Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 177 (2018) 405–425

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

Stable matching mechanisms are not obviously strategy-proof [☆]

Itai Ashlagi [a,*], Yannai A. Gonczarowski [b,c,*]

[a] *Management Science & Engineering, Stanford University, United States of America*
[b] *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, Israel*
[c] *Microsoft Research, Israel*

Received 30 July 2017; final version received 14 March 2018; accepted 9 July 2018

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 that was introduced by Li (2017). 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.

JEL classification: D4; D47

Keywords: Stable matching; Obviously strategy-proof; Obvious strategy-proofness; Matching; Mechanism

[☆] This paper greatly benefited from discussions with Sophie Bade, Elchanan Ben-Porath, 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.

[*] Corresponding authors.
E-mail addresses: iashlagi@stanford.edu (I. Ashlagi), yannai@gonch.name (Y.A. Gonczarowski).

https://doi.org/10.1016/j.jet.2018.07.001
0022-0531/© 2018 Elsevier Inc. All rights reserved.

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 agents on both sides of the market (or agents on one side of the market and objects on the other side of 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.[1]

Indeed, many clearinghouses have adopted in recent years the remarkable deferred acceptance (DA) mechanism (Gale and Shapley, 1962),[2] 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).[3,4] 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., 2017); this was observed in experiments (Chen and Sönmez, 2006), in surveys (Rees-Jones, 2016), and in the field (Hassidim et al., 2016; Shorrer and Sóvágó, 2017). This paper asks whether one can implement the deferred acceptance outcome via a mechanism whose description makes its strategy-proofness more apparent. Toward this goal, we adopt the notion of *obvious strategy-proofness*, an incentive property introduced by Li (2017) that is stronger than strategy-proofness.

Li (2017) 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.[5] 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 remained an open problem.

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 main 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, 2017, who shows that in a two-sided assignment market with agents and objects, serial

[1] See also Pathak and Sönmez (2008), which finds that non-strategy-proof mechanisms favor sophisticated players over more naïve players.

[2] Examples include the National Resident Matching Program (Roth, 1984), as well as school choice programs in Boston (Abdulkadiroğlu et al., 2005) and New York (Abdulkadiroğlu et al., 2009) (see also Abdulkadiroğlu and Sönmez, 2003).

[3] This mechanism is also approximately strategy-proof for all participants in the market (Immorlica and Mahdian, 2005; Kojima and Pathak, 2009; Ashlagi et al., 2017).

[4] 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 2005 (Abdulkadiroğlu et al., 2006).

[5] 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).

dictatorship, when implemented sequentially, is obviously strategy-proof.[6]) Generalizing to allow for weaker forms of alignment of women's preferences, we show that if women's preferences are acyclical (Ergin, 2002),[7] 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). These results apply to school choice settings even when schools are not strategic and have commonly known 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 that aim to implement a two-sided notion such as stability, and closely related two-sided market mechanisms that aim to implement some efficiency notion for one of the sides 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 (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 obviously strategy-proof for either side of the market.

Obvious strategy-proofness was introduced by Li (2017), who studies this property extensively in mechanisms with monetary transfers. In settings without transfers, Li (2017) studies this property in implementations of serial dictatorship and top trading cycles. Several papers further study this property in different settings. Closely related is Troyan (2016), who studies two-sided markets with agents and objects and asks for which priorities for objects one can implement in an obviously strategy-proof manner the Pareto-efficient top trading cycles algorithm. Pycia and Troyan (2016) characterize general obviously strategy-proof mechanisms without transfers

[6] 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 5 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.

[7] 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 $a \succ_x b \succ_x c \succ_y a$.

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. Bade and Gonczarowski (2017) constructively characterize Pareto-efficient social choice rules that admit obviously strategy-proof implementations in popular domains (object assignment, single-peaked preferences, and combinatorial auctions). It is worth noting that all three of these papers utilize machinery and observations that originated in this 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 also have private preferences. Section 6 concludes.

