Stable matching mechanisms are not obviously strategy-proof

Itai Ashlagi Yannai A. Gonczarowski *

December, 2016

Abstract

Many two-sided matching markets, from labor markets to school choice programs, use a clearinghouse based on the applicant-proposing deferred acceptance algorithm, which is well known to be strategy-proof for the applicants. Nonetheless, a growing amount of empirical evidence reveals that applicants misrepresent their preferences when this mechanism is used. This paper shows that no mechanism that implements a stable matching is *obviously strategy-proof* for any side of the market, a stronger incentive property than strategy-proofness introduced by Li (2015). A stable mechanism that is obviously strategy-proof for applicants is introduced for the case in which agents on the other side have acyclical preferences. Our findings suggest that strategic reasoning in two-sided markets requires more cognitive effort by participants than in one-sided markets.

1 Introduction

A number of labor markets and school admission programs that can be viewed as two-sided matching markets use centralized mechanisms to match between agents on both sides of

*First draft: November 2015. Ashlagi: Management Science & Engineering, Stanford University, *email*: iashlagi@stanford.edu. Gonczarowski: Einstein Institute of Mathematics, Rachel & Selim Benin School of Computer Science & Engineering, and the Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem; and Microsoft Research, *email*: yannai@gonch.name. This paper greatly benefited from discussions with Sophie Bade, Shengwu Li, Jordi Massó, Muriel Niederle, Noam Nisan, Assaf Romm, and Peter Troyan. Itai Ashlagi is supported by NSF grant SES-1254768. Yannai Gonczarowski is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities; his work is supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. 249159, by ISF grant 1435/14 administered by the Israeli Academy of Sciences, and by Israel-USA Bi-national Science Foundation (BSF) grant number 2014389.

the market. One important criterion in the design of such mechanisms is stability (Roth, 2002), requiring that no two agents, one from each side of the market, prefer each other over the partners with whom they are matched. Another highly desired property is strategy-proofness, which alleviates agents' incentives to behave strategically.

Indeed, many clearinghouses have adopted in recent years the remarkable deferred acceptance (DA) mechanism (Gale and Shapley, 1962), which finds a stable matching and is strategy-proof for one side of the market, namely the proposing side in the DA algorithm (Dubins and Freedman, 1981).[1,2] Interestingly, although participants are advised that it is in their best interest to state their true preferences, empirical evidence suggests that a significant fraction nonetheless attempt to strategically misreport their true preferences (Hassidim et al., 2016; Rees-Jones, 2016). This paper asks whether one can design a stable mechanism that is *obviously strategy-proof*, an incentive property introduced by Li (2015) that is stronger than strategy-proofness.

Li (2015) formulated the idea that it is "easier to be convinced" of the strategy-proofness of some mechanisms over others. He introduces, and characterizes, the class of *obviously strategy-proof* mechanisms. He shows that, roughly speaking, obviously strategy-proof mechanisms are those whose strategy-proofness can be proved even under a cognitively limited proof model that does not allow for contingent reasoning.[3] In his paper, Li studies whether various well-known auction and assignment mechanisms with attractive revenue or welfare properties for one side of the market can be implemented in an obviously strategy-proof manner. Whether one may implement stable matchings in an obviously strategy-proof manner remains unknown.

For the purpose of this paper, we adopt the Gale and Shapley (1962) one-to-one matching market with men and women to represent two-sided matching markets; our results naturally extend to many-to-one markets such as labor markets and school choice programs. When women's preferences over men are perfectly aligned, the unique stable matching may be recovered via serial dictatorship, where men, in their ranked order, choose their partners. In this case, a sequential implementation of such serial dictatorship is obviously strategy-proof. (This follows from Li (2015), who shows that in a one-sided assignment market with agents and objects, serial dictatorship, when implemented sequentially, is obviously strategy-

[1]This mechanism is also approximately strategy-proof for all participants in the market (Immorlica and Mahdian, 2005; Kojima and Pathak, 2009; Ashlagi et al., 2016).

[2]Indeed, removing the incentives to "game the system" was a key factor in the city of Boston's decision to replace its school assignment mechanism in 2006 (Abdulkadiroğlu et al., 2006).

[3]For instance, this notion separates sealed-bid second-price auctions from ascending auctions (where bidders only need to decide at any given moment whether to quit or not) and provides a possible explanation as to why more subjects have been reported to behave insincerely in the former than in the latter (Kagel et al., 1987).

proof.[4]) Generalizing to allow for weaker forms of alignment of women's preferences, we show that if women's preferences are acyclical (Ergin, 2002),[5] then the men-optimal stable matching can be implemented via an obviously strategy-proof mechanism. While the obvious truthfulness of the basic questions that we use to construct this implementation (questions of the form "do you prefer x the most out of all currently unmatched women?") draws from the same intuition upon which the serial dictatorship mechanism is based, the questions are considerably more flexible, and the order of the questions more subtle.

The main finding of this paper is that for general preferences, no mechanism that implements the men-optimal stable matching (or any other stable matching) is obviously strategy-proof for men. We first prove this impossibility in a specifically crafted matching market with 3 women and 3 men, in which women have fixed (cyclical) commonly known preferences and men have unrestricted private preferences. It is then shown that for the impossibility to hold in any market, it is sufficient for some 3 women to have this structure of preferences over some 3 men. Moreover, the same result holds even if women's preferences are privately known. An immediate implication of these results is that in a large market, in which women's preferences are drawn independently and uniformly at random, with high probability no implementation of any stable mechanism is obviously strategy-proof for all men (or even for most men).

To summarize, this paper finds that no stable matching mechanism is obviously strategy-proof for men as long as women's preferences are "sufficiently unaligned." The results apply to school choice settings even when schools are not strategic and simply have priorities over students. For example, unless schools' priorities over students are sufficiently aligned, no mechanism that is stable with respect to students' preferences and schools' priorities is obviously strategy-proof for students.

This paper sheds more light on fundamental differences between two-sided market mechanisms, which aim to implement a two-sided notion such as stability, and closely related one-sided market mechanisms, which aim to implement some efficiency notion for one side of the market. First, as noted, in assignment markets there exists an obviously strategy-proof ex-post efficient mechanism (serial dictatorship). Second, a variety of ascending auctions, from familiar multi-item auctions (Demange et al., 1986) to recently proposed clock auctions

[4]Since, after selecting an object, the agent quits the game, no contingent reasoning is needed in order to verify that she must ask for her favorite unallocated object. However, serial dictatorship (the same strategy-proof social choice rule), when implemented by having each agent simultaneously submit a ranking over all objects in advance, is not obviously strategy-proof. This example and the example in Footnote 3 both demonstrate that whereas strategy-proofness is a property of the social choice rule, obvious strategy-proofness is a property of the mechanism implementing the social choice rule.

[5]A preference profile for a woman over men is cyclical if there are three men a, b, c and two women x, y such that $a \succ_x b \succ_x c \succ_y a$.

(Milgrom and Segal, 2014), maximize welfare or revenue and are obviously strategy-proof, despite the latter's being based on deferred acceptance principles. In contrast, this paper shows that there is no way to achieve stability that is obvious for either side of the market.

Obvious strategy-proofness was introduced by Li (2015), who studied this property extensively in mechanisms with monetary transfers. In settings without transfers, Li studied this property in implementations of serial dictatorship and top trading cycles. Since the first online draft of our paper, quite a few papers on obvious strategy-proofness have subsequently followed: Perhaps most relevant to this paper is the work of Troyan (2016), who considers a one-sided market with agents and objects, and asks which preferences of the houses allow for an obviously strategy-proof implementation (for the agents) of the (Pareto efficient, not necessarily stable) top trading cycles algorithm (see the discussion concluding Section 4 below). Farther technically but nonetheless continuing in the spirit of search for obviously strategy-proof implementations of social choice rules that guarantee some attractive desideratum of the outcome, Bade and Gonczarowski (2016) constructively characterize Pareto-efficient (rather than stable) social choice rules that admit obviously strategy-proof implementations under three popular domains (house matching, single-peaked preferences, and combinatorial auctions). Also of note, Pycia and Troyan (2016) characterize general obviously strategy-proof mechanisms without transfers under a "richness" assumption on the preferences domain, and characterize the sequential version of random serial dictatorship under such an assumption via a natural set of axioms that includes obvious strategy-proofness. All three of these papers also utilize machinery and observations that originated in our paper.

