Strategic Simplicity in Jury Selection, Committee Selection, and Matching

By

Martin Van der Linden

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Economics

June 30, 2017

Nashville, Tennessee

Approved:

John A. Weymark, Ph.D.

Paul H. Edelman, Ph.D.

Eun Jeong Heo, Ph.D.

Myrna Wooders, Ph.D.

To Emily and Jack.

ACKNOWLEDGMENTS

My gratitude goes first and foremost to my advisor, John Weymark. There is a rumor in the profession that John is the most meticulous reviewer there is, and that no review is more valuable than when it comes from him. The rumor is true. John's patient and tireless reviews taught me how to think and write better, and many results in this thesis owe a lot to his excellent teaching.

John also excels at selecting dedicated committee members. Throughout the writing of this thesis, Paul Edelman, Eun Jeong Heo, and Myrna Wooders have gone above and beyond what can be expected from committee members and have provided invaluable comments, feedbacks and discussions. I am greatly indebted for their help and for accepting to be part of my committee.

Numerous other people generously offered their time to discuss and comment on the work in thesis, as well as other ideas and projects that did not make it into this thesis. With my sincere apologies to those who have been forgotten, I am thankful to Tommy Andersson, Geir Asheim, Andrew Daughety, Benoit Decerf, Andrew Dustan, Hanna Frank, Greg Leo, Clayton Masterman, Jordi Massó, John Nay, Claudia Rei, Roberto Serrano, Yves Sprumont and Rodrigo Velez. Some of the work in this thesis was presented at Université Aix-Marseille, Université catholique de Louvain, Katholiek Universiteit Leuven, Vanderbilt University, the 13th meeting of the Society for Social Choice and Welfare, the Fifth World Congress of the Game Theory Society and the Vanderbilt Mechanism Design Conference, 2016. I am thankful the participants to these presentations for their questions and suggestions. Needless to say, any remaining errors are mine.

Special thanks go to Patrick Prosser and his java code for finding stable matchings (http://www.dcs.gla.ac.uk/~pat/roommates/distribution/), which has been helpful in constructing some of the examples in the last chapter of this thesis, and to CEREC and CORE for their hospitality during visits.

I also want to express my gratitude to François Maniquet, whose passion for economics in general, and social choice theory in particular, made me to drift away from my previous interests and decide to study economics. It is François's work and teaching that first convinced me of the beauty, value, and usefulness of economics, and for that, I am indebted to him.

I could not have completed this thesis without the help of the administrative staff of the Department of Economics at Vanderbilt University. I am especially thankful to our Assistant to the DGS Kathleen Finn for her flexibility, and to our outstanding DGS Jennifer Reinganum.

Financial support from the Kirk Dornbush summer research grant and the National Science Foundation grant IIS-1526860 is gratefully acknowledged. I am deeply thankful to Yevgeniy Vorobeychik and Myrna Wooders, the PI and co-PI on the latter grant, for giving me the opportunity to work as an RA for the grant.

One of the privileges of academia is to be able to work with people that you admire and appreciate, and who eventually become friends. While I was working on this thesis, I have been fortunate to work on other projects with people like Benoit Decerf, Greg Leo, and John Nay. Not only did I learn (and keep learning) a lot from them, they made work feel like play. I can only wish for many more years of fruitful collaborations and friendship, and for more lucky encounters of this kind.

Of course, friendship goes beyond direct collaborations. While at Vanderbilt University and Université Catholique de Louvain, I have had the privilege to enjoy amazingly collegial graduate and PhD programs. From the days of my Masters in Louvain, to Graduate classes at Vanderbilt, all the way to the job market, my fellow graduate students have provided invaluable support and comradeship, and I could not thank them enough for that.

Finally, I want to thank my family, in the US and in Belgium, as well as my friends in Belgium. To my family in the US: you have been so welcoming from day one, and having you around has made being away from Belgium a lot sweeter. If it was not for you, the prospect of a life in the US would make much less sense. To my family and friends in Belgium: your visits and your patience awaiting our summer returns has made being away from Belgium a lot less bitter. If it was not for the continued feeling of never having left every time we return, the prospect of a life in the US would make much less sense too.

Last but not least, I want to thank Emily for her unconditional love and support. And Jack, for having decided that 3 months old is old enough to sleep through the night, and much more.

TABLE OF CONTENTS

Page

DEDICATION . . . ii

ACKNOWLEDGMENTS . . . iii

LIST OF TABLES . . . viii

LIST OF FIGURES . . . ix

1 Introduction . . . 1

2 Impossibilities for strategy-proof committee selection mechanisms with vetoes . . . 4

2.1 Introduction . . . 4

2.2 The model . . . 7

2.3 Preliminary results . . . 9

2.4 No sure committee in the range of strategy-proof 2-vetoers mechanisms . . . 10

2.4.1 Main result . . . 10

2.4.2 Applications . . . 14

2.5 No probability thresholds in strategy-proof 2-vetoers mechanisms . . . 15

2.5.1 Main result . . . 16

2.5.2 Applications . . . 20

2.6 Concluding remarks . . . 22

3 Bounded rationality and the choice of jury selection procedures . . . 33

3.1 Introduction . . . 33

3.2 Model and procedures . . . 37

3.2.1 The model . . . 38

3.2.2 Procedures . . . 39

3.3 Impossibility results . . . 41

3.4 A measure of strategic complexity 44

3.4.1 Motivating example 45

3.4.2 The dominance threshold 46

3.5 One-shot procedures 49

3.5.1 One-shot$_Q$ is always maximally simple 49

3.5.2 One-shot$_M$ is often complex: one-common profiles 51

3.5.2.1 Motivating example 51

3.5.2.2 One-common profiles 52

3.6 Alternating and other N-struck procedures 55

3.7 Conclusion 57

3.8 Prevalence of alternating$_M$-one-common profiles 71

4 Deferred acceptance is minimally manipulable 73

4.1 Introduction 73

4.2 Two partial orders for manipulability comparisons 77

4.3 The one-to-one environment 81

4.4 Preliminary results 82

4.5 Maximal and minimal manipulability among stable mechanisms 83

4.6 A conflict between fairness and manipulability 88

4.6.1 A conflict between miniworst and manipulability 89

4.6.2 A conflict between median stable mechanisms and manipulability . . . 92

4.7 Concluding remarks 95

BIBLIOGRAPHY 115

LIST OF TABLES

Table Page

3.1 Overlap in challenges in the 750 arbitration cases (de Clippel et al., 2014). . 69

3.2 Proportion of experimental plays of the veto-rank mechanism in de Clippel et al. (2014) in which *both* players challenged their two worst arbitrators. . . 70

4.1 Numerical values for the upper and lower-bounds of Propositions 4.4 and 4.5. 87

LIST OF FIGURES

Figure Page

3.1 Representation of first-, second-, and third-best procedures. 47

3.2 Computing the dominance threshold in Example 3.4. 50

3.3 Iterated best responses based on two of the L_p^0 strategies in the problem of Example 3.5. 53

3.4 Arbitrary hierarchical model for a given preference profile (R_d, R_p). 58

3.5 Iterated best responses based on two of the L_p^0 strategies. 62

3.6 Proportion of one-common profiles relative to the set of (a) all profiles and (b) almost juror inverse profiles. 66

3.7 Lower bounds on the percentage of one-common profiles in the 750 arbitration cases (blue line) as a function of the percentage of cases in which both parties were truthful. Estimates for the later percentage using experimental data from de Clippel et al. (2014) are shown by the dashed vertical lines (see Table 3.2). 71

4.1 Representation of the Pareto partial order. An arrow from mechanism *A* to mechanism *B* indicates that *A* Pareto dominates *B*. Because *DA* is Pareto efficient, the set of mechanisms that Pareto dominate *DA* is the empty set $\emptyset$. Mechanisms $A_1, \ldots, A_z$ are not Pareto efficient. 74

4.2 Representation of some of the results: An arrow from mechanism *A* to mechanism *B* indicates that *A* is less (AM- or PS-) manipulable than *B*. A box surrounding a list of mechanisms represent the relative size of that list. . . . 75

4.3 Representation of the results in Section 4.6.1. 92

4.4 Representation of the results in Section 4.6.2. The set *MS* is the set of *MS* mechanisms. 95

Chapter 1

Introduction

In many contexts, the choice of socially relevant outcomes is affected by the interaction of individuals in a way that matters to these individuals collectively. Examples include jury trials where each party can veto potential jurors that the other party likes, or auctions where one bidder's chance to win the auction depends on the bids of other bidders. The precise set of rules that relates individual interactions with the selection of a social outcome is called a *mechanism*.

When participating in a mechanism, individuals may behave strategically in order to influence the selected outcome in a way that best satisfies their preferences. For example, a party to a trial may withhold a veto on a potential juror if she thinks that the other party will veto that same juror in order to use that veto on another potential juror. In some auctions, bidders may also benefit from shading their bid if they believes that other bidders will follow suit.

If a mechanism is strategically complex and the stakes are high, considerable resources may be devoted to determining appropriate strategies. Resources spent on strategizing can be viewed as transaction costs inherent to the mechanism and should ideally be minimized. Strategic complexity also raises issues of fairness. In a number of cases, strategic complexity has been shown to favor strategically sophisticated individuals or individuals who can invest ample resources into strategic counseling, at the expense of more naive or less resourceful individuals (Basteck and Mantovani, 2016a; Pathak and Sönmez, 2008; Basteck and Mantovani, 2016b; Dur et al., 2017). For these reasons and others, strategic simplicity has long been viewed as a desirable property of mechanisms.

The best way to make a mechanism strategically simple is to eliminate strategic issues altogether. This can be done by ensuring that each individual has a strategy available that is a best response to the strategies of other individuals *regardless* of the strategies others choose. Such "universal best responses" are called *dominant strategies*. If the mechanism has individuals reporting preferences over the outcomes, dominant strategies are typically required to consist in reporting their true preferences. In this context, a mechanism that does not have truthful dominant strategies is called *manipulable*.

Since Gibbard (1973), it has been shown in a variety of problems that the goal of providing individuals with dominant strategies is irreconcilable with other objectives, such as efficiency. Negative results of this sort are called *impossibility* results. Impossibility results are essential in helping us identify the limits of what mechanisms can achieve.

In Chapter 2 of this dissertation, I provide new impossibility results for the problem of selecting a committee of a fixed number of members out of a set of candidates. Guaranteeing that no individual has veto power over potential outcomes has traditionally been viewed as a positive feature of a mechanism (Maskin, 1999). In practice, however, it is common and often desirable to endow individuals with veto power, especially in committee selection (see Chapter 2). I show that even limited veto power makes many committee selection mechanisms of interest manipulable. This applies in particular (i) to mechanisms the range of which contains a degenerate lottery in which a committee is chosen for sure and (ii) to mechanisms that are constructed from extensive game forms with a finite number of strategies. These impossibilities hold on a large set of domains including the domain of additive preferences and even when probabilistic mechanisms are allowed.

Dominant strategies are an important reference point in assessing the strategic simplicity of a mechanism. But the partition between dominant strategy mechanisms and mechanisms that do not have dominant strategies is coarse. Although dominant strategy mechanisms represent a first-best, one should not necessarily conclude that a mechanism that fails to have dominant strategies is strategically complex. Neither should one necessarily conclude that two mechanisms that both fail to have dominant strategies are equally complex.

