_{w_1}}(h-1,2))$ only one of which matches w_1 with her most preferred achievable mate.

It is easy to see how, for all $k \in \{2,\dots,h-1\}$, the above constructions extend to profiles $(R_{w_1}, R_{-w_1}^{R_{w_1}}(k,1))$ and $(R_{w_1}, R_{-w_1}^{R_{w_1}}(k,2))$ admitting k stable matchings only one of which matches w_1 with her most preferred achievable mate. In $(R_{w_1}, R_{-w_1}^{R_{w_1}}(h-2,1))$ for example, the first $h-2$ women rank m_h *and* m_{h-1} first and, among the first $h-2$ women, the sub-profile excluding m_h *and* m_{h-1} has a Latin Square structure of dimension $h-2$. Also, w_h ranks m_h first *and* w_{h-1} ranks m_{h-1} first.

Together with the original Latin Square profile, we have therefore identified $1+2(h-1)$ profiles with a (partial) Latin square structure in which i's preference is R_i^k. In other words, we have identified a set of sub-profiles $\{R_{-i}^1,\dots,R_{-i}^{1+2(h-1)}\}$ such that the set of profiles (R_i^k, R_{-i}^t) for $t \in \{1,\dots,1+2(h-1)\}$ consists of the Latin Square profile and the $2(h-1)$ profiles described above.

There are $h((h-1)!)^2$ ways to select stable matchings for these $1+2(h-1)$ profiles, only one of which always matches i with i's most preferred achievable mate. Hence,

$$\begin{aligned}\frac{1}{h((h-1)!)^2} &= \prod_{R_{-i}\in\{R_{-i}^1,\ldots,R_{-i}^{1+2(h-1)}\}} \sigma^{(R_i^k,R_{-i})}\big(\mu_i = f_i^{(R_i^k,R_{-i})}\big) \\ &\geq \prod_{R_{-i}\in\bar{\mathcal{D}}_{-i}} \sigma^{(R_i^k,R_{-i})}\big(\mu_i = f_i^{(R_i^k,R_{-i})}\big).\end{aligned} \tag{32}$$

Using (32) in (28) shows that (27) is at least

$$1 - \sum_{\{i\in N | i \text{ is acceptor}\}} \sum_{k\in\{1,\ldots,h!\}} \frac{1}{h((h-1)!)^2}. \tag{33}$$

Because the fraction in (33) is independent of the indices used in the summations, (33) is equal to $1 - \frac{h(h!)}{h((h-1)!)^2} = 1 - \frac{h}{(h-1)!}$.

Finally, we must account for the fact that DA itself might be one of the at most $1 - \frac{h}{(h-1)!}$ mechanisms A for which there exists no $R_* \in \cup_{i\in N}\bar{\mathcal{D}}_i$ and no acceptor $a \in N$ with $R_* \in \bar{\mathcal{D}}_a$ such that $A_a(R_*, R_{-a}) = f_a^{(R_*,R_{-a})}$ for all $R_{-a} \in \bar{\mathcal{D}}_{-a}$. Because DA is not less AM-manipulable than DA itself, we must not include it when computing the upper bound.

Clearly, DA by itself represents a very small proportion of the stable mechanisms. For example, only $\frac{1}{h((h-1)!)^2}$ of the mechanisms select a combination of stable matchings for the Latin Square profile and the $2(h-1)$ variants described above that is compatible with the mechanism being DA. Hence, overall, the proportion of stable mechanisms that are more AM-manipulable than DA is at least $1 - \big(\frac{h}{(h-1)!} + \frac{1}{h((h-1)!)^2}\big)$. ■

Proof of Proposition 7.

(i). Consider the profile in the proof of Proposition 4(ii). For any mini-worst mechanism A, the mechanism B constructed from A by changing the stable matching selected under this profile to the men optimal or women optimal stable matching is less PS-manipulable than A. As explained at the end of the proof of Proposition 4(ii), this profile can easily be extended to a profile with identical properties when $\#M > 3$ and $\#W > 3$.

(ii). See the proof of Proposition 9. ■

Proof of Proposition 8.