The paper is organized as follows. Section 2 provides the model and background, including the definition of obvious strategy-proofness in matching markets. Section 3 presents special cases for which an obviously strategy-proof implementation of the men-optimal stable matching exists. Section 4 provides the main impossibility result. Section 5 presents corollaries in a model where women's preferences are private and are not fixed in advance. We conclude in Section 6.

2 Preliminaries

2.1 Matching with one-sided preferences

For the bulk of our analysis it will be sufficient to consider (two-sided) markets in which only one side of the market is strategic. We begin by defining the notions of matching and strategy-proofness in such markets.

In a (one-sided) matching market, the participants are partitioned into a finite set of *men* M and a finite set of *women* W. A *preference list* (for some man m) over W is a

totally ordered subset of W (if some woman w does not appear on the preference list, we think of her as being unacceptable to m). Denote the set of all preference lists over W by $\mathcal{P}(W)$. A *preference profile* $\bar{p} = (p_m)_{m \in M}$ for M over W is a specification of a preference list p_m over W for each man $m \in M$. (So the set of all preference profiles for M over W is $\mathcal{P}(W)^M$.) Given a preference list p_m for some man m, we write $w \succ_m w'$ to denote that man m strictly prefers woman w over woman w', (i.e., either woman w is ranked higher than w' on m's preference list, or w appears on this list while w' does not), and write $w \succeq_m w'$ if it is not the case that $w' \succ_m w$.

A *matching* between M and W is a one-to-one mapping between a subset of M and a subset of W. Denote the set of all matchings between M and W by $\mathcal{M}$. Given a matching μ between M and W, for a participant $a \in M \cup W$ we write μ_a to denote a's match in μ, or write $\mu_a = a$ if a is unmatched.

A (one-sided) *matching rule* is a function $C : \mathcal{P}(W)^M \to \mathcal{M}$, from preference profiles for M over W to matchings between M and W.

A matching rule C is said to be *strategy-proof* for a man m if for every preference profile $\bar{p} = (p_m)_{m \in M} \in \mathcal{P}(W)^M$ and for every (alternate) preference list $p'_m \in \mathcal{P}(W)$, it is the case that $C_m(\bar{p}) \succeq_m C_m(p'_m, \bar{p}_{-m})$ according to p_m.[6] C is said to be *strategy-proof* if it is strategy-proof for every man.

2.2 Obvious strategy-proofness

This section briefly describes the notion of obvious strategy-proofness, developed in great generality by Li (2015). We rephrase these notions for the special case of deterministic matching mechanisms with finite preference and outcome sets. For ease of presentation, attention is restricted to mechanisms under complete information; however, the results in this paper still hold (*mutatis mutandis*) via the same proofs for the general definitions of Li (2015).[7]

Whereas strategy-proofness is a property of a given matching rule, obvious strategy-proofness is a property of a specific implementation, via a specific mechanism, of such a matching rule. A mechanism implements a matching rule by specifying, roughly speaking, an extensive-form game tree that implements the standard-form game associated (where

[6] As is customary, $(p'_m, \bar{p}_{-m})$ denotes the preference profile obtained from $\bar{p}$ by setting the preferences of m to be p'_m.

[7] Readers who are familiar with the general definitions of Li (2015) may easily verify that if a randomized stable obviously strategy-proof (OSP) mechanism exists, then derandomizing it by fixing in advance each choice of nature to some choice made with positive probability yields a deterministic stable OSP mechanism. Furthermore, if some stable mechanism is OSP under partial information, then it is also OSP under complete information.

strategies coincide with preferences) with the matching rule, where each action at each node of the extensive-form game tree corresponds to some set of possible preferences for the acting participant. We now formalize this definition.

Definition 1 (matching mechanism)**.** A (one-sided extensive-form) *matching mechanism* for M over W consists of:

1. A rooted tree T.
2. A map $X : L(T) \to \mathcal{M}$ from the leaves of T to matchings between M and W.
3. A map $Q : V(T) \setminus L(T) \to M$, from internal nodes of T to M.
4. A map $A : E(T) \to 2^{\mathcal{P}(W)}$, from edges of T to predicates over $\mathcal{P}(W)$, such that both of the following hold:
   - The predicates corresponding to edges outgoing from the same node are disjoint.
   - The disjunction (i.e., set union) of all predicates corresponding to edges outgoing from a node n equals the predicate corresponding to the last edge outgoing from a node labeled $Q(n)$ along the path from the root to n, or to the predicate matching all elements of $\mathcal{P}(W)$ if no such edge exists.

A preference profile $\bar{p} \in \mathcal{P}(W)^M$ is said to *pass through* a node $n \in V(T)$ if, for each edge e along the path from the root to n, it is the case that $p_{Q(n')} \in A(e)$, where n' is the source node of e.

Definition 2 (implemented matching rule)**.** Given an extensive-form matching mechanism $\mathcal{I}$, we denote by $C^{\mathcal{I}}$, called the matching rule *implemented by* $\mathcal{I}$, the (one-sided) matching rule mapping a preference profile $\bar{p} \in \mathcal{P}(W)^M$ to the matching $X(n)$, where n is the unique leaf through which $\bar{p}$ passes. Equivalently, n is the node in T obtained by traversing T from its root, and from each node n' following the edge outgoing from n' whose predicate matches the preference list of $Q(n')$.

Two preference lists $p, p' \in \mathcal{P}(W)$ are said to *diverge* at a node $n \in V(T)$ if there exist two distinct edges e, e' outgoing from n such that $p \in A(e)$ and $p' \in A(e')$.

Definition 3 (obvious strategy-proofness (OSP))**.** Let $\mathcal{I}$ be an extensive-form matching mechanism.

1. $\mathcal{I}$ is said to be *obviously strategy-proof (OSP) for a man* $m \in M$ if for every node n with $Q(n) = m$ and for every $\bar{p} = (p_{m'})_{m' \in M} \in \mathcal{P}(W)^M$ and $\bar{p}' = (p'_{m'})_{m' \in M} \in \mathcal{P}(W)^M$ that both pass through n such that p_m and p'_m diverge at n, it is the case that $C^{\mathcal{I}}_m(\bar{p}) \succeq_m$

$C^{\mathcal{I}}_m(\bar{p}')$ according to p_m. In other words, the worst possible outcome for m when acting truthfully (i.e., according to p_m) at n is no worse than the best possible outcome for m when misrepresenting his preference list to be p'_m at n.

2. $\mathcal{I}$ is said to be *obviously strategy-proof (OSP)* if it is obviously strategy-proof for every man $m \in M$.

Li (2015) shows that obviously strategy-proof mechanisms are, in a precise sense, mechanisms that can shown to implement strategy-proof matching rules under a cognitively limited proof model that does not allow for contingent reasoning. To observe how strategy-proofness of $C^{\mathcal{I}}$ for a man $m \in M$ is indeed a weaker condition than obvious strategy-proofness of $\mathcal{I}$ for m, note that $C^{\mathcal{I}}$ is strategy-proof for m if and only if for every node n with $Q(n) = m$ and for every $\bar{p} = (p_m)_{m \in M} \in \mathcal{P}(W)^M$ that passes through n and for every $\bar{p}'_m \in \mathcal{P}(W)$ that diverges from p_m at n,[8] it is the case that $C^{\mathcal{I}}_m(\bar{p}) \succeq_m C^{\mathcal{I}}_m(\bar{p})$ according to p_m.

Definition 4 (OSP-implementability)**.** A (one-sided) matching rule $C : \mathcal{P}(W)^M \to \mathcal{M}$ is said to be *OSP-implementable* if $C = C^{\mathcal{I}}$ for some obviously strategy-proof matching mechanism $\mathcal{I}$. In this case, we say that $\mathcal{I}$ *OSP-implements* C.