In Chapters 3 and 4, I contribute to a recent literature that has sought to go "beyond" dominant strategies impossibilities by (i) identifying second-best strategic simplicity properties, and (ii) providing criteria to compare the strategic simplicity of mechanisms that fail to have dominant strategies.

In Chapter 3, I introduce the *dominance threshold*, a new measure of strategic complexity based on "level-k" thinking. I use this measure to compare mechanisms used in practice to select juries in jury trials. In applying this measure, I overturn some commonly held beliefs about which jury selection mechanisms are strategically simple. In particular, I show that sequential mechanisms tend to be strategically simpler than mechanisms that involve simultaneous moves: By generating imperfect information games, simultaneous mechanisms increase the amount of guesswork needed to determine optimal strategies.

In Chapter 4, I show that, in the context of one-to-one two-sided matching, the deferred acceptance mechanism cannot be improved upon in terms of manipulability in the sense of Pathak and Sönmez (2013) or Arribillaga and Massó (2015) without compromising stability. I also identify conflicts between manipulability and fairness. Stable mechanisms that minimize the set of individuals who match with their least preferred achievable mate are shown to be maximally manipulable among the stable mechanisms. These mechanisms are also more manipulable than the deferred acceptance mechanism. I identify a similar conflict between fairness and manipulability in the case of the median stable mechanisms.

Chapter 2

Impossibilities for strategy-proof committee selection mechanisms with vetoes

2.1 Introduction

Guaranteeing that no individual has veto power over potential outcomes has traditionally been viewed as a positive feature of a mechanism (Maskin, 1999). In practice, however, it is common and often desirable to endow individuals with veto power. Vetoes can be required to protect individual rights, as forcefully argued by Sen (1970). Vetoes can also be used to avoid selecting outcomes that would be overly detrimental to some individuals. This latter use of vetoes is often found in procedures that select committees. Examples include jury selection (see Flanagan (2015) for a recent review) and other judicial procedures such as the selection of arbitrators (de Clippel et al., 2014) and Special Masters (see, e.g., Valdivia v. Schwarzenegger[1]). The right to veto *one* candidate to the papal throne was also exercised by France, Austria and Spain in various shapes and forms from the late 16th until the beginning of the 20th Century (O'Malley, 2015, p.41).

In many procedures, the individuals' actions are, in fact, limited to vetoes. For example, in sequential jury selection procedures, the defendant and the plaintiff take turns vetoing potential jurors until they have exhausted all of their vetoes.[2] Mueller (1978) and Moulin (1981) have shown that similar procedures in which individuals take turns vetoing outcomes can be used to implement desirable social choice functions using backward induction.[3]

One issue with veto procedures is that, despite having interesting equilibria, they are often manipulable. Examples of manipulable veto procedures include the procedures stud-

[1] Stipulation and Amended Order Re Special Master Order of Reference, Valdivia v. Schwarzenegger, No. CIV-S-94-0671 (E.D. Cal. filed Aug. 19, 2005)

[2] For the difference between sequential and simultaneous jury selection procedures, see Van der Linden (2017).

[3] Among other properties, the social choice functions in Mueller (1978) and Moulin (1981) are Pareto efficient and never select an individual's worst alternative.

ied in Mueller (1978), Moulin (1981), de Clippel et al. (2014), and Van der Linden (2017). Moreover, evidence from experimental and field data shows that individuals do attempt to manipulate procedures involving vetoes and sometimes fail to reach an equilibrium (Yuval, 2002; de Clippel et al., 2014). Because vetoes can be of normative importance, and because they are so pervasive in practice, it is of interest to ask whether reasonable veto procedures exist that leave no room for manipulation.

In this paper, I answer this question negatively for the standard problem in which a group of voters select a committee of k members out of a candidates. I show that endowing as few as two voters with the power to veto a *single* candidate makes many mechanisms of interest manipulable. This is true for a variety of domains, including the domain of additive preferences and any of its supersets (e.g., the domain of separable preferences), which are the domains of preferences most commonly studied in selection problems (Barberà et al., 1991, 2005). This is also the case when probabilistic mechanisms are allowed.

On these domains, I show that strategy-proof mechanisms with vetoers must have ranges that (i) do not contain degenerate lotteries in which a committee is chosen for sure and (ii) have as a limit point a lottery in which the probability of selecting a particular candidate is zero. Condition (i) implies that every *deterministic* mechanism with vetoers violates strategy-proofness. Condition (i) also restricts the efficiency of strategy-proof mechanisms with vetoers. For example, a strategy-proof mechanism with vetoers cannot always select committees that voters unanimously prefer. Condition (ii) implies that the range of strategy-proof mechanisms with vetoers must be in some sense "dense" around some lotteries. In particular, it implies that a wide class of selection mechanisms constructed from extensive game forms with a *finite* number of strategies violate strategy-proofness.

Related Literature

For deterministic mechanisms with an unrestricted domain, the Gibbard-Satterthwaite Theorem implies that every strategy-proof mechanism with more than three committees in its range is dictatorial. At least two approaches have been used to overcome this negative result.

The first approach weakens the unrestricted domain assumption. In the context of committee selection, this typically involves assuming that voters' preferences satisfy some separability condition. Unfortunately, Barberà et al. (2005) show that, even when preferences are separable, only a restricted set of non-dictatorial selection mechanisms are strategy-proof.[4]

The second approach allows for probabilistic selection mechanisms that select *lotteries* over committees rather than sure committees. In many problems, however, strategy-proofness cannot be combined with other desirable properties even using a probabilistic mechanism. In voting models with a finite set of outcomes, strategy-proofness is, for example, incompatible with unanimity except for (possibly random) dictatorial mechanisms (Hylland, 1980; Schummer, 1999; Benoît, 2002; Dutta et al., 2006; Nandeibam, 2012; Chatterji et al., 2012).

In this paper, I combine both approaches by considering probabilistic mechanisms on a variety of domains that include the domain of separable preferences. I show that strategy-proofness is, in general, incompatible with giving voters a minimal veto power over candidates. This finding contrasts with Ju (2003), who studies domain restrictions for which strategy-proof mechanisms exist that do *not* give voters veto power over candidates. My results provide new evidence of the difficulty of combining strategy-proofness with other desirable requirements, and of the limited freedom one gains by imposing domain restrictions and allowing for probabilistic mechanisms.

The paper is organized as follows. Section 2.2 presents the model and introduces preliminary results that are later used in the proofs. Section 2.3 introduces preliminary results that are later used in the proofs. Section 2.4 shows that strategy-proof mechanisms with

[4] The characterization in Barberà et al. (2005) is more permissive for *additive* preferences. But even on this smaller domain, the class of strategy-proof selection mechanisms remains a small subclass of the mechanisms known as *voting by committees* (the "committees" in *voting by committees* are committees of *voters* and should not be confused with the selected committee of k members). See also Barberà et al. (1991) for the problem of selecting a committee without size constraints.

vetoers have ranges that do not contain degenerate lotteries. Section 2.5 shows that the ranges of strategy-proof mechanisms with vetoers have as a limit point a lottery in which the probability of selecting a particular candidate is zero. Sections 2.4 and 2.5 also provide illustrative applications of these results. I conclude with some open questions. Omitted proofs may be found in the Appendix.

2.2 The model

The set of **voters** is $N := \{1,\dots,n\}$ with $n \geq 2$. The set of **candidates** is $A := \{1,\dots,a\}$ with $a \geq 2$. For any integer $k \in \{1,\dots,a\}$, the set of possible **committees** $\mathscr{A}_k$ is the set of subsets of A with k elements. Let $\Delta\mathscr{A}_k$ be the set of lotteries on $\mathscr{A}_k$. Slightly abusing the notation, let $C \in \mathscr{A}_k$ denote any degenerate *lottery* which yields committee C for sure.

For any lottery $L \in \Delta\mathscr{A}_k$ and any committee $C \in \mathscr{A}_k$, C's **selection probability** $L(C)$ is the probability that C is the chosen committee given L. A lottery L is **degenerate** if $L(C) = 1$ for some $C \in \mathscr{A}_k$. Similarly, for any lottery $L \in \Delta\mathscr{A}_k$ and any candidate $t \in A$, t's **selection probability** $L(t)$ is the probability that t is part of the chosen committee given L. Formally, $L(t) := \sum_{\{S \in \mathscr{A}_k | t \in S\}} L(S)$.

A preference on $\Delta\mathscr{A}_k$ is denoted R, with asymmetric counterpart P. A typical domain of preferences on $\Delta\mathscr{A}_k$ is denoted by $\mathscr{D}$. For every domain $\mathscr{D}$ in this paper, preferences in $\mathscr{D}$ are orderings that satisfy the expected utility axioms. A (preference) **profile** is an n-tuple $R_N := (R_1,\dots,R_n) \in \mathscr{D}^n$. For any profile R_N and any $i \in N$, $R_{-i} := (R_1,\dots,R_{i-1},R_{i+1},\dots,R_n)$ is the $(n-1)$-tuple that lists the preferences of every player but i.

The domain of additive preferences $\mathscr{R}_{add}$ has received considerable attention in committee selection (Barberà et al., 1991, 2005). The domain $\mathscr{R}_{add}$ consists of all the preferences R on $\Delta\mathscr{A}_k$ that can be represented by a utility function $u\colon A \to \mathbb{R}$ on the set of *candidates* in the following additive way: for all $L, L' \in \Delta\mathscr{A}_k$,

$$L\,R\,L' \Leftrightarrow \sum_{C\in\mathscr{A}_k} L(C) \sum_{t\in C} u(t) \geq \sum_{C\in\mathscr{A}_k} L'(C) \sum_{t\in C} u(t). \tag{2.1}$$

The larger domain of separable preferences $\mathscr{R}_{sep}$ is also often considered (Barberà et al., 1991, 2005; Arribillaga and Massó, 2017). The domain $\mathscr{R}_{sep}$ contains all the preferences R on $\Delta\mathscr{A}_k$ for which, for any $C, C' \in \mathscr{A}_k$, any $a \in C \cap C'$ and any $b \in A \backslash (C \cup C')$, $U(C \cup \{b\} \backslash \{a\}) \geq U(C)$ if and only if $U(C' \cup \{b\} \backslash \{a\}) \geq U(C')$, where U is a von Neumann–Morgenstern utility function on $\mathscr{A}_k$ representing R.

A **(selection) mechanism** is a function $M\colon \mathscr{D}^n \to \Delta\mathscr{A}_k$ that associates a lottery in $\Delta\mathscr{A}_k$ with every profile in $\mathscr{D}^n$. For any $R_N \in \mathscr{D}^n$, $M(R_N)$ is the lottery selected by M when R_N is reported. The **range** of M is the set of $L \in \Delta\mathscr{A}_k$ which can be selected using M; that is,

$$range(M) := \{L \in \Delta\mathscr{A}_k \mid M(R_N) = L \text{ for some } R_N \in \mathscr{D}^n\}. \tag{2.2}$$

The next definition introduces a relatively weak concept of a vetoer. A voter $i \in N$ is a vetoer if for each $t \in A$, voter i can report a preference $R_i^t \in \mathscr{D}$ which guarantees that candidate t is not part of the chosen committee whatever R_{-i} the other voters report. Formally, given a mechanism M, any voter $i \in N$ is a **vetoer** if for each $t \in A$, there exists $R_i^t \in \mathscr{D}$ with

$$M(R_i^t, R_{-i})(t) = 0 \qquad \text{for all } R_{-i} \in \mathscr{D}^{n-1}. \tag{2.3}$$

Although a vetoer can veto *any* candidate, a vetoer is only guaranteed the ability to veto *one* candidate at a time. For example, for i to be a vetoer, there does *not* need to be any pair of candidates (t, t') with $t \neq t'$ such that for some $\bar{R}_i \in \mathscr{D}$, $M(\bar{R}_i, R_{-i})(t) = M(\bar{R}_i, R_{-i})(t') = 0$ for all $R_{-i} \in \mathscr{D}^{n-1}$.