2. Preliminaries

2.1. Two-sided matching with one strategic side

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 two-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-side-querying) *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.[8] 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 (2017). We rephrase these notions for the special case of deterministic match-

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

ing mechanisms with finite preference and outcome sets. For ease of presentation, attention is restricted to mechanisms under perfect information; however, the results in this paper still hold (*mutatis mutandis*) via the same proofs for the general definitions of Li (2017).[9]

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

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

1. A rooted tree T. The nodes/vertices of the tree are denoted by $V(T)$. The edges of the tree are denoted by $E(T)$ and are directed away from the root: if an edge e is incident with a node n but is not on the path from the root of the tree to n, then e is outgoing from n. The leaves (nodes with no outgoing edges) of the tree are denoted by $L(T) \subset V(T)$.
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 all of the following hold:
   - Each such predicate must match at least one element in $\mathcal{P}(W)$.
   - 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 of T to n, it is the case that $p_{Q(n')} \in A(e)$, where n' is the source node of e. That is, the nodes through which $\bar{p}$ passes are the nodes of the path that starts from the root of T and follows, from each internal node n' that it reaches, the unique outgoing edge whose predicate matches the preference list of $Q(n')$.

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-side-querying) 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 internal node n' that is reached, following the unique outgoing edge whose predicate matches the preference list of $Q(n')$.

[9] Readers who are familiar with the general definitions of Li (2017) 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 perfect information.

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 (2017) shows that obviously strategy-proof mechanisms are, in a precise sense, mechanisms that can shown to implement strategy-proof rules under a cognitively limited proof model that does not allow for contingent reasoning. To observe how strategy-proofness of the matching rule $C^{\mathcal{I}}$ for a man $m \in M$ is indeed a weaker condition than obvious strategy-proofness of the mechanism $\mathcal{I}$ for m, note that the matching rule $C^{\mathcal{I}}$ 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 $p'_m \in \mathcal{P}(W)$ that diverges from p_m at n,[10] it is the case that $C^{\mathcal{I}}_m(\bar{p}) \succeq_m C^{\mathcal{I}}_m(p'_m, \bar{p}_{-m})$ according to p_m.[11]

Definition 4 *(OSP-implementability).* A (one-side-querying) 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 preferences (sometimes called priorities) for the (nonstrategic) women. Women's preference lists and preference profiles

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

[11] We emphasize that this rephrased definition is equivalent to the definition of strategy-proofness of the matching rule $C^{\mathcal{I}}$ that is given in Section 2.1, however it is not equivalent to standard definition of strategy-proofness of the extensive-form game underlying the mechanism $\mathcal{I}$, which would allow each man to condition the type he is "pretending to be" under any strategy on the information revealed by other men in preceding nodes. Once we move to the realm of obvious strategy-proofness, the restriction on each strategy to always consistently "pretend to be" of the same type is inconsequential, as the definition of OSP considers the case in which other men may play different types when the man in question acts truthfully or deviates. It is for this reason that we have chosen to implicitly define a strategy in the extensive-form game underlying $\mathcal{I}$ to be restricted to consistently "pretending to be" of the same type. This somewhat nonstandard implicit definition of a strategy considerably simplifies notation throughout this paper (by considering only consistent behavior on behalf of every agent) without changing the mathematical meaning of obvious strategy-proofness (or of strategy-proofness of a matching rule) and without limiting the generality of our results.

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-side-querying) 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 (one-side-querying) 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-side-querying) 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-side-querying, $\bar{q}$-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).[12] 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 strategy-proof $\bar{q}$-stable matching rule, it is the only candidate for OSP-implementability.)