2.3 Stability

We proceed to describe a simplified version of stability in matching markets as introduced by Gale and Shapley (1962). While, as stated in Section 2.1, for the bulk of our analysis it is sufficient to consider markets in which only men are strategic, to define the notion of stability one must consider not only preferences for the (strategic) men, but also for the (nonstrategic) women. Women's preference lists and preference profiles are defined analogously with those of men. We continue to denote a preference profile for men by $\bar{p} = (p_m)_{m \in M} \in \mathcal{P}(W)^M$, while denoting a preference profile for women by $\bar{q} = (q_w)_{m \in M} \in \mathcal{P}(M)^W$.

Let $\bar{p}$ and $\bar{q}$ be preference profiles of men and women respectively. A matching μ is said to be *unstable* with respect to $\bar{p}$ and $\bar{q}$ if there exist a man m and a woman w each preferring the other over the partner matched to them by μ, or if some participant $a \in M \cup W$ is matched with some other participant not on a's preference list. A matching that is not unstable is said to be *stable*. Gale and Shapley (1962) showed that a stable matching exists with respect to every pair of preference profiles and, furthermore, that for every pair of preference profiles there exists an *M-optimal stable matching*, i.e., a stable matching such that each man weakly prefers his match in this stable matching over his match in any other stable matching.

We now relate the concept of stability to the one-sided matching rules and mechanisms defined in the previous sections. Let $\bar{q} \in \mathcal{P}(M)^W$ be a preference profile for W over M. A

[8] These conditions imply that $(p'_m, \bar{p}_{-m})$ also passes through n.

(one-sided) matching rule C is said to be *$\bar{q}$-stable* if for every preference profile $\bar{p} \in \mathcal{P}(W)^M$ for M over W, the matching $C(\bar{p})$ is stable with respect to $\bar{p}$ and $\bar{q}$. A (one-sided) matching mechanism is said to be *$\bar{q}$-stable* if the matching rule that it implements is $\bar{q}$-stable.

We denote by $C^{\bar{q}} : \mathcal{P}(W)^M \to \mathcal{M}$ the *M-optimal stable matching rule*, i.e., the (one-sided, $\bar{q}$-stable) matching rule mapping each preference profile for men $\bar{p}$ to the M-optimal stable matching with respect to $\bar{p}$ and $\bar{q}$. It is well known that $C^{\bar{q}}$ is strategy-proof for all men (Dubins and Freedman, 1981). Moreover, no other matching rule is strategy-proof for all men (Gale and Sotomayor, 1985).[9] In the notation of this paper:

Theorem 1 (Gale and Sotomayor, 1985; Chen et al., 2016)**.** *For every preference profile $\bar{q} \in \mathcal{P}(M)^W$ for W over M, no $\bar{q}$-stable matching rule $C \neq C^{\bar{q}}$ is strategy-proof.*

In this paper, we ask whether $C^{\bar{q}}$ is not only strategy-proof, but also OSP-implementable. (As it is the unique $\bar{q}$-stable matching rule, it is the only candidate for OSP-implementability.)

3 OSP-implementable special cases

Before stating our main impossibility result, we first review a few special cases in which $C^{\bar{q}}$, the M-optimal stable matching rule for fixed women's preferences $\bar{q}$, is in fact OSP-implementable. For simplicity, we describe all of these cases under the assumption that the market is balanced (i.e., that $|W| = |M|$) and that all preference lists are full (i.e., that each participant prefers being matched to anyone over being unmatched); generalizing each of the below cases for unbalanced markets or for preference lists that are not full is straightforward.[10] The first case we consider is that in which women's preferences are perfectly aligned.

Example 1 ($C^{\bar{q}}$ is OSP-implementable when women's preferences are perfectly aligned)**.** Let $q \in \mathcal{P}(M)$ and let $\bar{q} = (q)_{w \in W}$ be the preference profile in which all women share the same preference list q. $C^{\bar{q}}$ is OSP-implementable by the following serial dictatorship mechanism: ask the man most preferred according to $\bar{q}$ which woman he prefers most, and assign that woman to this man (in all leaves of the subtree corresponding to this response), ask the man second-most preferred according to $\bar{q}$ which woman he prefers most out of those not yet assigned to any man, and assign that woman to this man (in all leaves of the subtree corresponding to this response), etc. This mechanism can be shown to be OSP by the same reasoning that Li (2015) uses to show that serial dictatorship is OSP.

[9] For a more general result, see Chen et al. (2016).

[10] Indeed, asking any man whether he prefers being unmatched over being matched with any (remaining not-yet-matched) woman never violates obvious strategy-proofness.

Another noteworthy example is that of arbitrary preferences in a very small matching market.

Example 2 ($C^{\bar{q}}$ is OSP-implementable when $|M| = |W| = 2$)**.** When $|M| = |W| = 2$, $C^{\bar{q}}$ is OSP-implementable for every $\bar{q} \in \mathcal{P}(M)^W$. Indeed, let $M = \{a, b\}$ and $W = \{x, y\}$. If $q_x = q_y$, then $C^{\bar{q}}$ is OSP-implementable as explained in Example 1. Otherwise, without loss of generality $a \succ_x b$ and $b \succ_y a$; for this case, Figure 1 describes an OSP mechanism that implements $C^{\bar{q}}$.

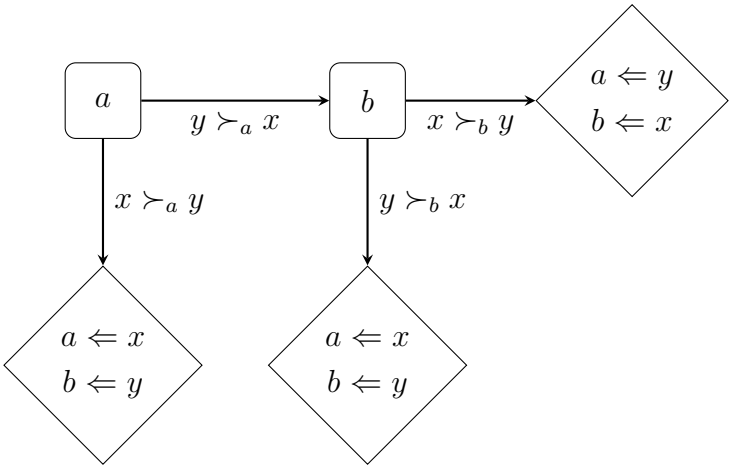

Figure 1: An OSP mechanism that implements $C^{\bar{q}}$ for $|W| = |M| = 2$ and for $\bar{q}$ where $a \succ_x b$ and $b \succ_y a$.

The preference profiles in Examples 1 and 2 are special cases of the class of acyclical preference profiles, whose structure was defined by Ergin (2002).

Definition 5 (acyclicality)**.** A preference profile $\bar{q} \in \mathcal{P}(M)^W$ for W over M is said to be *cyclical* if there exist $a, b, c \in M$ and $x, y \in W$ such that $a \succ_x b \succ_x c \succ_y a$. If $\bar{q}$ is not cyclical, then it is said to be *acyclical*.

Ergin (2002) shows that acyclicality of $\bar{q}$ is necessary and sufficient for $C^{\bar{q}}$ to be group strategy-proof (and not merely weakly group strategy-proof) and Pareto efficient. We now generalize Examples 1 and 2 by showing that acyclicality of $\bar{q}$ (as in both of these examples) is sufficient for $C^{\bar{q}}$ to be also OSP-implementable.