A selection mechanism M is an ***r*-vetoers mechanism** if there are *at least r distinct* vetoers in M.

A selection mechanism M is **strategy-proof** if for all $i \in N$, reporting i's true preference is a dominant strategy; that is, for all $R_i \in \mathscr{D}$

$$M(R_i, R_{-i}) \; R_i \; M(R_i', R_{-i}) \qquad \text{for all } R_i' \in \mathscr{D} \text{ and all } R_{-i} \in \mathscr{D}^{n-1}.$$

For any subset $B \subseteq \Delta\mathscr{A}^k$ and any $R \in \mathscr{D}$, the **top set** $top(R, B)$ is the set of best lotteries in B according to R. Formally, for all $B \subseteq \Delta\mathscr{A}_k$ and all $R \in \mathscr{D}$,

$$top(R,B) \coloneqq \{L \in B \mid L \, R \, L' \text{ for all } L' \in B\}.$$

A voter $j \in N$ is a **dictator** for mechanism M if the lottery that M chooses is always in j's top set; that is,

$$M(R_N) \in top\big(R_j, \Delta\mathscr{A}_k\big) \qquad \text{for all } R_N \in \mathscr{D}^n.$$

Finally, for any $i \in N$ and any preference $R_i \in \mathscr{D}$, the **option set** $O_{-i}(R_i)$ is the set of lotteries that M chooses for some report of the preferences of voters in $N \backslash \{i\}$ given that voters i report R_i (Barberà and Peleg, 1990). Formally, for all $R_i \in \mathscr{D}$,

$$O_{-i}(R_i) \coloneqq \left\{L \in \Delta\mathscr{A}_k \mid M(R_i, R_{-i}) = L \text{ for some } R_{-i} \in \mathscr{D}^{n-1}\right\}.$$

Note that i is a dictator if and only if $O_{-i}(R_i) \subseteq top\big(R_i, \Delta\mathscr{A}_k\big)$ for all $R_i \in \mathscr{D}$.

2.3 Preliminary results

This section introduces two propositions that I use repeatedly in the proofs. These propositions follow from results in Le Breton and Weymark (1999).

The first says that given a profile R_N, if M is strategy-proof and if some $i \in N$ has a unique top lottery L in the range of M, then L must be contained in the option set $O_{-i}(R_i)$.

Proposition 2.1. *Suppose that $M\colon \mathscr{D}^n \to \Delta\mathscr{A}_k$ is a strategy-proof mechanism. For all $i \in N$ and all $R_i \in \mathscr{D}$, if $top\big(R_i, range(M)\big) = \{L\}$, then $L \in O_{-i}(R_i)$.*

Proof. This proposition is a direct corollary of Le Breton and Weymark (1999, Proposition 3). □

The second proposition says that if M is strategy-proof and if all voters in $N\backslash\{i\}$ agree on the set of top lotteries B in the option set $O_{-i}(R_i)$, then the chosen lottery must be included in B.

Proposition 2.2. *Suppose that $M\colon \mathscr{D}^n \to \Delta\mathscr{A}_k$ is a strategy-proof mechanism. For all $R_N \in \mathscr{D}^n$ and all $i \in N$, if there exists a nonempty set $B \subseteq O_{-i}(R_i)$ such that $top(R_i, O_{-i}(R_i)) = B$ for all $i \in N\backslash\{i\}$, then $M(R_N) \in B$.*

Proof. This proposition is a direct corollary of Le Breton and Weymark (1999, Proposition 4). □

2.4 No sure committee in the range of strategy-proof 2-vetoers mechanisms

In this section, I show that no strategy-proof 2-vetoers mechanism can have in its range a degenerate lottery. This impossibility precludes the existence of deterministic strategy-proof 2-vetoers mechanisms and severely limits the efficiency of strategy-proof 2-vetoers mechanisms.

2.4.1 Main result

The main result in this section is the following.

Theorem 2.1. *If $M\colon \mathscr{D}^n \to \Delta\mathscr{A}_k$ is a 2-vetoers mechanism with a sure committee $C \in \mathscr{A}_k$ in its range and $\mathscr{D} \supseteq \mathscr{R}_{add}$, then M is not strategy-proof.*

The theorem applies, for instance, when $\mathscr{D} = \mathscr{R}_{sep}$. In order to identify the features of $\mathscr{R}_{add}$ that are responsible for the impossibility in Theorem 2.1, I establish a more general theorem. Theorem 2.2 is based on technical domain conditions that are satisfied by $\mathscr{R}_{add}$, and that are sufficient for the impossibility to hold. These technical conditions can also be used to identify domains that are *not* supersets of $\mathscr{R}_{add}$ on which the impossibility holds.

A minimin domain contains sequences of preferences for which the impact of the "worst" candidate on the value of a committee is more and more negative. In such a sequence, voters become increasingly concerned with *minimizing* the selection probability of their "worst" candidate.[5] In addition, for technical reasons that are made clear in the proof of Lemma 2.3 (see the Appendix), preferences in the sequences must have the same most preferred committee, with the "worst" candidate not a member of this committee.

Domain Property 2.1 (Minimin domain)**.** A domain of preferences $\mathscr{R}$ on the lotteries in $\Delta\mathscr{A}_k$ is **minimin** if for any candidate $t \in A$, for any committee $C \in \mathscr{A}_k$ with $t \notin C$, and for any $\varepsilon > 0$, there exists $R^\varepsilon \in \mathscr{R}$ such that

(i) $L \, P^\varepsilon \, L'$ for all $L, L' \in \Delta\mathscr{A}_k$ for which $L(t) < L'(t) - \varepsilon$, and

(ii) $C \, P^\varepsilon \, C'$ for all $C' \in \mathscr{A}_k \backslash \{C\}$.

Maximax is in a sense the inverse of minimin. A domain is maximax if it contains sequences of preferences for which the impact of the "best" candidate on the value of a committee is more and more positive. In such a sequence, voters become increasingly concerned with *maximizing* the selection probability of their "best" candidate.[6] In addition, for technical reasons that are made clear in the proof of Lemma 2.3 (see the Appendix), preferences in the sequences must have the same most preferred committee among the committees that do *not* contain the "best" candidate.

[5] Hence, the name "minimin", for "*mini*mizing the selection probability of the candidate whose contribution to the committee is *min*imal".

[6] Hence, the name "maximax", for "*maxi*mizing the selection probability of the candidate whose contribution to the committee is *max*imal".

Domain Property 2.2 (Maximax domain)**.** A domain of preferences $\mathscr{R}$ on the lotteries in $\Delta\mathscr{A}_k$ is **maximax** if for any candidate $t \in A$, for any committee $C \in \mathscr{A}_k$ with $t \notin C$, and for any $\varepsilon > 0$, there exists $R^{\varepsilon} \in \mathscr{R}$ such that

(i) $L\,P^{\varepsilon}\,L'$ for all $L, L' \in \Delta\mathscr{A}_k$ such that $L(t) > L'(t) + \varepsilon$, and

(ii) $C\,P^{\varepsilon}\,C'$ for all $C' \in \mathscr{A}_k \backslash \{C\}$ with $t \notin C'$.

Remark 2.1. By definition, a domain is minimin or maximax if it *contains* certain preferences. Hence, if a domain $\mathscr{R}^*$ is minimin or maximax, so is any superset of $\mathscr{R}^*$.

Lemma 2.1. *Any domain $\mathscr{D} \supseteq \mathscr{R}_{add}$ is both minimin and maximax.*

By Lemma 2.1, the following theorem generalizes Theorem 2.1.

Theorem 2.2. *If $M\colon \mathscr{D}^n \to \Delta\mathscr{A}_k$ is a 2-vetoers mechanism with a sure committee $C \in \mathscr{A}_k$ in its range and $\mathscr{D} \supseteq \mathscr{R}$ for some minimin and maximax domain $\mathscr{R}$, then M is not strategy-proof.*

The proof of Theorem 2.2 is established using Lemmas 2.2 and 2.3. Lemma 2.2 shows that, on a minimin domain, if a vetoer $j \in N$ has a sufficiently strong concern for minimizing the selection probability of some $t \in A$, then any lottery L in the option set generated by j must have $L(t)$ arbitrarily small. Otherwise, j would have an incentive to report preferences that veto t, contradicting strategy-proofness.

Lemma 2.2. *Suppose that $M : \mathscr{R}^n \to \Delta\mathscr{A}_k$ is a strategy-proof mechanism and $\mathscr{D} \subseteq \mathscr{R}$ for some minimin domain $\mathscr{R}$. If $j \in N$ is a vetoer for M, then for any $t \in A$, any $\varepsilon > 0$, and any preference $R^{\varepsilon}_j \in \mathscr{R}$ satisfying (i) in the definition of a minimin domain,*

$$L(t) \leq \varepsilon \qquad \textit{for all } L \in O_{-j}(R^{\varepsilon}_j).$$

Lemma 2.3 shows that on a domain that is both minimin and maximax, if a strategy-proof mechanism ever selects a sure committee $C \in \mathscr{A}_k$, then C must be chosen whenever any vetoer j likes C best. Informally, suppose that j likes C best and there is a lottery L with $L(t) > 0$ for some $t \notin C$ in the option set generated by j. Because the domain is maximax, there exist preferences for which the inclusion of t in a committee is essential. For any such preference R^*, some lottery L with $L(t) > 0$ will be chosen when everyone but j reports R^* (by Proposition 2.2). It is possible to choose such a preference, say R^{**}, so that C is the best committee among the committees that do not contain t. But then, when everybody but j reports R^{**}, j can report minimin preferences that force t to be chosen with arbitrarily small probability while keeping C as j's best committee. If j does so, C remains in the option set (by Proposition 2.1) and whenever everyone but j reports R^{**}, a lottery that selects C with arbitrarily large probability is chosen instead of L (by Proposition 2.2), contradicting strategy-proofness.

Lemma 2.3. *Suppose that $M\colon \mathscr{D}^n \to \Delta\mathscr{A}_k$ is a strategy-proof mechanism with a sure committee $C \in \mathscr{A}_k$ in its range and $\mathscr{D} \supseteq \mathscr{R}$ for some minimin and maximax domain $\mathscr{R}$. If $j \in N$ is a vetoer for M, then for all $R_j \in \mathscr{D}$,*

$$C \; P_j \; C' \text{ for all } C' \in \mathscr{A}_k \backslash \{C\} \tag{2.4}$$

implies

$$L(C) = 1 \qquad \text{for all } L \in O_{-j}(R_j). \tag{2.5}$$

It is now easy to prove Theorem 2.2.

Proof of Theorem 2.2. Let M be a strategy-proof 2-vetoers mechanism with $C \in \mathscr{A}_k$ in its range. Let $j \in N$ be any vetoer and $R_j \in \mathscr{D}$ be any preference with $top(R_j, \Delta\mathscr{A}_k) = C$. By Lemma 2.3, we have $O_{-j}(R_j) = \{C\}$. Consider any other vetoer $h \in N$. Clearly, h cannot veto any candidate in C whenever j reports R_j, contradicting the assumption that M is a 2-vetoers mechanism. □

The proof of Theorem 2.2 does not use the full strength of the 2-vetoers condition. The proof only requires the existence of *one* vetoer j and of some $h \neq j$ with the ability to veto *one* of the candidates in C (where C can be any sure committee in the range of M).