3. OSP-implementable special cases

Before stating our main impossibility result, we first present a few special cases in which $C^{\bar{q}}$, the M-optimal stable matching rule for a fixed women's preference profile $\bar{q}$, is in fact OSP-implementable. These are the first known OSP mechanisms without transfers that are not dictatorial.[13]

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 for men that are not full is straightforward.[14] The first case we consider is that in which women's preferences are perfectly aligned.

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

[13] All OSP mechanisms that are surveyed in the end of the introduction are based upon the query structure of the mechanisms of this Section 3.

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

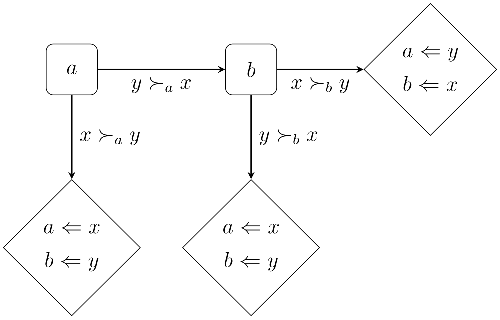

Fig. 1. An OSP mechanism that implements $C^{\bar{q}}$ for $|W| = |M| = 2$ and for $\bar{q}$ where $a \succ_x b$ and $b \succ_y a$. (The notation, e.g., $a \Leftarrow x$, indicates that x is matched to a in the matching corresponding to that leaf of the mechanism tree.)

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 for W over M 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 (2017) uses to show that serial dictatorship is OSP.

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 preference profile $\bar{q} \in \mathcal{P}(M)^W$ for W over M. 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, Fig. 1 describes an OSP mechanism that implements $C^{\bar{q}}$.

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 strongly 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. Much like the implementations in Examples 1 and 2, the strategy-proofness of the OSP implementation that emerges for acyclical preferences is far easier to understand than that of the standard deferred-acceptance implementation, thus showcasing the usefulness of obvious strategy-proofness in identifying easy-to-understand implementations. In each mechanism step, either a single man is given free pick out of all remaining $w \in W$, or two men are each given first priority over some subset of W (i.e., free pick if his favorite remaining $w \in W$ is there), and second priority over the rest (i.e., free pick out of all other remaining $w \in W$ except the one chosen by the other man if the latter invoked his first priority).

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. 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*.[15] We conclude somewhat similarly to Fig. 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.

To see that this implementation is OSP, consider a man $m \in M$ who is asked by this mechanism whether a woman $w \in W$ is his top choice (among the remaining women). If m really does prefer w most, then answering truthfully matches him to w, which he weakly prefers over any outcome that occurs if he is not truthful. Similarly, if m does not prefer w most, then answering truthfully may get m a more preferred choice, but also assures m that if he does not get such a preferred choice, then he would still be able to choose to get matched to w (he would do so if he fails to get his top choice, and w is his second-best); so, any outcome that results from truthfulness is weakly preferred by m over any outcome that results from nontruthfulness in this case as well. □

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 $\bar{q}$ for W over M (where each woman prefers being matched to any man over being unmatched), is OSP-implementable:

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

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 $\{b, c\}$).

[15] 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.

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.)
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 $\bar{q} \in \mathcal{P}(M)^W$ for W over M, such that no $\bar{q}$-stable (one-side-querying) matching rule is OSP-implementable.*

Observe that Theorem 3 applies to any $\bar{q}$-stable (one-side-querying) 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 $\bar{q} \in \mathcal{P}(M)^W$ for W over M such that no $\bar{q}$-stable (one-side-querying) 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):

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

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 the "pruning" technique in Li (2017), 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 $A(e) = \bot$,[16] 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 $\mathcal{P}_m$.[17] By a proposition in Li (2017), 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 of Theorem 3. The theorem 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 Equation (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.[18] 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,[19] and so, by Lemma 1, it is not OSP; therefore, by the same proposition in Li (2017) 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 Equation (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.