Theorem 2 (positive result for acyclical preferences)**.** *$C^{\bar{q}}$ is OSP-implementable for every acyclical preference profile $\bar{q} \in \mathcal{P}(M)^W$ for W over M.*

Proof sketch. We prove the result by induction over $|M| = |W|$. By acyclicality, at most two men are ranked by some woman as her top choice. If only one such man $m \in M$ exists,

then he is ranked by all women as their top choice—in this case, similarly to Example 1, we ask this man for his top choice $w \in W$, assign her to him, and then continue by induction (finding in an OSP manner the M-optimal stable matching between $M \setminus \{m\}$ and $W \setminus \{w\}$). Otherwise, there are precisely two men $a \in M$ and $b \in M$ who are ranked by some woman as her top choice. By acyclicality, each woman either has a as her top choice and b as her second-best choice, or *vice versa*.[11] We conclude somewhat similarly to Figure 1: for each woman $w \in W$ that prefers a most, we ask a whether he prefers w most; if so, we assign w to a and continue by induction. Otherwise, for each woman $w \in W$ that prefers b most, we ask b whether he prefers w most; if so, we assign w to b and continue by induction. Otherwise, we ask each of a and b for his top choice, assign each of them his top choice, and continue by induction. □

We conclude this section by noting, however, that acyclicality of $\bar{q}$ is not a necessary condition for OSP-implementability of $C^{\bar{q}}$, as demonstrated by the following example.

Example 3 (OSP-implementable $C^{\bar{q}}$ with cyclical $\bar{q}$)**.** Let $M = \{a, b, c\}$ and $W = \{x, y, z\}$. We claim that $C^{\bar{q}}$, for the following cyclical preference profile $\bar{q}$ (where each woman prefers being matched to any man over being unmatched), is OSP-implementable:

$$\begin{array}{ccccc} a & \succ_x & b & \succ_x & c \\ a & \succ_y & c & \succ_y & b \\ b & \succ_z & a & \succ_z & c. \end{array}$$

We begin by noting that $\bar{q}$ is indeed cyclical, as $a \succ_y c \succ_y b \succ_z a$. We now note that the following mechanism OSP-implements $C^{\bar{q}}$:

1. Ask a whether he prefers x the most; if so, assign x to a and continue as in Example 2 (finding in an OSP manner the M-optimal stable matching between $\{y, z\}$ and $\{b, c\}$).

2. Ask a whether he prefers y the most; if so, assign y to a and continue as in Example 2. (Otherwise, we deduce that 1) a prefers z the most and therefore 2) c will not end up being matched to z.)

3. Ask b whether he prefers z the most; if so, assign z to b and continue as in Example 2.

4. Ask b whether he prefers x the most; if so, assign x to b, z to a, and y to c. (Otherwise, we deduce that b prefers y the most.)

5. Ask c whether he prefers x over y. If so, assign x to c, y to b, and z to a. (Otherwise, we deduce that b will not end up being matched to y.)

[11] This is reminiscent of the priorities of the first two agents in bipolar serially dictatorial rules (Bogomolnaia et al., 2005), which are indeed included in the analysis of Theorem 2 as a special case.

6. Ask b whether he prefers z over x. Assign b to his preferred choice between z and x and continue as in Example 2.

Nonetheless, as we show in the next section, when there are more than 2 participants on each side and women's preferences are sufficiently unaligned, $C^{\bar{q}}$ is not OSP-implementable.

4 Impossibility result for general preferences

We now present our main impossibility result.

Theorem 3 (impossibility result for general preferences)**.** *If $|M| \geq 3$ and $|W| \geq 3$, then there exists a preference profile $\bar{q} \in \mathcal{P}(M)^W$, such that no $\bar{q}$-stable (one-sided) matching rule is OSP-implementable.*

Observe that Theorem 3 applies to any $\bar{q}$-stable (one-sided) matching rule, and not only to the M-optimal stable matching rule $C^{\bar{q}}$. Before proving the result, we first prove a special case that cleanly demonstrates the construction underlying our proof.

Lemma 1. *For $|M| = |W| = 3$, there exists a preference profile $\bar{q} \in \mathcal{P}(M)^W$ such that no $\bar{q}$-stable (one-sided) matching rule is OSP-implementable.*

Proof. Let $M = \{a, b, c\}$ and $W = \{x, y, z\}$. Let $\bar{q}$ be the following preference profile (where each woman prefers being matched to any man over being unmatched):

$$\begin{array}{lllll} a & \succ_x & b & \succ_x & c \\ b & \succ_y & c & \succ_y & a \\ c & \succ_z & a & \succ_z & b. \end{array} \tag{1}$$

Assume for contradiction that an OSP mechanism $\mathcal{I}$ that implements a $\bar{q}$-stable matching rule $C^{\mathcal{I}}$ exists. Therefore, $C^{\mathcal{I}}$ is strategy-proof, and so, by Theorem 1, $C^{\mathcal{I}} = C^{\bar{q}}$. In order to reach a contradiction by showing that such a mechanism (that OSP-implements $C^{\bar{q}}$) cannot possibly exist, we dramatically restrict the domain of preferences of all men, which results in a simpler mechanism, where the contradiction can be identified in a less cumbersome manner. We define:

$$\begin{array}{lll} p_a^1 \triangleq z \succ y \succ x & p_b^1 \triangleq x \succ z \succ y & p_c^1 \triangleq y \succ x \succ z \\ p_a^2 \triangleq y \succ x \succ z & p_b^2 \triangleq z \succ y \succ x & p_c^2 \triangleq x \succ z \succ y, \end{array}$$

and set $\mathcal{P}_a \triangleq \{p_a^1, p_a^2\}$, $\mathcal{P}_b \triangleq \{p_b^1, p_b^2\}$, and $\mathcal{P}_c \triangleq \{p_c^1, p_c^2\}$.

Following a proof technique in Li (2015), we note that if we "prune" the tree of $\mathcal{I}$ by replacing, for each edge e, the predicate $A(e)$ with the conjunction (i.e., set intersection) of $A(e)$ with the predicate matching all elements of $\mathcal{P}_{Q(n)}$, where n is the source node of e, and

by consequently deleting all edges e for which $A(e) = \bot$,[12] we obtain, in a precise sense, a mechanism that implements $C^{\bar{q}}$ where the preference list of each man $m \in M$ is *a priori* restricted to be in $\mathcal{P}_m$.[13] By a proposition in Li (2015), since the original mechanism $\mathcal{I}$ is OSP, so is the pruned mechanism as well.

Let n be the earliest (i.e., closest to the root) node in the pruned tree that has more than one outgoing edge (such a node clearly exists, since $C^{\mathcal{I}} = C^{\bar{q}}$ is not constant over $\mathcal{P}_a \times \mathcal{P}_b \times \mathcal{P}_c$). By symmetry of $\bar{q}, \mathcal{P}_a, \mathcal{P}_b, \mathcal{P}_c$, without loss of generality $Q(n) = a$. By definition of pruning, it must be the case that n has two outgoing edges, one labeled p_a^1, and the other labeled p_a^2. We claim that the mechanism of the pruned tree is in fact not OSP. Indeed, for $p_a = p_a^2$ (the "true preferences"), $p_b = p_b^2$, and $p_c = p_c^1$, we have that $C_a^{\mathcal{I}}(\bar{p}) = C_a^{\bar{q}}(\bar{p}) = x$, yet for $p'_a = p_a^1$ (a "possible manipulation"), $p'_b = p_b^1$, and $p'_c = p_c^2$, we have that $C_a^{\mathcal{I}}(\bar{p}') = C_a^{\bar{q}}(\bar{p}') = y$, even though $C_a^{\mathcal{I}}(\bar{p}') = y \succ_a x = C_a^{\mathcal{I}}(\bar{p})$ according to p_a (by definition of n, both $\bar{p}$ and $\bar{p}'$ pass through n, and p_a and p'_a diverge at n), and so the mechanism of the pruned tree indeed is not OSP — a contradiction. □

Proof sketch of Theorem 3. The proof follows from a reduction to Lemma 1. Indeed, let a, b, c be three distinct men and let x, y, z be three distinct women. Let $\bar{q} \in \mathcal{P}(W)^M$ be a preference profile such that the preferences of x, y, z satisfy Eq. (1) with respect to a, b, c (with arbitrary preferences over all other men), and with arbitrary preferences for all other women. Assume for contradiction that a $\bar{q}$-stable OSP mechanism $\mathcal{I}$ exists.