2.4.2 Applications

The next result is a direct corollary of Theorem 2.2 for deterministic mechanisms. A mechanism M has a **sure range** if for all $R_N \in \mathscr{D}^n$, $M(R_N) = C$ for some $C \in \mathscr{A}_k$. A mechanism M **only considers the ranking of sure committees** if for all $R_N, R'_N \in \mathscr{D}^n$ that induce the same rankings over sure committees, $M(R_N) = M(R'_N)$. A mechanism M is **deterministic** if it satisfies the two last properties.

Corollary 2.1. *If $M\colon \mathscr{D}^n \to \Delta\mathscr{A}_k$ is a deterministic 2-vetoers mechanism and $\mathscr{D} \supseteq \mathscr{R}$ for some minimin and maximax domain $\mathscr{R}$, then M is not strategy-proof.*

Proof. A deterministic 2-vetoers mechanism is a 2-vetoers mechanism with a sure committee in its range. Thus, Theorem 2.1 applies. □

Note that the sure range condition alone is sufficient to obtain the above impossibility. That is, Corollary 2.1 also applies to sure range mechanisms that take cardinal information about preferences on committees into account.

By Lemma 2.1, Corollary 2.1 applies, for example, when $\mathscr{D} \supseteq \mathscr{R}_{add}$. In the case of $\mathscr{D} = \mathscr{R}_{add}$, Corollary 2.1 might not come as a surprise given the characterization of deterministic strategy-proof mechanisms on $\mathscr{R}_{add}$ in Barberà et al. (2005, Proposition 2).[7] However, Corollary 2.1 holds on the larger class of minimin and maximax domains, which includes $\mathscr{R}_{add}$ and its supersets (by Lemma 2.1), but also includes domains that are *not* supersets

[7] Barberà et al. (2005, Proposition 2) show that the class of deterministic strategy-proof mechanisms on $\mathscr{R}_{add}$ is a subset of the mechanisms known as *voting by committees*. If a voting by committees mechanism is a 2-vetoers mechanism, then the two vetoers i and j are in all *winning coalitions* (see Barberà et al., 2005, for a definition). But then when $C^i := top(R_i, range(M)) \neq top(R_j, range(M)) =: C^j$, the chosen committee is $C^* \subseteq C^i \cap C^j$, which implies $\#C^* < k$, a contradiction.

of $\mathscr{R}_{add}$. The impossibility in Corollary 2.1 is therefore stronger than the one that can be derived from Barberà et al. (2005, Proposition 2).

Theorem 2.1 also has negative implications for the efficiency of a strategy-proof 2-vetoers mechanism. Consider the following weakening of Pareto efficiency. A mechanism M satisfies **minimal sure unanimity** if there exists *at least one* committee $C \in \mathscr{A}_k$ for which $top(R_i, \Delta\mathscr{A}_k) = C$ for all $i \in N$ implies $M(R_N) = C$. We then have the following impossibility.

Corollary 2.2. *If $M\colon \mathscr{D}^n \to \Delta\mathscr{A}_k$ is a 2-vetoers mechanism that satisfies minimal sure unanimity and $\mathscr{D} \supseteq \mathscr{R}$ for some minimin and maximax domain $\mathscr{R}$, then M is not strategy-proof.*

Proof. For any $C \in \mathscr{A}_k$, there exist many $R_N^C \in \mathscr{D}$ such that C is the unique best committee for all $i \in N$. Thus, minimal sure unanimity implies $M(R_N^C) = C$, and C is in the range of M. But then by Theorem 2.1, M cannot be a strategy-proof 2-vetoers mechanism. □

As mentioned in the Introduction, results showing that in probabilistic mechanisms, strategy-proofness is incompatible with unanimity requirements except for dictatorial mechanisms date back to Hylland (1980). Corollary 2.2 is independent from those results. First, unlike Corollary 2.2, those results do not rule out dictatorial mechanisms. Second they either (i) hold on domains that are larger than or independent from the smallest minimin and maximax domain or (ii) rely on unanimity conditions that are stronger than or independent from minimal sure unanimity. Corollary 2.1 is not a generalization of any of those results either because it relies on the 2-vetoers condition.

2.5 No probability thresholds in strategy-proof 2-vetoers mechanisms

In this section, I show that if M is a strategy-proof 2-vetoers mechanism, then one limit point of the range of M must be a lottery that selects some candidate t with probability zero. This implies that for some t, there exists a lottery L in the range of M with $0 < L(t) \leq \varepsilon$ for every $\varepsilon > 0$. Equivalently, if for all $t \in A$ there exists a threshold $\varepsilon_t > 0$ such that t is never chosen with a positive probability smaller than ε_t, then a 2-vetoers mechanism cannot be strategy-proof.

This result may seem innocuous as there is *a priori* no reason to impose such thresholds. It however implies that a large class of mechanisms constructed from extensive game forms with a finite number of strategies violate strategy-proofness (Corollary 2.3). These extensive game forms include many that are used in practice, notably in jury selection procedures.

2.5.1 Main result

The main result in this section is the following.

Theorem 2.3. *If $M\colon \mathscr{D}^n \to \Delta\mathscr{A}_k$ is a strategy-proof 2-vetoers mechanism and $\mathscr{D} \supseteq \mathscr{R}_{add}$, then there exists $t \in A$ such that, for all $\varepsilon > 0$,*

$$0 < L^{\varepsilon}(t) \leq \varepsilon \qquad \textit{for some } L^{\varepsilon} \textit{ in the range of } M. \tag{2.6}$$

Again, in order to identify the features of $\mathscr{R}_{add}$ that are responsible for the impossibility in Theorem 2.3, I prove a more general theorem. Theorem 2.4 is based on a technical domain condition that is satisfied by $\mathscr{R}_{add}$ and that is sufficient for the impossibility to hold.

A domain is *negative leximin* if, as with minimin preferences, some voters are primarily concerned with minimizing the selection probability of a "worst" candidate. But if the selection probability of the "worst" candidate is fixed, then "negative leximin voters" become primarily concerned with minimizing the selection probability of the "*second* worst" candidate, and so on.

For Theorem 2.4 to hold, it is sufficient for the domain to include preferences that are close to a lexicographic assessment of any lottery for up to $(a-k)$ of the "worst" candidates (recall that a is the number of candidates and k the number of committee members that are to be selected). For technical reasons, it is also important that these preferences satisfy the defining properties of a maximax domain for some $t \in A$ whenever the selection probability of the candidates these preferences treat in a leximin fashion is unchanged. The importance of the last requirement is made clear in the proof of Lemma 2.5 (see the Appendix).

For any set S, let $\#S$ denote the cardinality of S. For any strict ordering $\succ$ of a finite set S with $s := \#S$, let $\succ_1, \succ_2, \ldots, \succ_s$ denote respectively the best element in S according to $\succ$, the second best element in S according to $\succ$, ..., the worst element in S according to $\succ$.

Domain Property 2.3 (Negative leximin)**.** A domain of preferences $\mathscr{R}$ on the lotteries in $\Delta\mathscr{A}_k$ is *negative leximin* if for any subset of candidates $X \subset A$ with $x := \#X \leq (a-k)$, any strict ordering $\succ$ of the candidates in X, any $t \in A \backslash X$, and any $\varepsilon > 0$, there exists $R^{\varepsilon} \in \mathscr{R}$ such that for all $L, L' \in \Delta\mathscr{A}_k$,

$$\begin{bmatrix} \left[L(\succ_x) < L'(\succ_x) - \varepsilon\right] \text{ or} \\ \left[L(\succ_x) = L'(\succ_x), L(\succ_{x-1}) < L'(\succ_{x-1}) - \varepsilon\right] \text{ or} \\ \vdots \\ \left[L(\succ_x) = L'(\succ_x), \ldots, L(\succ_1) < L'(\succ_1) - \varepsilon\right] \end{bmatrix} \Rightarrow \left[L \, P^{\varepsilon} \, L'\right] \tag{2.7}$$

and

$$\begin{bmatrix} \left[L(\succ_x) = L'(\succ_x), \ldots, L(\succ_1) = L'(\succ_1)\right] \text{ and} \\ \left[L(t) > L'(t) + \varepsilon\right] \end{bmatrix} \Rightarrow \left[L \, P^{\varepsilon} \, L'\right]. \tag{2.8}$$

Remark 2.2. As for minimin or maximax domains (Remark 2.1), a domain is negative leximin if, by definition, it contains certain preferences. Hence, if a domain $\mathscr{R}^*$ is negative leximin, so is any superset of $\mathscr{R}^*$.

Lemma 2.4. *Any domain $\mathscr{D} \supseteq \mathscr{R}_{add}$ is negative leximin.*

By Lemma 2.4, the following theorem generalizes Theorem 2.3.

Theorem 2.4. *If $M: \mathscr{D}^n \to \Delta\mathscr{A}_k$ is a strategy-proof 2-vetoers mechanism and $\mathscr{D} \supseteq \mathscr{R}$ for some negative leximin domain $\mathscr{R}$, then there exists $t \in A$ such that, for all $\varepsilon > 0$,*

$$0 < L^{\varepsilon}(t) \leq \varepsilon \qquad \textit{for some } L^{\varepsilon} \textit{ in the range of } M. \tag{2.9}$$

The proof of Theorem 2.4 is established using Lemmas 2.5 and 2.6. For any candidate $t \in A$, there is a **probability threshold $\varepsilon_t > 0$ for t** if $L(t) > 0$ implies $L(t) > \varepsilon_t$ for all $L \in range(M)$. Lemma 2.5 shows that in the presence of probability thresholds, vetoers can generate *singleton* option sets containing any sure committee by reporting appropriate preferences. The proof of Lemma 2.5 proceeds by induction on $A \backslash C$. Let $j, h \in N$ be two vetoers and C be j's best committee. The proof shows that if there are probability thresholds for all $t \in A$, then for larger and larger subsets of $A \backslash C$, j can reveal particular preferences which guarantee that no candidate in the subset is ever chosen with positive probability. The lemma then follows by strategy-proofness.

For a single $t \in A \backslash C$, this is true because j is a vetoer. Now consider any $t, t' \in A \backslash C$ with $t \neq t'$. Because the domain is negative leximin, there exists a preference $R_h^{t;t'}$ such that h cares primarily about *min*imizing the selection probability of t, and secondarily about *max*imizing the selection probability of t'. By an argument similar to the one used in Lemma 2.2, when such a preference is sufficiently extreme (i.e., for ε sufficiently small in the definition of a negative leximin domain), the selection probability of t must tend to zero (otherwise h would want to veto t). But because of the threshold assumption, this implies that t's selection probability *is* actually zero for some sufficiently extreme preference. That is, h effectively vetoes t when reporting this extreme preference.

When j has negative leximin preferences $R_j^{t,t'}$ that focus on *min*imizing the selection probability of both t and t', the option set generated by j cannot contain lotteries in which the selection probability of t' is positive. Otherwise, t' is selected with positive probability when everyone but j reports $R_h^{t;t'}$ (Proposition 2.2) and j would prefer to veto t' because h

already vetoes t by reporting $R_h^{t,t'}$.