Corollary 1 *(impossibility result for random markets). If $|M| \geq 3$ and $|W| \geq 3$, then as $|M| + |W|$ grows, we have for a randomly drawn preference profile $\bar{q} \sim U\big(\mathcal{P}(M)^W\big)$ for W over M that*[20]*:*

[16] 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.

[17] 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.

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

[19] 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.

[20] 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).

a. *With high probability no $\bar{q}$-stable (one-side-querying) 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-side-querying) 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-side-querying) 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 Equation (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 Equation (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 Equation (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.[21] 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.

5. Matching with two strategic sides

So far, this paper has studied two-sided matching markets in which only men are strategic and women’s preference lists 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 when both men and women behave strategically, no stable matching mechanism is OSP-implementable. To formalize this result, we introduce a few definitions. A *two-sides-querying 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-sides-querying 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-sides-querying matching

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

mechanism[22] is *stable* if the two-sides-querying matching rule that it implements is stable. Theorem 3 implies the following impossibility result for two-sides-querying matching mechanisms:

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

As with Theorem 3, we note that Corollary 2 applies to any stable two-sides-querying matching rule, and not only to the *M-optimal two-side-querying stable matching rule* (i.e., the two-sides-querying matching rule that maps each pair of preference profiles to the corresponding M-optimal stable matching). Similarly, Theorem 2 implies the following possibility result for two-sides-querying matching mechanisms:

Corollary 3 *(positive result for $|M| = 2$ for two-sides-querying mechanisms). If $|M| = 2$, then the two-sides-querying M-optimal stable matching rule is OSP-implementable (by first querying the women, and then, given their preferences, continuing as in Theorem 2).*

A precise argument that relates the results for markets with one strategic side and those for markets with two strategic sides is given in Appendix D.

6. Discussion

This paper finds that no stable matching mechanism is obviously strategy-proof for the participants even on one of the sides of the market. This suggests that there may not be any alternative way to describe the deferred acceptance procedure that makes its strategy-proofness more apparent, implying that strategic mistakes observed in practice (Chen and Sönmez, 2006; Rees-Jones, 2016; Hassidim et al., 2016; Shorrer and Sóvágó, 2017) may not be avoidable by better explaining the mechanism. This highlights the importance of gaining the trust of the agents who participate in stable mechanisms, so that they both act as advised (even when it is hard to verify that no strategic opportunities exist) and are assured that the social planner will not deviate from the prescribed procedure after preferences are elicited.

For the case in which women's preferences are acyclical, we describe an OSP mechanism that implements the men-optimal stable matching. As may be expected, the strategy-proofness of this OSP implementation is easier to understand than that of deferred acceptance. 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 (2017) 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

[22] The definition of mechanisms and OSP for markets where both sides are strategic extends naturally from that given in Section 2.2 for markets where only one side is strategic. The interested reader is referred to Appendix C for precise details.

m most among all yet-to-be-matched men. This novel construction has come to be utilized by various OSP implementations, such as all of those that are surveyed in the end of the introduction.

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.[23]

Interestingly, while deferred acceptance is 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 objective (concerning the preferences of agents on both sides of the market), 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 (measured in the number of messages) 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.

While direct-revelation stable mechanisms are ubiquitous, there is growing usage of sequential-like implementations of deferred acceptance (or close variations thereof).[24] Our results imply that none of these variants, however presented to students and however conducted, can be OSP. Moreover, when DA is implemented sequentially according to its traditional description, sincere behavior is no longer even a dominant strategy but only induces an ex-post equilibrium (Bó and Hakimov, 2016a).[25] Nonetheless, seemingly contrasting these theoretical results, experimental evidence shows that such a sequential implementation of DA leads more often to sincere behavior and stable outcomes than the static implementation (Bó and Hakimov, 2016b; Pais et al., 2016). While sequential-like implementations do not possess stronger incentive properties than static implementations (and sometimes even possess weaker incentive properties), sequential-like implementations do ease the cognitive tasks of participants in various ways: they simplify strategic interactions by allowing students to break-down their decisions into smaller decisions that each requires somewhat less contingent reasoning than in the static implementation (as it is taken after receiving more information and feedback, such as the updated cutoff at each college); they allow students to focus on their next choice rather than to dwell on tentative choices that may never be reached; and they reduce the necessary preference communication and preference learning due to the information that is released throughout the mechanism (Bó and Hakimov, 2016a; Ashlagi et al., 2017). Our findings formally demonstrate

[23] While a technical challenge, we find it unlikely that resolving this problem will yield interesting economic insights.