We prune (see the proof of Lemma 1 for an explanation of pruning) the tree of $\mathcal{I}$ such that the only possible preference lists for a, b, c are those in which they prefer each of x, y, z, over all other women, and the only possible preference list for all other men is empty.[14] Let $\bar{q}'$ be the preference profile given in Lemma 1; the resulting (pruned) mechanism is a $\bar{q}'$-stable matching mechanism for a, b, c over x, y, z,[15] and so, by Lemma 1, it is not OSP; therefore, by the same proposition in Li (2015) that is used in Lemma 1, neither is $\mathcal{I}$. □

As Theorem 3 shows, it is enough that *some three women* have preferences that satisfy Eq. (1) with respect to *some three men* in order for obvious strategy-proofness to be unattainable. This implies that obvious strategy-proofness in also unattainable in large random markets with high probability.

[12] The standard notation $\bot$ stands for "false" (mnemonic: an upside-down "true" $\top$), i.e., the predicate that matches nothing, so an edge for which $A(e) = \bot$ will never be followed.

[13] The definition of mechanisms and OSP when the domain of preferences is restricted extends naturally from that given in Section 2.2 for unrestricted preferences. The interested reader is referred to Appendix A for precise details.

[14] Alternatively, one could set for all other men arbitrary preference lists that do not contain x, y, z.

[15] Formally, it is a matching mechanism for W over M with respect to the pruned preferences, but can be shown to always leave all participants but a, b, c and x, y, z, unmatched, and so can be thought of as a matching mechanism for a, b, c over x, y, z.

Corollary 1 (impossibility result for random markets)**.** *If* $|M| \geq 3$ *and* $|W| \geq 3$*, then as* $|M| + |W|$ *grows, we have for* $\bar{q} \sim U\big(\mathcal{P}(M)^W\big)$ *that:*[16]

a. With high probability no $\bar{q}$*-stable (one-sided) matching rule is OSP-implementable.*

b. For every three distinct men $a, b, c \in M$*, as* $|W|$ *grows, with high probability no* $\bar{q}$*-stable (one-sided) matching mechanism is OSP for* a*,* b*, and* c*.*

c. If $|M| \leq \text{poly}\big(|W|\big)$*, then with high probability no* $\bar{q}$*-stable (one-sided) matching mechanism is OSP for more than two men.*

Corollary 1 follows from an argument similar to the one in the proof of Theorem 3. Indeed, our proof of Theorem 3 in fact shows that if $\bar{q}$ satisfies Eq. (1) with respect to three men a, b, c and three women x, y, z, then no $\bar{q}$-stable matching mechanism is OSP for a, b, and c. For Part c, for instance, we note that for a fixed triplet of distinct men $a, b, c \in M$, the probability that Eq. (1) is not satisfied by $\bar{q}$ with respect to a, b, c and any three women x, y, z decreases exponentially with $|W|$, while the number of triplets of men increases polynomially with $|M|$.

We conclude this section by noting that while the aesthetic preference profile defined in Eq. (1) is sufficient for proving Theorem 3 and even Corollary 1, it is by no means the unique preference profile that eludes an obviously strategy-proof implementation, even when $|M| = |W| = 3$. Indeed, Proposition 1 in Appendix B gives an additional example of such a preference profile, which could be described as "less cyclical," in some sense.[17] In this context, it is worth noting that following up on our paper, Troyan (2016) gives a necessary and sufficient condition, "weak acyclicality" (weaker, indeed, than acyclicality as defined in Definition 5), on the preferences of objects in the (Pareto efficient, not necessarily stable) top trading cycles algorithm for this algorithm to be OSP-implementable for the agents. The example given in Proposition 1 also demonstrates that Troyan's condition does not suffice for the existence of an OSP-implementable stable mechanism. A comparison of the respective preference profiles used for the positive result of Example 3 and the negative result of Proposition 1, noting that the former is obtained by taking the latter and arguably making it "more aligned" by modifying the preference list of woman x to equal that of woman y, suggests that an analogous succinct "maximal domain" characterization of preference profiles that admit OSP-implementable stable mechanisms may be delicate, and obtaining it may be challenging.

[16] This result also holds, with the same proof, if $\bar{q}$ is drawn uniformly at random from the set of all full preferences (i.e., where each woman prefers being matched to any man over being unmatched).

[17] While the proof of Proposition 1 also follows a pruning argument, the reasoning is more involved than the in proof given for Lemma 1 above.

5 Two-sided mechanisms

So far, this paper has studied one-sided matching markets, in which only men are strategic and women's preferences are commonly known. This allowed us to ask questions such as, For which preference profiles of women one can OSP-implement the M-optimal stable matching rule? This setting is furthermore practically relevant in school choice where, for example, schools do not act strategically but have priorities over students.

Our analysis, however, also immediately yields that in two-sided markets, i.e., where both men and women behave strategically, no stable matching mechanism is OSP-implementable. To formalize this result, we introduce a few definitions. A *two-sided matching rule* is a function $C : \mathcal{P}(W)^M \times \mathcal{P}(M)^W \to \mathcal{M}$, from preference profiles for both men and women to a matching between M and W. A two-sided matching rule C is *stable* if for any preference profiles $\bar{p}$ and $\bar{q}$ for men and women, $C(\bar{p}, \bar{q})$ is stable with respect to $\bar{p}$ and $\bar{q}$. A two-sided matching mechanism[18] is *stable* if the two-sided matching rule that it implements is stable. Theorem 3 implies the following impossibility result for two-sided matching mechanisms:

Corollary 2 (impossibility result for two-sided mechanisms)**.** *If $|M| \geq 3$ and $|W| \geq 3$, then no stable two-sided matching rule is OSP-implementable for M. Moreover, no stable two-sided matching mechanism is OSP for more than two men.*

As with Theorem 3, we note that Corollary 2 applies to any stable two-sided matching rule, and not only to the M-optimal stable matching. Similarly, Theorem 2 implies the following possibility result for two-sided matching mechanisms:

Corollary 3 (positive result for $|M| = 2$ for two-sided mechanisms)**.** *If $|M| = 2$, then the two-sided M-optimal stable matching rule (i.e., the two-sided matching rule mapping each pair of preference profiles to the corresponding M-optimal stable matching) is OSP-implementable (by first querying the women, and then, given their preferences, continuing as in Theorem 2).*

A precise argument that relates the one-sided and two-sided results is given in Appendix D.

6 Conclusion

This paper finds that no stable matching mechanism is obviously strategy-proof for participants on one side of the market, unless the preferences of the various participants on the other side are strongly aligned with each other. This suggests that the strategic mistakes observed

[18]The definition of mechanisms and OSP for two-sided markets extends naturally from that given in Section 2.2 for one-sided markets. The interested reader is referred to Appendix C for precise details.

in practice (Hassidim et al., 2016; Rees-Jones, 2016) may not be avoided by implementing the men-optimal stable matching via any other procedure. This highlights the importance of clearinghouses gaining the unwavering trust of participating agents, so that participants both act accordingly when they are advised that no strategic opportunities exist, and trust that the mechanism will be run as stated after preferences are collected.

For the case in which women's preferences are acyclical, we describe an OSP mechanism that implements the men-optimal stable matching. It is interesting to compare and contrast this mechanism with OSP mechanisms for auctions. In binary allocation problems, such as private-value auctions with unit demand, procurement auctions with unit supply, and binary public good problems, Li (2015) shows that in every OSP mechanism, each buyer chooses, roughly speaking, between a fixed option (i.e., quitting) and a "moving" option that is *worsening* over time (i.e., its price is increasing). In contrast, in the OSP mechanism that we construct for the men-optimal stable matching with acyclical women's preferences, each man m either is assigned his (current) top choice or chooses between a fixed option (i.e., being unmatched) and a "moving" option that is *improving* over time: choosing any woman who prefers m most among all remaining (yet-to-be-matched) men.

Bridging the negative and positive results via an exact, succinct characterization of how aligned the preference profile of the proposed-to side needs to be in order to support an obviously strategy-proof implementation remains an open question. A comparison of the respective preference profiles used for the positive result of Example 3 and the negative result of Proposition 1 (in Appendix B) suggests that such a succinct "maximal domain" characterization may be delicate, and obtaining it may be challenging.