Thus both vetoers can in fact veto *two* different candidates. Extending the argument by induction, if j has a preference that focuses on *min*imizing the selection probability of t, t' and t'', then the three candidates must be vetoed. Otherwise, whenever h vetoes t and t' while caring about *max*imizing the selection probability of t'' (which is possible by the previous step) and everyone but h and j reports the same preferences as h, j would be better off vetoing t'' than revealing her true preference.

Lemma 2.5. *Suppose that $M\colon \mathscr{D}^n \to \Delta\mathscr{A}_k$ is a strategy-proof 2-vetoers mechanism and $\mathscr{D} \supseteq \mathscr{R}$ for some negative leximin domain $\mathscr{R}$. Let $j \in N$ be a vetoer for M. If for all $t \in A$ there exists a probability threshold $\varepsilon_t > 0$, then for all $C \in \mathscr{A}_k$ there exists $R_j^* \in \mathscr{D}$ such that $O_{-j}(R_j^*) = \{C\}$.*

Using Lemma 2.5, we can prove the following result.

Lemma 2.6. *Suppose that $M\colon \mathscr{D}^n \to \Delta\mathscr{A}_k$ is a strategy-proof 2-vetoers mechanism and $\mathscr{D} \supseteq \mathscr{R}$ for some negative leximin domain $\mathscr{R}$. If for all $t \in A$ there exists a probability threshold $\varepsilon_t > 0$, then every vetoer is a dictator.*

Proof. Let $j \in N$ be any vetoer and $h \in N$ be any other vetoer. Consider any $R_N \in \mathscr{D}^n$. Because R_j satisfies the expected utility axioms, there exists a sure committee $C \in \mathscr{A}_k$ such that $C \in top(R_j, \Delta\mathscr{A}_k)$. By Lemma 2.5, there exists R_j^* such that $M(R_j^*, R_{-j}) = C$. If $M(R_N) \notin top(R_j, \Delta\mathscr{A}_k)$, we have $M(R_j^*, R_{-j})\ P_j\ M(R_N)$, contradicting strategy-proofness. Hence, we must have $M(R_N) \in top(R_j, \Delta\mathscr{A}_k)$ and thus j is a dictator. □

We can now prove Theorem 2.4.

Proof of Theorem 2.4. Let

$$l := \inf\left\{p \in (0,1] \;\middle|\; p = L^*(t) \text{ for some } L^* \in \mathit{range}(M) \text{ and some } t \in A\right\}.^{8}$$

If $l > 0$, then all $t \in A$ have a probability threshold, and by Lemma 2.6 every vetoer is a dictator for M. But then, because there are at least two vetoers, there are at least two dictators, which is impossible.[9] Thus, we must have $l = 0$ and (2.9) holds. □

The fact that there exists a lottery L^* with $L^*(t) = 0$ that is a limit point of the range follows from Theorem 2.4 by the Bolzano-Weierstrass Theorem, which shows that every bounded sequence has a converging subsequence (see, e.g., Rudin, 1976, Theorem 3.6(b)).[10]

2.5.2 Applications

The rest of this section illustrates the usefulness of Theorem 2.4 by showing how it rules out strategy-proofness for a wide class of mechanisms constructed from sequential procedures. Constructing direct mechanisms from sequential procedures is common in mechanism design. A simple example in the case of a selection mechanism on $\mathscr{R}_{add}$ is presented below.

Example 2.1 (Repeatedly veto the worst)**.** Choose two vetoers $j, h \in N$. For any profile of preferences $R_N \in \mathscr{R}^n_{add}$, select a committee by repeating the following two steps until there are only k candidates left. Let u_j and u_h be any utility functions on the set of candidates representing R_j and R_h as in (2.1).

(i) Remove the worst candidate according to u_j among the candidates in N that have not yet been removed (break ties randomly).

[8] Because M always selects a well-defined lottery over committees, there exists $t \in A$ and $L \in \Delta\mathscr{A}_k$ such that L is in the range of M and $L(t) > 0$, and this set is non-empty.

[9] Alternatively, if one of the vetoers is a dictator, then the other vetoers cannot always veto every alternative. For example, when a vetoer has a favorite sure committee, a second vetoer cannot veto any candidate in the dictator's favorite sure committee. Hence, the mechanism is not a 2-vetoers mechanism, which again yields a contradiction.

[10] By Theorem 2.4, there exists a sequence of lotteries $\{L^r\}_{r=1}^{\infty}$ in the range such that for some $t \in A$ we have $L^r(t) > 0$ for all $r > 0$ and $\lim_{r \to \infty} L^r(t) = 0$. By the Bolzano-Weierstrass Theorem, this sequence has a converging subsequence (the sequence is bounded because all lotteries belong to a 2^a-dimensional simplex). Clearly, the limit of that subsequence must be a lottery L^* with $L^*(t) = 0$. Also, L^* is, by definition, a limit point of the range.

(ii) Remove the worst candidate according to u_k among the candidates in N that have not yet been removed (break ties randomly).

To every $R_N \in \mathscr{R}^n_{add}$, the above algorithm associates a unique lottery in $\Delta\mathscr{A}_k$ and therefore defines a (direct) mechanism $M\colon \mathscr{R}^n_{add} \to \Delta\mathscr{A}_k$.

The algorithm in Example 2.1 can be viewed as an extensive game form in which the strategies of the players have been fixed as a function of their preferences.

In general, let a (selection) **procedure** be an extensive game form Γ in which

(a) the set of players is $I := N \cup \{Nature\}$ and

(b) every terminal node is a committee $C \in \mathscr{A}_k$.

For any domain of profiles $\mathscr{D}^n$ and any procedure Γ, a **generalized strategy profile** g associates every preference profile $R_N \in \mathscr{D}^n$ with a strategy profile $g(R_N)$ in the space of strategy profiles of Γ. A mechanism $M^{\Gamma}_{g,\rho}$ is **constructed from** procedure Γ if there exists a generalized strategy profile g and an assignment of probabilities ρ for Nature's moves such that

$$M^{\Gamma}_{g,\rho}(R_N) = \Gamma\big(g(R_N),\rho\big) \qquad \text{for all } R_N \in \mathscr{D}^n, \tag{2.10}$$

where $\Gamma\big(g(R_N),\rho\big)$ is the lottery resulting from Γ when strategy profile $g(R_N)$ is played and the probabilities associated with Nature's moves are ρ.

For example, the mechanism described in Example 2.1 is constructed from the extensive game form in which two vetoers take turns vetoing candidates, which is similar in spirit to procedures used in Mueller (1978), Moulin (1981), and in jury selection. In Example 2.1, the generalized strategy is what Moulin (1981) defines as the *prudent* strategy. At each decision node, a vetoer j chooses the action that maximizes his or her utility assuming that all further actions will be chosen in such a way as to minimize her utility.

As in Example 2.1, procedures used to construct mechanisms are often **finite**, in the sense that they have a finite number of *nodes*. For mechanisms constructed from such procedures, the following is an implication of Theorem 2.4.

Corollary 2.3. *If $M\colon \mathscr{D}^n \to \Delta\mathscr{A}_k$ is a 2-vetoers mechanism constructed from a finite procedure and $\mathscr{D} \supseteq \mathscr{R}$ for some negative leximin domain $\mathscr{R}$, then M is not strategy-proof.*

Proof. Because there is a finite number of nodes in Γ, there is a finite number of strategy profiles in Γ. Because for all $R_N \in \mathscr{D}^n$, $M^{\Gamma}_{g,\rho}(R_N) = \Gamma\big(s_N, \rho\big)$ for some strategy profile s_N, there is also a finite number of lotteries in the range of M. Thus, there must exist probability thresholds for all $t \in A$. But then, Theorem 2.4 applies because M is a 2-vetoers mechanism. Hence, M cannot be strategy-proof. □

A special class of finite procedures extensively used in jury selection feature two voters $j, h \in N$ (the prosecutor and the defense) sequentially vetoing candidates (potential jurors) among sets of candidates drawn at random from A (the pool).[11] Corollary 2.3 shows that any 2-vetoers mechanism constructed from such a procedure cannot be strategy-proof.

Finally, observe that, by Corollary 2.3, no procedure Γ in which two players can veto a candidate can have dominant strategies on a negative leximin domain. By a revelation principle argument, if such a procedure Γ has dominant strategies for some choice ρ of Nature's moves, then there exists a generalized strategy g^* that makes $M^{\Gamma}_{g^*,\rho}$ a 2-vetoers strategy-proof mechanism, contradicting Corollary 2.3.[12]

2.6 Concluding remarks

Many open questions remain. One concerns the necessity of the sure range condition in Theorem 2.2 and of the threshold conditions in Theorem 2.4. Whether there exists any strategy-proof 2-vetoers mechanism in the absence of these conditions is unknown.

[11] See Flanagan (2015) and Van der Linden (2017).

[12] See Van der Linden (2017, Proposition 2) for more details and Hylland (1980, Section 4) for similar results.

Another question is whether strategy-proof mechanisms exist for weaker veto conditions. In a 2-vetoers mechanism, vetoers are allowed to veto a single candidate, but this candidate can be any candidate. What happens when vetoers can only veto a subset of candidates is an open question.

Finally, Theorems 2.2 and 2.4 rely extensively on domains containing preferences that are *arbitrarily* close to lexicographic, maximax, and minimin preferences. How much flexibility can be gained by further constraining the domain of preferences has not been determined.[13] The proofs of Theorems 2.2 and 2.4 suggest that any possibility result would depend on a combination of restrictions on the richness of (i) the domain of preferences of the vetoers and (ii) the range of the mechanism.

Appendix

Proof of Lemma 2.1. By Remark 2.1, it is sufficient to show that $\mathscr{R}_{add}$ is both minimin and maximax.

$\mathscr{R}_{add}$ is minimin. For any $t \in A$ and any $C \in \mathscr{A}_k$ with $t \notin C$, consider any preference $R^r \in \mathscr{R}_{add}$ for $r > 0$ that can be represented by a utility function on candidates u^r satisfying

(a) $u^r(t) = -r$,

(b) for all $t' \in A$ with $t' \neq t$, $u^r(t') = c_{t'}$ for some constant $c_{t'} \in \mathbb{R}$, and

(c) for all $a \in C$ and all $b \in A \backslash C$, $c_a > c_b$.

By (a) and (b), for any $\varepsilon > 0$, there exists r sufficiently large such that (i) is satisfied in the definition of a minimin domain. Also, (ii) is satisfied by (c).

$\mathscr{R}_{add}$ is maximax. For any $t \in A$ and $C \in \mathscr{A}_k$ with $t \notin C$, consider any preference $R^r \in \mathscr{R}_{add}$ for $r > 0$ that can be represented by a utility function on candidates u^r satisfying

(a) $u^r(t) = r$,

[13] In this respect, see Dutta et al. (2006) who study the extension of an impossibility result of Hylland (1980) to domains in which utility functions must take values in a discrete utility grid.

(b) for all $t' \in A$ with $t' \neq t$, $u^r(t') = c_{t'}$ for some constant $c_{t'} \in \mathbb{R}$, and

(c) for all $a \in C$ and all $b \in A \backslash (C \cup \{t\})$, $c_a > c_b$.