(i). **Sufficiency**. Consider any profile R and any stable matching μ. By assumption, (15) is false. That is, either

$$\{i \in N \mid A_i(R) = \ell_i^R\} = \{i \in N \mid \mu_i = \ell_i^R\}, \tag{34}$$

or there exists an $i^* \in N$ such that

$$\mu_{i^*} = \ell_{i^*}^R \text{ and } A_{i^*}(R) \neq \ell_{i^*}^R. \tag{35}$$

By Lemmas 1 and 4, (34) implies that

$$\{i \in N \mid A_i(R) = f_i^R \text{ and } A_i(R) \neq i\} = \{i \in N \mid \mu_i = f_i^R \text{ and } \mu_i \neq i\}.$$

Then by Lemma 4 again,

$$\{i \in N \mid A_i(R) = f_i^R = i\} = \{i \in N \mid \mu_i = f_i^R = i\}.$$

Thus, if (34) holds, Lemma 1 implies

$$\{i \in N \mid A_i(R) = f_i^R\} = \{i \in N \mid \mu_i = f_i^R\}. \tag{36}$$

On the other hand, if (35) holds, we have

$$A_{\ell_{i^*}^R}(R) \neq f_{\ell_{i^*}^R}^R = i^* \quad \text{and} \quad \mu_{\ell_{i^*}^R} = f_{\ell_{i^*}^R}^R = i^*. \tag{37}$$

If (36) holds, then the set of individuals who are matched with their most preferred achievable mate is the same in $A(R)$ and μ. On the other hand, if (37) holds, then there is an individual $\ell_{i^*}^R$ who is matched with $f_{\ell_{i^*}^R}^R$ in μ, but not in $A(R)$. In both cases, the set of individuals who are matched with their most preferred achievable mates in A is not a superset of the set of individuals who are matched with their most preferred achievable mates in μ, that is,

$$\{i \in N \mid A_i(R) = f_i^R\} \not\supset \{i \in N \mid \mu_i = f_i^R\}. \tag{38}$$

But because (34) implies (36) and (35) implies (37), (38) must be true. Finally, because (38) implies that (12) does not hold for stable matching μ and profile R and because μ and R were chosen arbitrarily, A is maximally PS-manipulable among the stable mechanisms by Proposition 3.

Necessity. In order to derive a contradiction, assume that there exists a profile R^* and a matching μ^* such that (15) holds. By an argument similar to the one used in the sufficiency part of the proof, Lemma 1 implies

$$\{i \in N \mid A_i(R^*) = f_i^{R^*}\} \supset \{i \in N \mid \mu_i^* = f_i^{R^*}\}. \tag{39}$$

Now, construct mechanism B from A by setting $B(R) = A(R)$ for all $R \in \mathcal{D}$ with $R \neq R^*$, and $B(R^*) = \mu^*$. By Proposition 1, because $B(R) = A(R)$ for all $R \neq R^*$, we have

$$\begin{aligned} &\{i \in N \mid A \text{ is manipulable for } i \text{ given } R\} = \\ &\{i \in N \mid B \text{ is manipulable for } i \text{ given } R\} \text{ for all } R \in \mathcal{D} \text{ with } R \neq R^*. \end{aligned} \tag{40}$$

Also, by (39) and Proposition 1 again,

$$\begin{aligned} &\{i \in N \mid A \text{ is manipulable for } i \text{ given } R^*\} \\ &\quad \subset \{i \in N \mid B \text{ is manipulable for } i \text{ given } R^*\}. \end{aligned} \tag{41}$$

Together, (40) and (41) imply that that A is less PS-manipulable than B and therefore A is not minimally PS-manipulable, a contradiction.

(ii). For any $i \in N$ and any $R_i \in \mathcal{D}_i^3$, it is possible to construct a Latin Square profile similar to (26) among the acceptable mates given R_i.[22] Slightly abusing the notation, this profile is also denoted $(R_i, R_{-i}^{R_i})$.

If mechanism A is miniworst, then for any $i \in N$ and any $R_i \in \mathcal{D}_i^3$, because R_i admits at least three acceptable mates, $A_i(R_i, R_{-i}^{R_i}) \neq f_i^{(R_i, R_{-i}^{R_i})}$. Hence, for any $R_* \in \cup_{i \in N} \mathcal{D}_i^3$,

$$\{i \in N \text{ with } R_* \in \mathcal{D}_i^3 \mid A \text{ is AM-manipulable for } i \text{ given } R_*\} = N$$

and A is clearly maximally AM-manipulable.

(iii). Consider any mechanism A such that

(a) for any $i \in N$ and any $R_i \in \mathcal{D}_i^3$, $A_i(R_i, R_{-i}^{R_i}) \neq f_i^{(R_i, R_{-i}^{R_i})}$ (i.e, in any Latin Square profile, A selects an stable matching that matches no individual with her or his most preferred achievable mate), but

(b) for any $R^* \in \mathcal{D}^3$ with $R^* \notin \{(R_i, R_{-i}^{R_i}) \in \mathcal{D}^3 \mid R_i \in \mathcal{D}_i^3 \text{ for some } i \in N\}$, let $A(R^*) = DA(R^*)$ (i.e., for any profile that is *not* a Latin Square, let A select the same matching as DA).