Interestingly, while deferred acceptance is (even weakly group) strategy-proof and has an ascending flavor similar to that of ascending unit-demand auctions or clock auctions (which are all obviously strategy-proof), deferred acceptance is in fact not OSP-implementable. It seems that the fact that stability is a two-sided notion, in contrast with maximizing efficiency or welfare for one side, increases the difficulty of employing strategic reasoning over stable mechanisms. In this context, it is worth noting a line of work (Segal, 2007; Gonczarowski et al., 2015) that highlights a similar message in terms of complexity rather than strategic reasoning, by showing that the communication complexity of finding (or even verifying) an approximately stable matching is significantly higher than the communication complexity of approximate welfare maximization for one of the sides of the market (Dobzinski et al., 2014). Indeed, in more than one way, stability is not an "obvious" objective.

References

A. Abdulkadiroğlu, P. Pathak, A. E. Roth, and T. Sönmez. Changing the Boston school choice mechanism. Working paper 11965, National Bureau of Economic Research, 2006.

I. Ashlagi, Y. Kanoria, and J. D. Leshno. Unbalanced random matching markets: The stark effect of competition. *Journal of Political Economy*, 2016. Forthcoming.

S. Bade and Y. A. Gonczarowski. Gibbard-Satterthwaite success stories and obvious strategyproofness. Mimeo, 2016.

A. Bogomolnaia, R. Deb, and L. Ehlers. Strategy-proof assignment on the full preference domain. *Journal of Economic Theory*, 123(2):161–186, 2005.

P. Chen, M. Egesdal, M. Pycia, and M. B. Yenmez. Manipulability of stable mechanisms. *American Economic Journal: Microeconomics*, 8(2):202–214, 2016.

G. Demange, D. Gale, and M. Sotomayor. Multi-item auctions. *Journal of Political Economy*, 94(4):863–872, 1986.

S. Dobzinski, N. Nisan, and S. Oren. Economic efficiency requires interaction. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC 2014)*, pages 233–242, 2014. Full version available at arXiv:1311.4721.

L. E. Dubins and D. A. Freedman. Machiavelli and the Gale-Shapley algorithm. *American Mathematical Monthly*, 88(7):485–494, 1981.

H. I. Ergin. Efficient resource allocation on the basis of priorities. *Econometrica*, 70(6): 2489–2497, 2002.

D. Gale and L. S. Shapley. College admissions and the stability of marriage. *American Mathematical Monthly*, 69(1):9–15, 1962.

D. Gale and M. Sotomayor. Ms. Machiavelli and the stable matching problem. *American Mathematical Monthly*, 92(4):261–268, 1985.

Y. A. Gonczarowski, N. Nisan, R. Ostrovsky, and W. Rosenbaum. A stable marriage requires communication. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2015)*, pages 1003–1017, 2015.

A. Hassidim, A. Romm, and R. I. Shorrer. "Strategic" behavior in a strategy-proof environment. Mimeo, 2016.

N. Immorlica and M. Mahdian. Marriage, honesty, and stability. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2005)*, pages 53–62, 2005.

J. H. Kagel, R. M. Harstad, and D. Levin. Information impact and allocation rules in auctions with affiliated private values: A laboratory study. *Econometrica*, 55(6):1275–1304, 1987.

F. Kojima and P. A. Pathak. Incentives and stability in large two-sided matching markets. *American Economic Review*, 99(3):608–627, 2009.

S. Li. Obviously strategy-proof mechanisms. Mimeo, 2015.

P. Milgrom and I. Segal. Deferred-acceptance auctions and radio spectrum reallocation. In *Proceedings of the 15th ACM Conference on Economics and Computation (EC 2014)*, pages 185–186, 2014.

M. Pycia and P. Troyan. Obvious dominance and random priority. Mimeo, 2016.

A. Rees-Jones. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. Mimeo, 2016.

A. E. Roth. The economist as engineer: Game theory, experimentation and computation as tools for design economics. *Econometrica*, 70(4):1341–1378, 2002.

I. Segal. The communication requirements of social choice rules and supporting budget sets. *Journal of Economic Theory*, 136(1):341–378, 2007.

P. Troyan. Obviously strategyproof implementation of allocation mechanisms. Mimeo, 2016.

A Mechanisms with restricted domains

In this appendix, we explicitly adapt the definitions in Section 2.2 to a restricted domain of preferences, as used in the proof of Lemma 1. The differences from the definitions in Section 2.2 are marked with an underscore. We emphasize that these definitions, like those in Section 2.2, are also a special case of the definitions in Li (2015). For every $m \in M$, fix a subset $\mathcal{P}_m \subseteq \mathcal{P}(W)$. Furthermore, define $\mathcal{P} \triangleq \bigtimes_{m \in M} \mathcal{P}_m$.

Definition 6 (matching mechanism)**.** A (one-sided extensive-form) *matching mechanism* for M over W with respect to $\mathcal{P}$ consists of:

1. A rooted tree T.
2. A map $X : L(T) \to \mathcal{M}(M, W)$ from the leaves of T to matchings between M and W.

3. A map $Q : V(T) \setminus L(T) \to M$, from internal nodes of T to M.

4. A map $A : E(T) \to 2^{\mathcal{P}(W)}$, from edges of T to predicates over $\mathcal{P}(W)$, such that both of the following hold:

   - The predicates corresponding to edges outgoing from the same node are disjoint.
   - The disjunction (i.e., set union) of all predicates corresponding to edges outgoing from a node n equals the predicate corresponding to the last edge outgoing from a node labeled $Q(n)$ along the path from the root to n, or to the predicate matching all elements of $\underline{\mathcal{P}_{Q(n)}}$ if no such edge exists.[19]

A preference profile $\bar{p} \in \underline{\mathcal{P}}$ is said to *pass through* a node $n \in V(T)$ if, for each edge e along the path from the root to n, it is the case that $p_{Q(n')} \in A(e)$, where n' is the source node of e.

Definition 7 (implemented matching rule)**.** Given an extensive-form matching mechanism $\mathcal{I}$ with respect to $\mathcal{P}$, we denote by $C^{\mathcal{I}}$, called the matching rule *implemented by* $\mathcal{I}$, the (one-sided) matching rule mapping a preference profile $\bar{p} \in \underline{\mathcal{P}}$ to the matching $X(n)$, where n is the unique leaf through which $\bar{p}$ passes. Equivalently, n is the node in T obtained by traversing T from its root, and from each node n' following the edge outgoing from n' whose predicate matches the preference list of $Q(n')$.

Two preference lists $p, p' \in \mathcal{P}(W)$ are said to *diverge* at a node $n \in V(T)$ if there exist two distinct edges e, e' outgoing from n such that $p \in A(e)$ and $p' \in A(e')$.[20]

Definition 8 (obvious strategy-proofness (OSP))**.** Let $\mathcal{I}$ be an extensive-form matching mechanism with respect to $\mathcal{P}$.

1. $\mathcal{I}$ is said to be *obviously strategy-proof (OSP) for a man* $m \in M$ if for every node n with $Q(n) = m$ and for every $\bar{p} = (p_{m'})_{m' \in M} \in \underline{\mathcal{P}}$ and $\bar{p}' = (p'_{m'})_{m' \in M} \in \underline{\mathcal{P}}$ that both pass through n such that p_m and p'_m diverge at n, it is the case that $C^{\mathcal{I}}_m(\bar{p}) \succeq_m C^{\mathcal{I}}_m(\bar{p}')$ according to p_m. In other words, the worst possible outcome for m when acting truthfully (i.e., according to p_m) at n is no worse than the best possible outcome for m when misrepresenting his preference list to be p'_m at n.

2. $\mathcal{I}$ is said to be *obviously strategy-proof (OSP)* if it is obviously strategy-proof for every man $m \in M$.

[19] In particular, this implies that the predicates corresponding to edges outgoing from a node n are predicates over $\underline{\mathcal{P}_{Q(n)}}$.

[20] In particular, this implies that $p, p' \in \underline{\mathcal{P}_{Q(n)}}$.