[24] These include college admissions in Brazil (Bó and Hakimov, 2016a), Inner Mongolia (Chen and Pereyra, 2015; Gong and Liang, 2016), and Tunisia (Luflade, 2017), and school choice in Wake County (Dur et al., 2018). These implementations differ in various dimensions including the type of information provided to students, the timing, and how students can revise their choices; such differences may very well impact the students' behavior and therefore the outcome.

[25] Bó and Hakimov (2016a) require that only rejected agents may revise their proposals at each step in order to eliminate possible manipulations that appeared in the mechanism for college admissions in Brazil.

the cognitive complexity of reasoning in stable mechanisms; this suggests a possible explanation as to why, within the context of such mechanisms, the benefits from reducing cognitive load that are offered by sequential implementations outweigh the negative effects of the slightly weaker incentive properties of these implementations.[26] In a sense, in the absence of an OSP mechanism to reduce the cognitive load while strengthening the incentive properties, the "next best thing" may well be an extensive-form mechanism that eases cognitive load in different manners than OSP mechanisms, and moreover gives students a "feeling" similar to that of OSP mechanisms by not relinquishing control to the mechanism and by being able to constantly witness that the mechanism is run as promised.[27]

Appendix 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 $\underline{\text{underscore}}$. We emphasize that these definitions, like those in Section 2.2, are also a special case of the definitions in Li (2017). 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-side-querying extensive-form) *matching mechanism* for M over W $\underline{\text{with respect to } \mathcal{P}}$ 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 all of the following hold:
   - Each such predicate must match at least one element in $\mathcal{P}(W)$.
   - 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.[28]

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 of T to n, it is the case that $p_{Q(n')} \in A(e)$, where n' is the source node of e. That is, the nodes through which $\bar{p}$ passes are the nodes of the path that starts from the root of T and follows, from each internal node n' that it reaches, the unique outgoing edge whose predicate matches the preference list of $Q(n')$.

[26] The impact on students' welfare (which is of major importance) is beyond the scope of this paper, but see, for example, Luflade (2017); Dur et al. (2018).

[27] For a recent definition of a very strong sense of witnessing that the mechanism is run as promised, see Akbarpour and Li (2017).

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

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-side-querying) 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 internal node n' that is reached, following the unique outgoing edge 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 $p' \in A(e')$.[29]

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 $m \in M$.

Appendix 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-side-querying) matching rule, for the following preference profile $\bar{q} \in \mathcal{P}(M)^W$ for M over W (where each woman prefers being matched to any man over being unmatched):*

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

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 $\mathcal{I}$

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

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 (2017), 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 (2017), 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.[30] 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.

[30] To our knowledge, the first instance of an impossibility-by-pruning proof with more than two possible preference lists/types for any of the agents is in an impossibility result for OSP-implementation of combinatorial auctions in Bade and Gonczarowski (2017). 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.

$i=1$ 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.