By (a) and (b), for any $\varepsilon > 0$, there exists r sufficiently large such that (i) is satisfied in the definition of a maximax domain. Also, (ii) is satisfied by (c). ■

Proof of Lemma 2.2. Consider any such preference $R_j^{\varepsilon} \in \mathscr{D}$. Because j is a vetoer, there exists $R_j^t \in \mathscr{D}$ such that

$$M(R_j^t, R_{-j})(t) = 0 \qquad \text{for all } R_{-j}. \tag{2.11}$$

If $M(R_j^t, R_{-j}^*)(t) < M(R_j^{\varepsilon}, R_{-j}^*)(t) - \varepsilon$ for some R_{-j}^*, then by (i) in the definition of a minimin domain, $M(R_j^t, R_{-j}^*)\ P_j^{\varepsilon}\ M(R_j^{\varepsilon}, R_{-j}^*)$, contradicting strategy-proofness. Thus, for all R_{-j} we must have $M(R_j^t, R_{-j})(t) \geq M(R_j^{\varepsilon}, R_{-j})(t) - \varepsilon$. Hence, by (2.11), $M(R_j^{\varepsilon}, R_{-j})(t) \leq \varepsilon$ for all R_{-j}. ■

Proof of Lemma 2.3. Consider any $R_j \in \mathscr{D}$ such that (2.4) holds. Such an R_j must exist in $\mathscr{D}$ by (ii) in the definition of a minimin domain. In order to derive a contradiction, assume that

$$L^*(C) < 1 \qquad \text{for some } L^* \in O_{-j}(R_j). \tag{2.12}$$

By the definition of a lottery, $L^*(C) < 1$ implies $L^*(t) = \varepsilon + \gamma$ for some $t \in A \backslash C$ and some $\varepsilon > 0$ and $\gamma > 0$. By assumption, for all $\delta > 0$, there exists a preference $R^{\delta} \in \mathscr{D}$ satisfying (i) and (ii) in the definition of a maximax domain, with (a) t as the "best" candidate, (b) C as the best committee not containing t, and (c) δ replacing ε in the definition. But then Proposition 2.2 and $L^* \in O_{-j}(R_j)$ imply

$$M\left(R_j, R^{\delta}, \ldots, R^{\delta}\right)(t) \geq \varepsilon \qquad \text{for all } \delta < \gamma. \tag{2.13}$$

This inequality holds because the preference R^{δ} has a tolerance of $0 < \delta < \gamma$ for a decrease in the selection probability of t. Thus, $M\left(R_j, R^{\delta}, \ldots, R^{\delta}\right)(t) < \varepsilon$ implies that the lottery selected by M is worse for the preferences R^{δ} than $L^* \in O_{-j}(R_j)$ because $L^*(t) = \varepsilon + \gamma$, contradicting Proposition 2.2.

By the definition of a lottery and because $t \notin C$, (2.13) implies

$$M\left(R_j, R^{\delta}, \ldots, R^{\delta}\right)(C) < 1. \tag{2.14}$$

Because $\mathscr{R}$ is a minimin domain, by Lemma 2.2, there exists a sequence of preferences $\{R_j^r\}_{r=1}^{\infty}$ in $\mathscr{D}$ such that C is the most preferred committee for all $r > 0$ (see (2.4)) and

$$\lim_{r\to\infty} M\left(R_j^r, R^{\delta}, \ldots, R^{\delta}\right)(t) = 0. \tag{2.15}$$

Let $L^r := M\left(R_j^r, R^{\delta}, \ldots, R^{\delta}\right)$ for all $r > 0$.

We now show that

$$\lim_{r\to\infty} L^r(C) = 1. \tag{2.16}$$

By Proposition 2.1, because C is the most preferred committee for R_j^r and because C is in the range, $C \in O_{-j}(R_j^r)$ for all $r > 0$. But then by Proposition 2.2, we must have

$$L^r\ R^{\delta}\ C \qquad \text{for all } r > 0.^{14} \tag{2.17}$$

Let $\widehat{C} \in \mathscr{A}_k$ be (one of) the second most preferred committee(s) according to R^{δ} among the committees that do not contain t; that is,

$$\widehat{C}\ R^{\delta}\ C' \qquad \text{for all } C' \in \mathscr{A}_k \backslash \{C\} \text{ with } t \notin C'. \tag{2.18}$$

[14] The argument is similar to the one used to prove (2.13).

Because R^{δ} satisfies (ii) in the definition of a maximax domain with C as the best committee not containing t, we have

$$C\, P^{\delta}\, \widehat{C}. \tag{2.19}$$

We can now rewrite (2.17) in utility terms as follows:

$$L^r(C)U^{\delta}(C) + \sum_{\{S\in\mathscr{A}_k|t\in S\}} L^r(S)U^{\delta}(S) + \sum_{\{S\in\mathscr{A}_k|t\notin S \text{ and } S\neq C\}} L^r(S)U^{\delta}(S) \geq U^{\delta}(C).$$

By (2.18), this implies

$$\begin{aligned} L^r(C)U^{\delta}(C) &+ \sum_{\{S\in\mathscr{A}_k|t\in S\}} L^r(S)U^{\delta}(S) \\ &+ \left(1 - L^r(C) - \sum_{\{S\in\mathscr{A}_k|t\in S\}} L^r(S)\right) U^{\delta}(\widehat{C}) \geq U^{\delta}(C). \end{aligned}$$

Finally, because $U^{\delta}(C) - U^{\delta}(\widehat{C}) > 0$ by (2.19), we have

$$\begin{aligned} &L^r(C) \geq \\ &\frac{U^{\delta}(C) - \left(1 - \sum_{\{S\in\mathscr{A}_k|t\in S\}} L^r(S)\right) U^{\delta}(\widehat{C})}{\left(U^{\delta}(C) - U^{\delta}(\widehat{C})\right)} - \frac{\sum_{\{S\in\mathscr{A}_k|t\in S\}} L^r(S)U^{\delta}(S)}{\left(U^{\delta}(C) - U^{\delta}(\widehat{C})\right)}. \end{aligned} \tag{2.20}$$

By (2.15), $L^r(t)$ tends to 0 as $r \to \infty$, which implies that

$$\lim_{r\to\infty} \sum_{\{S\in\mathscr{A}_k|t\in S\}} L^r(S) = 0. \tag{2.21}$$

By (2.21), the first term on the right-hand side of (2.20) tends to 1 as $r \to \infty$. Similarly, the second term on the right-hand side of (2.20) tends to 0 as $r \to \infty$. Overall, the right-hand side of (2.20) tends to 1 as $r \to \infty$ and therefore $L^r(C)$ must also tend to 1 as $r \to \infty$, which proves (2.16).

Together (2.4), (2.14) and (2.16) imply that there exists r sufficiently large such that

$$M\left(R_j^r, R^\delta, \ldots, R^\delta\right)\ P_j\ M\left(R_j, R^\delta, \ldots, R^\delta\right) \qquad \text{for all } \delta < \gamma \tag{2.22}$$

contradicting strategy-proofness. ■

Proof of Lemma 2.4. By Remark 2.2, it is again sufficient to show that $\mathscr{R}_{add}$ is negative leximin. For any X, any $\succ$, and any $t \in A \backslash X$, consider any preference $R^r \in \mathscr{R}_{add}$ for $r > 0$ that can be represented by a utility function on candidates satisfying

(a) $u^r(t) = r$,

(b) $u^r(b) = 0$ for all $b \in A \backslash (X \cup \{t\})$, and

(c) $u^r(\succ_h) = -(r^{(h+1)})$, for all $h \in \{1, \ldots, x\}$.

For any $\varepsilon > 0$, there exists r sufficiently large such that R^r satisfies both (2.7) and (2.8).

It is easy to see that R^r satisfies (2.8) for all $r > 0$. Let us illustrate the argument for the first part of (2.7). To prove the first part of (2.7), we need to show that

$$[L(\succ_x) < L'(\succ_x) - \varepsilon] \Rightarrow L\ P^r\ L'. \tag{2.23}$$

For any $\varepsilon > 0$ and any $L, L' \in \Delta\mathscr{A}_k$ with

$$L(\succ_x) < L'(\succ_x) - \varepsilon, \tag{2.24}$$

we need to find some $r_\varepsilon > 0$ which guarantees that

$$\sum_{S \in \mathscr{A}_k} L(S) \sum_{t \in S} u^{r_\varepsilon}(t) > \sum_{S \in \mathscr{A}_k} L'(S) \sum_{t \in S} u^{r_\varepsilon}(t). \tag{2.25}$$

Note that (2.25) is equivalent to

$$\sum_{t\in\{1,\dots,a\}} L(t)u^{r_\varepsilon}(t) > \sum_{t\in\{1,\dots,a\}} L'(t)u^{r_\varepsilon}(t),$$

which implies that

$$\big(L(\succ_x)-L'(\succ_x)\big)u^{r_\varepsilon}(\succ_x) > \sum_{t\in A\setminus\{\succ_x\}} L'(t)u^{r_\varepsilon}(t) - \sum_{t\in A\setminus\{\succ_x\}} L(t)u^{r_\varepsilon}(t). \tag{2.26}$$

First, consider the left-hand side of (2.26). By the construction of R^r and (2.24),

$$\big(L(\succ_x)-L'(\succ_x)\big)u^{r_\varepsilon}(\succ_x) < (-\varepsilon)\big(-({r_\varepsilon}^{(x+1)})\big) = \varepsilon {r_\varepsilon}^{(x+1)}. \tag{2.27}$$

Second, consider the first term on the right-hand side of (2.26). Observe that for all $L\in\Delta\mathscr{A}_k$, the sum of the candidates' selection probabilities is $\sum_{t\in\{1,\dots,a\}} L(t)=k$. Thus, by the construction of R^r,

$$\sum_{t\in A\setminus\{\succ_x\}} L'(t)u^{r_\varepsilon}(t) \le r_\varepsilon \sum_{t\in A\setminus\{\succ_x\}} L'(t) \le r_\varepsilon k.$$

Finally, consider the second term on the right-hand side of (2.26). We have

$$\left(-\sum_{t\in A\setminus\{\succ_x\}} L(t)u^{r_\varepsilon}(t)\right) \le -\big({r_\varepsilon}^{(x)}\big)\left(-\sum_{t\in A\setminus\{\succ_x\}} L(t)\right) \le {r_\varepsilon}^{(x)}k.$$

From the two last displayed inequalities we obtain

$$\sum_{t\in A\setminus\{\succ_x\}} L'(t)u^{r_\varepsilon}(t) - \sum_{t\in A\setminus\{\succ_x\}} L(t)u^{r_\varepsilon}(t) \le k\big(r_\varepsilon + {r_\varepsilon}^{(x)}\big). \tag{2.28}$$

Together, (2.27) and (2.28) imply that (2.26) holds provided

$$\varepsilon {r_\varepsilon}^{(x+1)} > k\big(r + {r_\varepsilon}^{(x)}\big)$$

or, equivalently,

$$\frac{r_\varepsilon^x}{1+r_\varepsilon^{(x-1)}} > \frac{k}{\varepsilon}. \tag{2.29}$$

Because the left-hand side of (2.29) tends to ∞ as $r_\varepsilon \to \infty$, there must exist a r_ε sufficiently large such that this inequality holds.