B A "less cyclical" non-OSP-implementable example

In this appendix, we give an additional example of a preference profile $\bar{q} \in \mathcal{P}(M)^W$, for three women over three men, for which no $\bar{q}$-stable matching rule is OSP-implementable. This preference profile could be described, in some sense, as "less cyclical" than the one used above to drive the proof of the results of Section 4. (Indeed, as noted above, this non-OSP-implementable preference profile is obtained by taking the OSP-implementable preference profile from Example 3 and arguably making it "more aligned" by modifying the preference list of woman x to equal that of woman y.) While, similarly to the proof of Lemma 1, we show the impossibility of OSP-implementation of this example via a pruning argument, the reasoning in this argument is more involved than in the one in the proof given for Lemma 1 in Section 4.

Proposition 1. *For $|M| = |W| = 3$, no OSP mechanism implements a $\bar{q}$-stable (one-sided) matching rule, for the preference profile $\bar{q} \in \mathcal{P}(M)^W$ defined as follows (where each woman prefers being matched to any man over being unmatched):*

$$\begin{array}{llllll} a & \succ_x & c & \succ_x & b \\ a & \succ_y & c & \succ_y & b \\ b & \succ_z & a & \succ_z & c. \end{array}$$

Proof. The proof starts similarly to that of Lemma 1. Let $M = \{a, b, c\}$ and $W = \{x, y, z\}$. Let $\bar{q}$ be the above preference profile, and assume for contradiction that an OSP mechanism $\mathcal{I}$ that implements a $\bar{q}$-stable matching rule $C^{\mathcal{I}}$ exists. Therefore, $C^{\mathcal{I}}$ is strategy-proof, and so, by Theorem 1, $C^{\mathcal{I}} = C^{\bar{q}}$. In order to reach a contradiction we dramatically restrict the domain of preferences of all men, however in this proof to a slightly richer domain than in the proof of Lemma 1. We define:

$$\begin{array}{lll} p_a^1 \triangleq z \succ x \succ y & p_b^1 \triangleq y \succ z \succ x & p_c^1 \triangleq x \succ y \succ z \\ p_a^2 \triangleq z \succ y \succ x & p_b^2 \triangleq x \succ z \succ y & p_c^2 \triangleq y \succ x \succ z, \\ & p_b^3 \triangleq x \succ y \succ z & \end{array}$$

and set $\mathcal{P}_a \triangleq \{p_a^1, p_a^2\}$, $\mathcal{P}_b \triangleq \{p_b^1, p_b^2, p_b^3\}$, and $\mathcal{P}_c \triangleq \{p_c^1, p_c^2\}$.

Following a proof technique in Li (2015), we prune (see the proof of Lemma 1 for more details) the tree of $\mathcal{I}$ according to $\mathcal{P}_a, \mathcal{P}_b, \mathcal{P}_c$, to obtain a mechanism that implements $C^{\bar{q}}$ where the preference list of each man $m \in M$ is *a priori* restricted to be in $\mathcal{P}_m$. By a proposition in Li (2015), since the original mechanism $\mathcal{I}$ is OSP, so is the pruned mechanism as well.

Let n be the earliest (i.e., closest to the root) node in the pruned tree that has more than one outgoing edge (such a node clearly exists, since $C^{\mathcal{I}} = C^{\bar{q}}$ is not constant over $\mathcal{P}_a \times \mathcal{P}_b \times \mathcal{P}_c$).

While the lack of symmetry of $\bar{q}$ does requires a slightly longer argument compared to the proof of Lemma 1 to complete this proof (reasoning by cases according to $Q(n)$ below), what makes the reasoning in this argument more involved (see the reasoning in the case $Q(n) = b$ below) than in its counterpart in the proof of Lemma 1 is the fact that we have left possible three preference lists for man b.[21] We conclude the proof by reasoning by cases according to the identity of $Q(n)$, in each case obtaining a contradiction by showing that the pruned tree is in fact not OSP.

$Q(n) = a$ By definition of pruning, it must be the case that n has two outgoing edges, one labeled p_a^1, and the other labeled p_a^2. In this case, for $p_a = p_a^1$ (the "true preferences"), $p_b = p_b^1$, and $p_c = p_c^2$, we have that $C_a^{\mathcal{I}}(\bar{p}) = C_a^{\bar{q}}(\bar{p}) = x$, yet for $p'_a = p_a^2$ (a "possible manipulation"), $p'_b = p_b^2$, and $p'_c = p_c^2$, we have that $C_a^{\mathcal{I}}(\bar{p}') = C_a^{\bar{q}}(\bar{p}') = z$, even though $C_a^{\mathcal{I}}(\bar{p}') = z \succ_a x = C_a^{\mathcal{I}}(\bar{p})$ according to p_a (by definition of n, both $\bar{p}$ and $\bar{p}'$ pass through n, and p_a and p'_a diverge at n), and so the mechanism of the pruned tree indeed is not OSP — a contradiction.

$Q(n) = c$ By definition of pruning, it must be the case that n has two outgoing edges, one labeled p_c^1, and the other labeled p_c^2. In this case, for $p_c = p_c^1$ (the "true preferences"), $p_a = p_a^1$, and $p_b = p_b^2$, we have that $C_c^{\mathcal{I}}(\bar{p}) = C_c^{\bar{q}}(\bar{p}) = y$, yet for $p'_c = p_c^2$ (a "possible manipulation"), $p'_a = p_a^2$, and $p'_b = p_b^1$, we have that $C_c^{\mathcal{I}}(\bar{p}') = C_c^{\bar{q}}(\bar{p}') = x$, even though $C_c^{\mathcal{I}}(\bar{p}') = x \succ_c y = C_c^{\mathcal{I}}(\bar{p})$ according to p_c (by definition of n, both $\bar{p}$ and $\bar{p}'$ pass through n, and p_c and p'_c diverge at n), and so the mechanism of the pruned tree indeed is not OSP — a contradiction.

$Q(n) = b$ By definition of pruning, it must be the case that n has at least two outgoing edges, and therefore has at least one edge labeled by a singleton preference list p_b^i. We prove this case by reasoning by subcases according to the value of i.

i=1 In this case, for $p_b = p_b^i = p_b^1$ (the "true preferences"), $p_a = p_a^1$, and $p_c = p_c^2$, we have that $C_b^{\mathcal{I}}(\bar{p}) = C_b^{\bar{q}}(\bar{p}) = z$, yet for $p'_b = p_b^3$ (a "possible manipulation"), $p'_a = p_a^1$, and $p'_c = p_c^1$, we have that $C_b^{\mathcal{I}}(\bar{p}') = C_b^{\bar{q}}(\bar{p}') = y$, even though $C_b^{\mathcal{I}}(\bar{p}') = y \succ_b z = C_b^{\mathcal{I}}(\bar{p})$ according to p_b (by definition of n, both $\bar{p}$ and $\bar{p}'$ pass through n, and since $i = 1$ we have that $p_b = p_b^i$ and $p'_b \neq p_b^i$ diverge at n), and so the mechanism of the pruned tree indeed is not OSP — a contradiction.