$i=2$ 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. □

Appendix C. Two-sides-querying mechanisms

In this appendix, we explicitly adapt the definitions in Section 2.2 for two-sides-querying mechanisms, where the (strategic) 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 (2017). 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-sides-querying matching mechanism).* A two-sides-querying (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 $Q : V(T) \setminus L(T) \to \underline{M \cup W}$, from internal nodes of T to participants $\underline{M \cup W}$.
4. A map $A : E(T) \to 2^{\mathcal{P}(W)} \underline{\cup 2^{\mathcal{P}(M)}}$, from edges of T to predicates over $\mathcal{P}(W)$ or over $\mathcal{P}(M)$, such that all of the following hold:
   - Each such predicate must match at least one element in $\mathcal{P}(W)$ if $Q(n) \in M$ and at least one element in $\mathcal{P}(M)$ if $Q(n) \in W$.
   - 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 $Q(n) \in W$.[31]

A two-sides-querying 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 of T to n, it is the case that $r_{Q(n')} \in A(e)$, where n' is the source node of e. That is, the nodes through which $\bar{r}$ passes are the nodes of the path that starts from the root of T and follows, from each internal node n' that it reaches, the unique outgoing edge whose predicate matches the preference list of $Q(n')$.

Definition 10 *(implemented matching rule).* Given a two-sides-querying extensive-form matching mechanism $\mathcal{I}$, we denote by $C^{\mathcal{I}}$, called the two-sides-querying matching rule *implemented by* $\mathcal{I}$, the two-sides-querying matching rule mapping a two-sides-querying 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 internal node n' that is reached, following the unique outgoing edge whose predicate matches the preference list of $Q(n')$.

Two preference lists $r, r' \in \mathcal{P}(W) \cup \mathcal{P}(M)$ 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 $r' \in A(e')$.[32]

Definition 11 *(obvious strategy-proofness (OSP)).* Let $\mathcal{I}$ be a two-sides-querying extensive-form matching mechanism. $\mathcal{I}$ is said to be *obviously strategy-proof (OSP) for a participant $a \in M \cup W$* 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-sides-querying matching rule $C : \underline{\mathcal{P}} \to \mathcal{M}(M, W)$ is said to be *OSP-implementable* for a set of participants $A \subseteq M \cup W$ if $C = C^{\mathcal{I}}$ for some two-sides-querying matching mechanism $\mathcal{I}$ that is OSP for (every participant in) A.

Appendix D. From one strategic side to two strategic sides

The next lemma allows us to obtain results in the two-strategic-sides model from the results obtained in the one-strategic-side 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-sides-querying mechanisms/matching rules). Indeed, Corollaries 2 and 3 both follow via this lemma from the respective analogous results for one-side-querying mechanisms/matching rules.

Lemma 2 *(relation between one-side-querying and two-sides-querying OSP mechanisms). For every $M' \subseteq M$, there exists a stable two-sides-querying matching mechanism that is OSP for M'*

[31] 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 $Q(n) \in W$.

[32] 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 $Q(n) \in W$.

if and only if for every $\bar{q} \in \mathcal{P}(W)^M$ there exists a $\bar{q}$-stable one-side-querying matching mechanism that is OSP for M'.

Proof. $\Rightarrow$: Assume that there exists a stable two-sides-querying 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-side-querying* matching mechanism that is $\bar{q}$-stable and (by the same proposition in Li (2017) 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-side-querying matching mechanism $\mathcal{I}^{\bar{q}}$ that is OSP for M'. We construct a stable *two-sides-querying* matching mechanism $\mathcal{I}$ as follows: first ask all women, in some order, for all of their preference lists; 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'. □

References

Abdulkadiroğlu, A., Sönmez, T., 2003. School choice: a mechanism design approach. Am. Econ. Rev. 93 (3), 729–747.

Abdulkadiroğlu, A., Pathak, P.A., Roth, A.E., Sönmez, T., 2005. The Boston public school match. Am. Econ. Rev. 95 (2), 368–371.

Abdulkadiroğlu, A., Pathak, P., Roth, A.E., Sönmez, T., 2006. Changing the Boston School Choice Mechanism. Working paper 11965. National Bureau of Economic Research.