It is relatively straightforward to adapt the above argument to the $(x-1)$ other parts of (2.7). For example, for

$$\left[L(\succ_x) = L'(\succ_x),\ L(\succ_{x-1}) > L'(\succ_{x-1}) - \varepsilon\right] \Rightarrow L\,P^r\,L$$

the argument can be repeated with the second line of (2.26) simplified to

$$\begin{aligned}&\big(L(\succ_{x-1}) - L'(\succ_{x-1})\big)u^{r_\varepsilon}(\succ_{x-1})\\ &\qquad\qquad > \sum_{t\in A\setminus\{\succ_x,\succ_{x-1}\}} L'(t)u^{r_\varepsilon}(t) - \sum_{t\in A\setminus\{\succ_x,\succ_{x-1}\}} L(t)u^{r_\varepsilon}(t)\end{aligned}$$

because the terms $L'(\succ_x)u^{r_\varepsilon}(\succ_x)$ and $L(\succ_x)u^{r_\varepsilon}(\succ_x)$ cancel out.

After repeating this argument for each of the x components of (2.7), one obtains a set of thresholds $\{r_\varepsilon^x, r_\varepsilon^{x-1}, \ldots, r_\varepsilon^1\}$ for each of the components of (2.7) to hold. Recall that (2.8) is satisfied whenever $r > 0$. Because x is finite, $\bar{r}_\varepsilon := \max\{0.1, r_\varepsilon^x, r_\varepsilon^{x-1}, \ldots, r_\varepsilon^1\}$ is well defined. Then, because the left-hand side of (2.29) and its counter-parts for the other components of (2.7) are increasing in r_ε, $R^{\bar{r}_\varepsilon}$ satisfies (2.7) and (2.8), the desired result. ■

Proof of Lemma 2.5. We need to show that there exists $R_j^* \in \mathscr{D}$ such that for all $t \in A\setminus C$

$$L(t) = 0 \qquad \text{for all } L \in O_{-j}(R_j^*). \tag{2.30}$$

Let $X = A\setminus C$ and $\succ$ be any strict ordering of X. Because $\mathscr{R}$ is a negative leximin domain, for all $\varepsilon > 0$ there exists $R^{\varepsilon}_{j} \in \mathscr{D}$ that ranks lotteries by a lexicographic order of the selection probability of candidates in X as defined in (2.7).[15] We will prove that there exists $\varepsilon > 0$ such that R^{ε}_{j} satisfies (2.30). The argument is by induction on the elements of X. We provide the two first steps in detail.

Step 1: There exists $\varepsilon > 0$ such that $L(\succ_x) = 0$ for all $L \in O_{-j}(R^{\varepsilon}_{j})$.

By the threshold assumption, it is sufficient to show that for any $\varepsilon > 0$,

$$L(\succ_x) \leq \varepsilon \qquad \text{for all } L \in O_{-j}(R^{\varepsilon}_{j}). \tag{2.31}$$

The claim in Step 1 then follows by choosing ε with $0 < \varepsilon < \tau_{\succ_x}$, where $\tau_{\succ_x}$ is the threshold for $\succ_x$.

The proof of (2.31) is similar to the proof of Lemma 2.2. Recall that because j is a vetoer, there exists $R^{\succ_x}_{j}$ such that $L(\succ_x) = 0$ for all $L \in O_{-j}(R^{\succ_x}_{j})$. By strategy-proofness, j can never benefit from declaring $R^{\succ_x}_{j}$ instead of her true preference R^{ε}_{j}. In particular, for any $L \in O_{-j}(R^{\varepsilon}_{j})$, voter j must weakly prefer L to the worst possible lottery for which $L(\succ_x) = 0$, say $\underline{L}$. By the definition of a negative leximin preference in (2.7), if $\underline{L}(\succ_x) < L(\succ_x) - \varepsilon$ then $\underline{L}(\succ_x)\ P^{\varepsilon}_{j}\ L(\succ_x) - \varepsilon$. Because $L(\succ_x)\ R^{\varepsilon}_{j}\ \underline{L}(\succ_x) - \varepsilon$, we thus have $\underline{L}(\succ_x) \geq L(\succ_x) - \varepsilon$. But because $\underline{L}(\succ_x) = 0$, this implies that $\varepsilon \geq L(\succ_x)$, the desired result.

Step 2: There an exists $\varepsilon > 0$ such that $L(\succ_x) = L(\succ_{x-1}) = 0$ for all $L \in O_{-j}(R^{\varepsilon}_{j})$.

By Step 1, it is sufficient to show that there exists an ε with $0 < \varepsilon < \tau_{\succ_x}$ such that

$$L(\succ_{x-1}) = 0 \qquad \text{for all } L \in O_{-j}(R^{\varepsilon}_{j}). \tag{2.32}$$

Applying the threshold assumption again, (2.32) holds provided that for all ε with $0 < \varepsilon <$

[15] Note that $\#C = k$ and, hence, $\#X \leq (a-k)$ in accordance with the definition of a negative leximin domain. Here the choice of $t \in C$ in (2.8) is irrelevant.

$\tau_{\succ_x}$,

$$L(\succ_{x-1}) \leq \varepsilon \qquad \text{for all } L \in O_{-j}(R_j^\varepsilon). \tag{2.33}$$

In order to derive a contradiction, assume that there exists an ε with $0 < \varepsilon < \tau_{\succ_x}$ and some $\gamma > 0$ such that

$$L^\varepsilon(\succ_{x-1}) = \varepsilon + \gamma \qquad \text{for some } L^\varepsilon \in O_{-j}(R_j^\varepsilon). \tag{2.34}$$

Let h be any vetoer with $h \neq j$. Because $\mathscr{R}$ is a negative leximin domain, for all $\delta > 0$ there exists $R_h^\delta \in \mathscr{D}$ satisfying (2.7) for $X = \{\succ_x\}$, and satisfying (2.8) for $t = (x-1)$ (where δ replaces ε in both (2.7) and (2.8)). Because h is a vetoer, the argument in Step 1 applies to h, and for any δ with $0 < \delta < \tau_{\succ_x}$,

$$M(R_h^\delta, R_{-h})(\succ_x) = 0 \qquad \text{for all } R_{-h}. \tag{2.35}$$

Together, (2.34) and $0 < \varepsilon < \tau_{\succ_x}$ imply

$$L^\varepsilon(\succ_{x-1}) = \varepsilon + \gamma \text{ and } L^\varepsilon(\succ_x) = 0. \tag{2.36}$$

But then, because $L^\varepsilon \in O_{-j}(R_j^\varepsilon)$, by Proposition 2.2,

$$M\left(R_j^\varepsilon, R_h^\delta, \ldots, R_h^\delta\right) \; R_h^\delta \; L^\varepsilon. \tag{2.37}$$

Because $\gamma > 0$, there exists δ with $0 < \delta < \min\{\gamma, \tau_{\succ_x}\}$. Observe that by the construction of negative leximin preferences in (2.7) and by the threshold assumption, for any such δ,

$$L^\varepsilon \; P_h^\delta \; L \text{ for any } L \text{ with } L(\succ_x) > 0 \text{ or with } L(\succ_x) = 0 \text{ and } L(\succ_{x-1}) \leq \varepsilon. \tag{2.38}$$

Hence, (2.37) and (2.38) imply that

$$M\left(R_j^{\varepsilon}, R_h^{\delta}, \ldots, R_h^{\delta}\right)(\succ_x) = 0 \text{ and } M\left(R_j^{\varepsilon}, R_h^{\delta}, \ldots, R_h^{\delta}\right)(\succ_{x-1}) > \varepsilon. \tag{2.39}$$

Because j is a vetoer, by (2.35), there exists $R_j^{\succ_{x-1}}$ such that

$$M\left(R_j^{\succ_{x-1}}, R_h^{\delta}, \ldots, R_h^{\delta}\right)(\succ_{x-1}) = M\left(R_j^{\succ_{x-1}}, R_h^{\delta}, \ldots, R_h^{\delta}\right)(\succ_x) = 0. \tag{2.40}$$

But then by the construction of a negative leximin preference,

$$M\left(R_j^{\succ_{x-1}}, R_h^{\delta}, \ldots, R_h^{\delta}\right) \; P_j^{\varepsilon} \; M\left(R_j^{\varepsilon}, R_h^{\delta}, \ldots, R_h^{\delta}\right), \tag{2.41}$$

contradicting strategy-proofness. The remaining steps follow the same inductive pattern.

■

Chapter 3

Bounded rationality and the choice of jury selection procedures

3.1 Introduction

It is customary to let the parties involved in a jury trial dismiss some of the potential jurors without justification. Procedures for dismissal are known as *peremptory challenge* procedures. Such procedures are used in many countries, including the United States.[1] A variety of procedures are used in practice. These procedures differ notably in their strategic complexity. More complex procedures give an unfair advantage to parties that are strategically skilled or can devote ample resources to hiring jury consultants. It is therefore important to identify the procedures that are strategically simple in order to level the playing field among parties. In this paper, I introduce the concept of a *dominance threshold*, a new measure of strategic complexity based on *level-k* thinking, and I use this measure to compare the complexity of some challenge procedures commonly used in practice (see Crawford et al., 2013, for a survey of the level-k literature).

Fairness is an important issue in jury selection. One feature of a procedure that impacts fairness is its strategic complexity. If a procedure is complex, parties with better strategic skills are likely to secure more favorable juries. This is particularly relevant in jury selection where the parties invest significant resources for developing an effective strategy. For example, jury selection consultancy has become a well-established industry.[2] Using

[1] In *Swain v. Alabama*, the Supreme Court affirmed that "the [peremptory] challenge is one of the most important of the rights secured to the accused." (LaFave et al., 2009). Following *Batson v. Kentucky*, a party can disqualify a peremptory challenge by her opponent if she can prove that the challenge was based on a set of characteristics including race or gender. However, *Batson v. Kentucky* is notoriously hard to implement and judges rarely rule in favor of Batson challenges (Marder, 2012; Daly, 2016).

[2] The widespread use of jury consultants is evidenced by the existence of the American Society of Trial Consultants, and its publication "The Jury Expert: The Art and Science of Litigation Advocacy". Jury consultants explicitly describe how part of their job is concerned with the strategic use of challenges. Jury consultant Roy Futterman for example writes: "Caditz argues that [...] jury selectors pay [...] little to no attention to the strategic use of strikes [i.e., peremptory challenges]. [...] it is a bit of a reach to say that strategy is barely utilized. In my experience, [...] [jury selection] comes closer to a long battle of stealth,

strategically simple procedures limits the impact on the selected jury of differences in the parties' ability to strategize or in their financial means to hire jury consultants.

Comparing the strategic complexity of jury selection procedures presents two challenges. First, jury selection procedures are indirect mechanisms because the parties' actions consist of dismissing jurors rather than revealing their preferences.[3] Second, in some procedures commonly used in practice, the parties submit their challenges *simultaneously*, which induces games of imperfect information. These two difficulties make it impossible to apply measures of strategic complexity previously developed in the literature.[4]

I overcome these difficulties by introducing the concept of a *dominance threshold*. Given some assumption about the strategies her opponent could play — henceforth, a *model* of her opponent — a party has a dominant strategy if one of her strategies is a best response to *any* strategy of her opponent that is consistent with her model. The objective is to identify models for which the parties have dominant strategies. This is accomplished by iteratively eliminating strategies that are never best responses. The dominance threshold is the number of rounds of elimination needed to reach models in which both parties have a dominant strategy. The dominance threshold measures the complexity of the model of their opponent that the parties need in order to have a dominant strategy. For example, a dominance threshold of 1 corresponds to the parties having a dominant strategy given *any* model of their opponent. When the dominance threshold is 2, the parties only need to know that their opponent is a best responder in order to have a dominant strategy.[5]

counter-punches, misdirection, and hand-to-hand combat than a lofty academic experience." (Excerpt from Roy Futterman's answer to David Caditz's August 20, 2014 post on http://www.thejuryexpert.com/).