[21] To our knowledge, the first instance of an impossibility-by-pruning proof with more than two possible preferences/types for any of the agents is in an impossibility result for OSP-implementation of combinatorial auctions in Bade and Gonczarowski (2016). While that paper is much newer than any other result in our paper, the first draft of that proof predated the proof given in this appendix.

i=2 In this case, for $p_b = p_b^i = p_b^2$ (the "true preferences"), $p_a = p_a^2$, and $p_c = p_c^1$, we have that $C_b^{\mathcal{I}}(\bar{p}) = C_b^{\bar{q}}(\bar{p}) = z$, yet for $p'_b = p_b^3$ (a "possible manipulation"), $p'_a = p_a^1$, and $p'_c = p_c^2$, we have that $C_b^{\mathcal{I}}(\bar{p}') = C_b^{\bar{q}}(\bar{p}') = x$, even though $C_b^{\mathcal{I}}(\bar{p}') = x \succ_b z = C_b^{\mathcal{I}}(\bar{p})$ according to p_b (by definition of n, both $\bar{p}$ and $\bar{p}'$ pass through n, and since $i = 2$ we have that $p_b = p_b^i$ and $p'_b \neq p_b^i$ diverge at n), and so the mechanism of the pruned tree indeed is not OSP — a contradiction.

i=3 In this case, for $p_b = p_b^i = p_b^3$ (the "true preferences"), $p_a = p_a^1$, and $p_c = p_c^1$, we have that $C_b^{\mathcal{I}}(\bar{p}) = C_b^{\bar{q}}(\bar{p}) = y$, yet for $p'_b = p_b^2$ (a "possible manipulation"), $p'_a = p_a^1$, and $p'_c = p_c^2$, we have that $C_b^{\mathcal{I}}(\bar{p}') = C_b^{\bar{q}}(\bar{p}') = x$, even though $C_b^{\mathcal{I}}(\bar{p}') = x \succ_b y = C_b^{\mathcal{I}}(\bar{p})$ according to p_b (by definition of n, both $\bar{p}$ and $\bar{p}'$ pass through n, and since $i = 3$ we have that $p_b = p_b^i$ and $p'_b \neq p_b^i$ diverge at n), and so the mechanism of the pruned tree indeed is not OSP — a contradiction. □

C Two-sided mechanisms

In this appendix, we explicitly adapt the definitions in Section 2.2 for two-sided mechanisms, where the participants include not only the men but also the women, as in Section 5. The differences from the definitions in Section 2.2 are marked with an underscore. We emphasize that these definitions, like those in Section 2.2, are also a special case of the definitions in Li (2015). Define $\mathcal{P} \triangleq \mathcal{P}(W)^M \times \mathcal{P}(M)^W$. For every two-sided preference profile $\bar{r} = (\bar{p}, \bar{q}) \in \mathcal{P}$, we write $r_m = p_m$ for every $m \in M$ and $r_w = q_w$ for every $w \in W$.

Definition 9 (two-sided matching mechanism)**.** A *two-sided* (extensive-form) *matching mechanism* for M and W consists of:

1. A rooted tree T.
2. A map $X : L(T) \to \mathcal{M}(M, W)$ from the leaves of T to matchings between M and W.
3. A map $Q : V(T) \setminus L(T) \to M \cup W$, from internal nodes of T to participants $M \cup W$.
4. A map $A : E(T) \to 2^{\mathcal{P}(W)} \cup 2^{\mathcal{P}(M)}$, from edges of T to predicates over $\mathcal{P}(W)$ or over $\mathcal{P}(M)$, such that both of the following hold:
   - The predicates corresponding to edges outgoing from the same node are disjoint.
   - The disjunction (i.e., set union) of all predicates corresponding to edges outgoing from a node n equals the predicate corresponding to the last edge outgoing from a node labeled $Q(n)$ along the path from the root to n, or, if no such edge exists,

to the predicate matching all elements of $\mathcal{P}(W)$ if $Q(n) \in M$ and all elements of $\mathcal{P}(M)$ if $Q(n) \in W$.[22]

A two-sided preference profile $\bar{r} \in \underline{\mathcal{P}}$ is said to *pass through* a node $n \in V(T)$ if, for each edge e along the path from the root to n, it is the case that $r_{Q(n')} \in A(e)$, where n' is the source node of e.

Definition 10 (implemented matching rule)**.** Given a two-sided extensive-form matching mechanism $\mathcal{I}$, we denote by $C^{\mathcal{I}}$, called the two-sided matching rule *implemented by* $\mathcal{I}$, the two-sided matching rule mapping a two-sided preference profile $\bar{r} \in \underline{\mathcal{P}}$ to the matching $X(n)$, where n is the unique leaf through which $\bar{r}$ passes. Equivalently, n is the node in T obtained by traversing T from its root, and from each node n' following the edge outgoing from n' whose predicate matches the preference list of $Q(n')$.

Two preference lists $r, r' \in \mathcal{P}(W) \underline{\cup \mathcal{P}(M)}$ are said to *diverge* at a node $n \in V(T)$ if there exist two distinct edges e, e' outgoing from n such that $r \in A(e)$ and $r' \in A(e')$.[23]

Definition 11 (obvious strategy-proofness (OSP))**.** Let $\mathcal{I}$ be a two-sided extensive-form matching mechanism. $\mathcal{I}$ is said to be *obviously strategy-proof (OSP) for a participant* $a \in M \underline{\cup W}$ if for every node n with $Q(n) = a$ and for every $\bar{r}, \bar{r}' \in \underline{\mathcal{P}}$ that both pass through n such that p_a and p'_a diverge at n, it is the case that $C^{\mathcal{I}}_a(\bar{r}) \succeq_a C^{\mathcal{I}}_a(\bar{r}')$ according to r_a. In other words, the worst possible outcome for a when acting truthfully (i.e., according to r_a) at n is no worse than the best possible outcome for a when misrepresenting his or her preference list to be r'_a at n.

Definition 12 (OSP-implementability)**.** A two-sided matching rule $C : \underline{\mathcal{P}} \to \mathcal{M}(M, W)$ is said to be *OSP-implementable* for a set of participants $A \subseteq M \underline{\cup W}$ if $C = C^{\mathcal{I}}$ for some two-sided matching mechanism $\mathcal{I}$ that is OSP for (every participant in) A.

D From one-sided to two-sided markets

The next lemma allows us to obtain results in the two-sided model from the results obtained in the one-sided model (as alluded to in the discussion opening Section 5, the converse is not as immediate, e.g., neither Theorem 2 nor Corollary 1 is an immediate corollary of results that are naturally stated for two-sided mechanisms/matching rules). Indeed, Corollaries 2 and 3 both follow via this lemma from the respective analogous one-sided results.

[22] In particular, this implies that the predicates corresponding to edges outgoing from a node n are predicates over $\mathcal{P}(W)$ if $Q(n) \in M$ and over $\mathcal{P}(M)$ if $Q(n) \in W$.

[23] In particular, this implies that $r, r' \in \mathcal{P}(W)$ if $Q(n) \in M$ and that $r, r' \in \mathcal{P}(M)$ if $Q(n) \in W$.

Lemma 2 (relation between one-sided and two-sided OSP mechanisms)**.** *For every $M' \subseteq M$, there exists a stable two-sided matching mechanism that is OSP for M' if and only if for every $\bar{q} \in \mathcal{P}(W)^M$ there exists a $\bar{q}$-stable one-sided matching mechanism that is OSP for M'.*

Proof sketch. $\Rightarrow$: Assume that there exists a stable two-sided matching mechanism $\mathcal{I}$ that is OSP for M', and let $\bar{q} \in \mathcal{P}(W)^M$. We prune (see the proof of Lemma 1 for an explanation of pruning) the tree of $\mathcal{I}$ such that the women's preference profile is fixed to be $\bar{q}$. The resulting (pruned) mechanism is a *one-sided* matching mechanism that is $\bar{q}$-stable and (by the same proposition in Li (2015) that is used in Lemma 1) OSP for M', as required.

$\Leftarrow$: Assume that for every $\bar{q} \in \mathcal{P}(M)^W$ there exists a $\bar{q}$-stable one-sided matching mechanism $\mathcal{I}^{\bar{q}}$ that is OSP for M'. We construct a stable *two-sided* matching mechanism $\mathcal{I}$ as follows: first ask all women, in some order, for all of their preferences; the leaves of the tree so far are thus in one-to-one correspondence with preference profiles $\bar{q} \in \mathcal{P}(M)^W$ that pass through them. Next, at each "interim leaf" $n^{\bar{q}}$ corresponding to a preference profile $\bar{q} \in \mathcal{P}(M)^W$ (that passes through it), construct a subtree that is identical to the tree of $\mathcal{I}^{\bar{q}}$, with $n^{\bar{q}}$ as its root. It is straightforward to verify that the fact that each $\mathcal{I}^{\bar{q}}$ is $\bar{q}$-stable and OSP for M' implies that $\mathcal{I}$ is stable and OSP for M'. $\square$