By (a), for any $R_* \in \cup_{i \in N} \mathcal{D}_i^3$,

$$\{i \in N \text{ with } R_* \in \mathcal{D}_i^3 \mid A \text{ is AM-manipulable for } i \text{ given } R_*\} = N,$$

and A is maximally AM-manipulable. However, for many $R^* \in \mathcal{D}^3$ with $R^* \notin \{(R_i, R_{-i}^{R_i}) \in \mathcal{D}^3 \mid R_i \in \mathcal{D}_i^3 \text{ for some } i \in N\}$, there exists stable matchings μ such that (14) holds. This is the case, for example, in the profile presented in the proof of Proposition 12. For this profile, DA selects either μ^9 or μ^1 (depending on the variant of DA that is used). Hence, by construction, A selects either μ^9 or μ^1 although μ^4 satisfies (14). Thus, A is maximally AM-manipulable but not miniworst. ■

Proof of Proposition 9.

As shown in the proof of Proposition 8.(ii), if a mechanism is miniworst, then for any $R_* \in \cup_{i \in N} \mathcal{D}_i^3$,

$$\{i \in N \text{ with } R_* \in \mathcal{D}_i^3 \mid A \text{ is AM-manipulable for } i \text{ given } R_*\} = N,$$

and DA is less AM-manipulable than A. ■

Proof of Proposition 10.

(i). Consider any Latin Square profile $R^{LS} \in \mathcal{D}$ that admits more than 3 stable matchings. Given R^{LS}, MSM select a stable matching in which no individual matches with her or his best achievable mate (among the individuals

[22] For example, if three mates are acceptable given R_i, reproduce (26) using two additional individuals on i's side and the three acceptable mates given R_i on the other side. Other individuals have no acceptable mate.

that are not single in all stable matching). The mechanism A constructed from any MSM by only changing the stable matching selected under R^{LS} to either of the optimal stable matchings is less PS-manipulable than this MSM. Hence, MSM are not PS-minimally manipulable.

(ii). Like miniworst mechanisms, in Latin Square profiles admitting more than three stable matchings, MSM select a stable mechanism in which no-one matches with her or his most preferred achievable mate. Hence, for any $i \in N$ and any $R_i \in \mathcal{D}_i^3$, because R_i admits at least three acceptable mates, we have $MSM_i(R_i, R_{-i}^{R_i}) \neq f_i^{(R_i, R_{-i}^{R_i})}$. Thus, for any $R_* \in \cup_{i \in N} \mathcal{D}_i^3$,

$$\{i \in N \mid R_* \in \mathcal{D}_i^3 \text{ and } MSM \text{ is AM-manipulable for } i \text{ given } R_*\} = N \quad (42)$$

and MSM is clearly not minimally AM-manipulable. ■

Proof of Proposition 11.(i).

See (42) in the proof of Proposition 10.(ii). ■

Proof of Proposition 12.

(i). See (42) in the proof of Proposition 10.(ii).

(ii). By Proposition 8, it is enough to show that MSM are not miniworst on the set of stable matchings.

Consider the following profile:

$$\begin{array}{lllllllll}
R_{w_1}: & m_3 & m_8 & m_7 & m_6 & m_5 & m_4 & m_2 & m_1 \\
R_{w_2}: & m_2 & m_8 & m_7 & m_6 & m_5 & m_4 & m_1 & m_3 \\
R_{w_3}: & m_1 & m_8 & m_7 & m_6 & m_5 & m_4 & m_3 & m_2 \\
R_{w_4}: & m_8 & m_7 & m_6 & m_5 & m_4 & w_4 & & \\
R_{w_5}: & m_7 & m_6 & m_5 & m_4 & m_8 & w_5 & & \\
R_{w_6}: & m_6 & m_5 & m_4 & m_8 & m_7 & w_6 & & \\
R_{w_7}: & m_5 & m_4 & m_8 & m_7 & m_6 & w_7 & & \\
R_{w_8}: & m_4 & m_8 & m_7 & m_6 & m_5 & w_8 & &
\end{array}$$

$$\begin{array}{lllllllll}
R_{m_1}: & w_1 & w_2 & w_3 & m_1 & & & & \\
R_{m_2}: & w_3 & w_1 & w_2 & m_2 & & & & \\
R_{m_3}: & w_2 & w_3 & w_1 & m_3 & & & & \\
R_{m_4}: & w_4 & w_5 & w_1 & w_2 & w_3 & w_6 & w_7 & w_8 \\
R_{m_5}: & w_8 & w_4 & w_1 & w_2 & w_3 & w_5 & w_6 & w_7 \\
R_{m_6}: & w_7 & w_8 & w_1 & w_2 & w_3 & w_4 & w_5 & w_6 \\
R_{m_7}: & w_6 & w_7 & w_1 & w_2 & w_3 & w_8 & w_4 & w_5 \\
R_{m_8}: & w_5 & w_6 & w_1 & w_2 & w_3 & w_7 & w_8 & w_4
\end{array}$$