Abdulkadiroğlu, A., Pathak, P.A., Roth, A.E., 2009. Strategy-proofness versus efficiency in matching with indifferences: Redesigning the NYC high school match. Am. Econ. Rev. 5 (99), 1954–1978.

Akbarpour, M., Li, S., 2017. Credible mechanisms. Mimeo.

Ashlagi, I., Kanoria, Y., Leshno, J.D., 2017. Unbalanced random matching markets: the stark effect of competition. J. Polit. Econ. 125 (1), 69–98.

Ashlagi, I., Braverman, M., Kanoria, Y., Shi, P., 2017. Communication requirements and informative signaling in matching markets. In: Proceedings of the 18th ACM Conference on Economics and Computation (EC 2017), p. 263.

Bade, S., Gonczarowski, Y.A., 2017. Gibbard-Satterthwaite success stories and obvious strategyproofness. In: Proceedings of the 18th ACM Conference on Economics and Computation (EC 2017), p. 565.

Bó, I., Hakimov, R., 2016a. The iterative deferred acceptance mechanism. Mimeo.

Bó, I., Hakimov, R., 2016b. Iterative versus standard deferred acceptance: Experimental evidence. Mimeo.

Bogomolnaia, A., Deb, R., Ehlers, L., 2005. Strategy-proof assignment on the full preference domain. J. Econ. Theory 123 (2), 161–186.

Chen, L., Pereyra, J., 2015. Time-constrained school choice. Mimeo.

Chen, P., Egesdal, M., Pycia, M., Yenmez, M.B., 2016. Manipulability of stable mechanisms. Am. Econ. J. Microecon. 8 (2), 202–214.

Chen, Y., Sönmez, T., 2006. School choice: an experimental study. J. Econ. Theory 127 (1), 202–231.

Demange, G., Gale, D., Sotomayor, M., 1986. Multi-item auctions. J. Polit. Econ. 94 (4), 863–872.

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

Dubins, L.E., Freedman, D.A., 1981. Machiavelli and the Gale–Shapley algorithm. Am. Math. Mon. 88 (7), 485–494.

Dur, U., Hammond, R.G., Morrill, T., 2018. Identifying the harm of manipulable school-choice mechanisms. Am. Econ. J.: Econ. Pol. 10 (1), 187–213.

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

Gale, D., Shapley, L.S., 1962. College admissions and the stability of marriage. Am. Math. Mon. 69 (1), 9–15.

Gale, D., Sotomayor, M., 1985. Ms. Machiavelli and the stable matching problem. Am. Math. Mon. 92 (4), 261–268.

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

Gong, B., Liang, Y., 2016. A dynamic college admission mechanism in Inner Mongolia: Theory and experiment. Mimeo.

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

Hassidim, A., Marciano, D., Romm, A., Shorrer, R.I., 2017. The mechanism is truthful, why aren't you? Am. Econ. Rev. 107 (5), 220–224.

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

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

Kojima, F., Pathak, P.A., 2009. Incentives and stability in large two-sided matching markets. Am. Econ. Rev. 99 (3), 608–627.

Li, S., 2017. Obviously strategy-proof mechanisms. Am. Econ. Rev. 107 (11), 3257–3287.

Luflade, M., 2017. The value of information in centralized school choice systems. Mimeo (job market paper).

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

Pais, J., Klijn, F., Vorsatz, M., 2016. Static Versus Dynamic Deferred Acceptance in School Choice: Theory and Experiment. Working Paper 926. Barcelona GSE.

Pathak, P.A., Sönmez, T., 2008. Leveling the playing field: sincere and sophisticated players in the Boston mechanism. Am. Econ. Rev. 98 (4), 1636–1652.

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

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

Roth, A.E., 1984. The evolution of the labor market for medical interns and residents: a case study in game theory. J. Polit. Econ. 92 (6), 991–1016.

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

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

Shorrer, R.I., Sóvágó, S., 2017. Obvious mistakes in a strategically simple college-admissions environment. Mimeo.

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