Observe that because no two women have the same most preferred men, the matching that matches every women with her favorite men is stable. Thus, because the set of individuals who are married is the same in every stable matching (Lemma 4), every stable matching matches every individual.

Thus, because none of $\{m_1, m_2, m_3\}$ are acceptable to any of $\{w_4, w_5, w_6, w_7, w_8\}$, but are acceptable to any of $\{w_1, w_2, w_3\}$, all men in $\{m_1, m_2, m_3\}$ must match with women in $\{w_1, w_2, w_3\}$ in any stable matching. As a consequence, all men in $\{m_4, m_5, m_6, m_7, m_8\}$ also match with women from $\{w_4, w_5, w_6, w_7, w_8\}$ in every stable matching.

Among $\{m_1, m_2, m_3\} \cup \{w_1, w_2, w_3\}$, the stable sub-matchings are

$$\begin{array}{lccc} \mu^1_{123}\colon & m_1 & m_3 & m_2 \\ \mu^2_{123}\colon & m_2 & m_1 & m_3 \\ \mu^3_{123}\colon & m_3 & m_2 & m_1 \end{array}$$

Among $\{m_4, m_5, m_6, m_7, m_8\} \cup \{w_4, w_5, w_6, w_7, w_8\}$, the stable sub-matchings are

$$\begin{array}{lccccc} \mu^1_{45678}\colon & m_4 & m_8 & m_7 & m_6 & m_5 \\ \mu^2_{45678}\colon & m_5 & m_4 & m_8 & m_7 & m_6 \\ \mu^3_{45678}\colon & m_6 & m_5 & m_4 & m_8 & m_7 \\ \mu^4_{45678}\colon & m_7 & m_6 & m_5 & m_4 & m_8 \\ \mu^5_{45678}\colon & m_8 & m_7 & m_6 & m_5 & m_4 \end{array}$$

Observe that in any stable matching including μ^1_{123} or μ^2_{123}, the sub-matching among $\{m_4, m_5, m_6, m_7, m_8\} \cup \{w_4, w_5, w_6, w_7, w_8\}$ must be either μ^1_{45678} or μ^2_{45678}. Indeed, in any other combination including μ^1_{123} or μ^2_{123}, e.g. $(\mu^1_{123}, \mu^4_{45678})$, every women in $\{w_4, w_5, w_6, w_7, w_8\}$ forms a blocking pair with every men in $\{m_1, m_2, m_3\}$.

On the other hand, in stable matchings including μ^3_{123}, the sub-matching among $\{m_4, m_5, m_6, m_7, m_8\} \cup \{w_4, w_5, w_6, w_7, w_8\}$ can be any of the stable sub-matchings $(\mu^1_{45678}, \ldots, \mu^5_{45678})$.

Overall, the stable matchings are

$$\begin{array}{ll} \mu^1 \coloneqq (\mu^1_{123}, \mu^1_{45678}) & \mu^5 \coloneqq (\mu^3_{123}, \mu^1_{45678}) \\ \mu^2 \coloneqq (\mu^1_{123}, \mu^2_{45678}) & \mu^6 \coloneqq (\mu^3_{123}, \mu^2_{45678}) \\ \mu^3 \coloneqq (\mu^2_{123}, \mu^1_{45678}) & \mu^7 \coloneqq (\mu^3_{123}, \mu^3_{45678}) \\ \mu^4 \coloneqq (\mu^2_{123}, \mu^2_{45678}) & \mu^8 \coloneqq (\mu^3_{123}, \mu^4_{45678}) \\ & \mu^9 \coloneqq (\mu^3_{123}, \mu^5_{45678}) \end{array}$$

Every women in $\{w_1, w_2, w_3\}$ ranks the stable matchings in the same way

$$\mu^9 \; R_w \; \mu^8 \; R_w \; \mu^7 \; R_w \; \mu^6 \; R_w \; \mu^5 \; R_w \; \mu^4 \; R_w \; \mu^3 \; R_w \; \mu^2 \; R_w \; \mu^1, \quad \text{for all } w \in \{w_1, w_2, w_3\}.$$

Hence, given profile R, MSM match every women in $\{w_1, w_2, w_3\}$ with her match under μ^5.

Every women in $\{w_4, w_5, w_6, w_7, w_8\}$ also ranks the stable matchings in

the same way

$$\mu^9 \; R_w \; \mu^8 \; R_w \; \mu^7 \; R_w \; \mu^6 \; R_w \; \mu^4 \; R_w \; \mu^2 \; R_w \; \mu^5 \; R_w \; \mu^3 \; R_w \; \mu^1,$$
$$\text{for all } w \in \{w_4, w_5, w_6, w_7, w_8\}.$$

Hence, given profile R, MSM matches every women in $\{w_4, w_5, w_6, w_7, w_8\}$ with her match under μ^4.

Thus, MSM selects matching μ^6 given profile R. Under μ^6, the set of individuals i who match with ℓ_i^R is $\{w_1, w_2, w_3\}$. But note that under μ^4 the set of individual i who match with ℓ_i^R is empty. Hence, MSM are not miniworst on the set of stable matchings, the desired result. ■

References

Aleskerov, F., Kurbanov, E., 1999. Degree of manipulability of social choice procedures, in: Alkan, P.A., Aliprantis, P.C.D., Yannelis, P.N.C. (Eds.), Current Trends in Economics. Springer, Berlin Heidelberg, pp. 13–27.

Andersson, T., Ehlers, L., Svensson, L.G., 2014. Least manipulable envy-free rules in economies with indivisibilities. Mathematical Social Sciences 69, 43–49.

Arribillaga, R.P., Massó, J., 2015. Comparing generalized median voter schemes according to their manipulability. Theoretical Economics 11, 547–586.

Chen, P., Egesdal, M., Pycia, M., Yenmez, M.B., 2016. Manipulability of Stable Mechanisms. American Economic Journal: Microeconomics 8, 202–214.

Coles, P., Shorrer, R., 2014. Optimal truncation in matching markets. Games and Economic Behavior 87, 591–615.

Decerf, B., Van der Linden, M., 2016. Manipulability and tie-breaking in constrained school choice. SSRN Working Paper No. 2809566.

Dubins, L.E., Freedman, D.A., 1981. Machiavelli and the Gale-Shapley algorithm. American Mathematical Monthly 88, 485–494.

Fujinaka, Y., Wakayama, T., 2012. Maximal manipulation in fair allocation. SSRN Working Paper No. 2051296.

Gale, D., Shapley, L.S., 1962. College admissions and the stability of marriage. American Mathematical Monthly 69, 9–15.

Gerber, A., Barberà, S., 2016. Sequential voting and agenda manipulation. Theoretical Economics .

Irving, R.W., Leather, P., Gusfield, D., 1987. An efficient algorithm for the "optimal" stable marriage. Journal of the ACM (JACM) 34, 532–543.

Klaus, B., Klijn, F., 2006. Median stable matching for college admissions. International Journal of Game Theory 34, 1–11.

Knuth, D.E., 1997. Stable Marriage and Its Relation to Other Combinatorial Problems: An Introduction to the Mathematical Analysis of Algorithms. American Mathematical Society, Providence, RI.

Maus, S., Peters, H., Storcken, T., 2007. Anonymous voting and minimal manipulability. Journal of Economic Theory 135, 533–544.

Pathak, P.A., Sönmez, T., 2013. School admissions reform in Chicago and England : Comparing mechanisms by their vulnerability to manipulation. American Economic Review 103, 80–106.

Pittel, B., 1992. On likely solutions of a stable marriage problem. Annals of Applied Probability 2, 358–401.

Pittel, B., Shepp, L., Veklerov, E., 2008. On the number of fixed pairs in a random instance of the stable marriage problem. SIAM Journal on Discrete Mathematics 21, 947–958.

Roth, A., 1982. The economics of matching: Stability and incentives. Mathematics of Operations Research 7, 617–628.

Roth, A., 1985. The college admissions problem is not equivalent to the marriage problem. Journal of Economic Theory 288, 277–288.

Roth, A., Sotomayor, M., 1992. Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis. Cambridge University Press, Cambridge.

Roth, A.E., 1984. The evolution of the labor market for medical interns and residents: A case study in game theory. Journal of Political Economy 92, 991–1016.

Roth, A.E., Vande Vate, J.H., 1991. Incentives in two-sided matching with random stable mechanisms. Economic Theory 1, 31–44.

Teo, C.P., Sethuraman, J., 1998. The geometry of fractional stable matchings and its applications. Mathematics of Operations Research 23, 874–891.