[3] A procedure based on direct preference revelation would go against the idea of allowing the parties to challenge jurors *without justification*.

[4] For example, the measures in Pathak and Sönmez (2013) and Arribillaga and Massó (2015) are defined for direct games only and cannot be applied to existing jury selection procedures. In indirect games, de Clippel et al. (2014) recommend focusing on procedures that can be solved in two rounds of backward induction. More generally, this suggests using the number of rounds of backward induction needed to solve a procedure as a measure of its complexity. By its nature, such a measure is only applicable to games of perfect information and therefore excludes simultaneous moves (when games are modeled in extensive form, any simultaneous move implies that the game is of imperfect information). It also has the disadvantage of being sensitive to the addition of inconsequential actions. Neither of these limitations is shared by the *dominance threshold*, the new measure I propose.

[5] The related concept of a "rationality threshold" was introduced by Ho et al. (1998). For a *given* assump-

Many judges appear to share the concern for selecting strategically simple procedures. They have developed procedures that attempt to limit the parties' ability to strategize. In a report on judges' practices regarding peremptory challenges, Shapard and Johnson (1994, p. 6) write:

> "Some judges require that peremptories be exercised [following procedure X] [...]. This approach [...] makes it more difficult to pursue a strategy prohibited by *Batson* (or any other strategy). [...] A more extreme approach to the same end is [procedure Y] [...]. This approach imposes maximum limits on counsel's ability to employ peremptories in a strategic manner."[6]

Using the dominance threshold as a measure of strategic complexity provides new insights and overturns some commonly held beliefs about jury selection procedures. Shapard and Johnson (1994, p. 6) write:

> "Other judges, for the same purposes [(limiting the parties' ability to strategize)], allow all peremptories to be exercised after all challenges for cause, but with the parties making their choices 'blind' to the choices made by opposing parties (in contrast to alternating "strikes" from a list of names of panel members)."[7]

I show that, contrary to these judges' beliefs, procedures in which challenges are sequential tend to be strategically simpler than procedures in which challenges are simultaneous: By generating imperfect information games, simultaneous procedures increase the amount of guesswork needed to determine optimal strategies.

I also study the design of "maximally simple" jury selection procedures. I show that it is *im*possible to construct a "reasonable" procedure that allows the parties to challenge

tion about the strategies of unsophisticated players, the rationality threshold measures the number of rounds of iterated *best responses* needed to reach an equilibrium. In contrast, the dominance threshold relies on iterated *elimination of never best responses* and does not require a specific assumption about the nature of unsophisticated plays. See Section 3.4 for a formal definition.

[6] See footnote 1 regarding *Batson v. Kentucky*.

[7] Unlike peremptory challenges, challenges *for cause* must be based on biases recognized by law, such as being a direct relative of one of the parties.

jurors and always have a dominant strategy. Hence, the smallest achievable dominance threshold is 2. Such a minimal dominance threshold is attained by a procedure that I call *sequential one-shot* in which the parties sequentially submit a single list of jurors that they want to challenge.

Although the focus of this paper is the study of jury selection procedures, the dominance threshold applies more generally as a general measure of strategic complexity. Unlike previous measures in the literature, the dominance threshold can be used to compare the strategic complexity of *any* pair of games, including indirect games and games of imperfect information. The dominance threshold is the first such measure of strategic complexity proposed in the literature.

Related Literature.— This paper differs from the previous game theoretic literature on jury selection procedures in at least two ways (see Flanagan (2015) for a recent review). First, the literature has focused on subgame perfect equilibrium as a solution concept.[8] Subgame perfection requires a high level of strategic sophistication, especially in complex procedures. By relying on the concept of a dominance threshold, this paper accounts for the possibility of boundedly rational parties. I show how the dominance threshold, which measures the "amount" of common knowledge and rationality needed to have a dominant strategy, can be used to measure the strategic complexity of a procedure.

Second, here, jury selection is studied from the point of view of mechanism design.[9] Most of the literature focuses on the characterization and properties of equilibria of different procedures.[10] When the performance of procedures is compared, it is typically in terms of their effects on the composition of the jury. These comparisons have yielded few policy recommendations.[11] In contrast, this paper adopts a traditional mechanism design approach and compares procedures with respect to the standard objective of limiting the

[8] Two exceptions are Bermant (1982) and Caditz (2015).

[9] In this respect, the closest paper is de Clippel et al. (2014), which takes a mechanism design perspective but studies the selection of a *single* arbitrator.

[10] See Roth et al. (1977), Brams and Davis (1978), DeGroot and Kadane (1980), Kadane et al. (1999), Alpern and Gal (2009), and Alpern et al. (2010).

[11] See, however, Bermant (1982) and Flanagan (2015, Section 4.2).

parties' ability to strategize. This later approach enables a clear comparison of some of the procedures used in practice.

The paper is organized as follows. Section 3.2 introduces the model as well as several examples of jury selection procedures and a general class thereof. In Section 3.3, I show that most "reasonable" jury selection procedures do not have dominant strategies. Section 3.4 formally introduces the concept of a dominance threshold. The dominance threshold is then applied to comparing the strategic complexity of jury selection procedures in Sections 3.5 and 3.6. Proofs may be found in the Appendix.

3.2 Model and procedures

I focus on *struck* procedures. In addition to peremptory challenges which require no justification, the parties can raise challenges *for cause* which must be based on some bias recognized by law, such as being a direct relative of one of the parties. As explained by Bermant and Shapard (1981, p. 92), the defining feature of a struck procedure "is that the judge rules on all challenges for cause before the parties claim any peremptories. Enough potential jurors are examined to allow for the size of the jury plus the number of peremptory challenges allotted to both sides. In a federal felony trial, for example, the jury size is twelve; the prosecution has six peremptories, and the defense has ten. Under the struck jury method, therefore, 28 potential jurors are cleared through challenges for cause before the exercise of peremptories."[12]

Struck procedures are commonly used in federal courts. In a 1977 survey of judges' practices regarding the exercise of peremptory challenges, 55% of federal district judges reported using a struck procedure (Bermant and Shapard, 1981). Today, the use of a struck procedure is, for example, recommended by law as the preferred method for criminal cases

[12] This contrasts with *strike and replace* procedures in which challenges for cause and peremptory challenges are intertwined. In a strike and replace procedure, prospective jurors who are challenged (either for cause or peremptorily) are replaced by new jurors from the pool and "to one degree or another, counsel exercise their challenges without knowing the characteristics of the next potential juror to be interviewed" (Bermant and Shapard, 1981, p. 93).

other than first-degree murder in Minnesota.[13]

3.2.1 The model

The set of prospective jurors left after all challenges for cause have been raised is $N = \{1, \ldots, n\}$. The defendant is d and the plaintiff is p. The defendant and the plaintiff are allowed c_d and c_p peremptory challenges, respectively. Out of N, a jury J of b jurors must be selected. The jurors in J are the **impaneled** jurors. As explained above, when struck procedures are considered, $n = b + c_d + c_p$ in order to allow the parties to challenge up to c_d and c_p jurors.

Let $\mathscr{J}$ be the set of juries containing b jurors and $\Delta\mathscr{J}$ the set of lotteries on $\mathscr{J}$. In some cases, it is possible that after all challenges have been raised, more than b jurors remain unchallenged. In this case, I will assume that the b impaneled jurors are chosen at random among the unchallenged jurors. As a consequence, the parties have expected utility preferences R_d and R_p on $\Delta\mathscr{J}$, respectively, with corresponding Bernoulli utility functions u_d and u_p on $\mathscr{J}$ (instead of preferences on $\mathscr{J}$). A pair of preferences (R_p, R_d) is called a **(preference) profile** and a quintuple (R_d, R_p, c_d, c_p, b) a **(jury selection) problem**.

Example 3.1. If $\pi(J)$ is the probability that jury J convicts the defendant, then party i's preference is represented by a utility function of the form $u_i = v_i(\pi(J))$, where v_p is increasing and v_d decreasing.

Throughout, I assume that preferences on juries are ***separable***, i.e., if replacing juror h by juror j in jury J is an improvement according to u_i, then the same is true when h is replaced by j in any other jury J'. Formally, for any $i \in \{d, p\}$ and any $J, J' \in \mathscr{J}$, $h \in J \cap J'$ and $j \in N \backslash (J \cup J')$, $u_i(J \cup \{j\} \backslash \{h\}) \geq u_i(J)$ if and only if $u_i(J' \cup \{j\} \backslash \{h\}) \geq u_i(J')$.

Most of the results in this paper also hold when separability is not assumed. Separability eases the exposition because it implies that the preferences R_d and R_p induce well-defined

[13] Minnesota Court Rules, Criminal Procedure, Rule 26.02, Subd.4.(3)b).

preferences for *individual* jurors. It is also assumed that the preferences for jurors induced by R_d and R_p are *strict*. Slightly abusing the notation, R_i serves to denote i's preference for individual jurors as well as i's preference for juries.

An extreme kind of profile is a **juror inverse** profile. A profile is juror inverse if R_d and R_p induce inverse preferences on jurors (i.e., for all $j, h \in N$, $j \; R_p \; h$ if and only if $h \; R_d \; j$). Unlike separability, which is assumed throughout the paper, juror inverse profiles are only considered as a special case.

3.2.2 Procedures

As attested to by Bermant and Shapard (1981), a wide variety of struck procedures are used by judges. One common type of struck procedure are procedures that I call **one-shot**. In a one-shot procedure, each party $i \in \{d, p\}$ submits a *single* list of up to c_i jurors in N that i wants to challenge. Depending on the procedure, the parties submit their lists simultaneously (**one-shot**$_M$) or sequentially (**one-shot**$_Q$). The impaneled jurors are the jurors in N who have not been challenged. If more than b jurors are left unchallenged, the b impaneled jurors are drawn at random from among the unchallenged jurors.[14]

Another common type of struck procedure are the procedures that I call **alternating**. Alternating procedures proceed through a succession of rounds in which the parties can challenge as many jurors in N as they have challenges left. Again, an alternating procedure can be either simultaneous (**alternating**$_M$) or sequential (**alternating**$_Q$) depending on whether challenges are submitted simultaneously or sequentially in each round. In alternating$_M$, if both parties challenge the same juror in a given round, both parties are charged with the challenge and can challenge one less juror.

Alternating procedures stop when neither of the parties has challenges left, or when

[14] The use of one-shot$_M$ is documented by Bermant (1982, Step. 5, Comments by Judges Feikens and Voorhees). Bermant (1982, Step. 5, Comments by Judge Enright) shows that a procedure in which the parties alternate challenges *twice* has been used in practice, with each party allowed to challenge up to $\frac{c_i}{2}$ jurors in each round.

both parties abstain from challenging a juror in a single round. The impaneled jurors are the jurors left unchallenged in N, or a random draw of b of these jurors if more than b jurors are left unchallenged.[15]

One-shot and alternating procedures are members of the class of **N-struck procedures** in which parties take turns challenging jurors from N for a number of rounds.[16] Formally, every N-struck procedure consists of a maximum of $f \geq 1$ rounds, where f differs between procedures. Each round $r \in \{1,\ldots,f\}$ is characterized by a maximum number of challenges $x_i^r \geq 1$ for each party, with $\sum_{r=1}^{f} x_i^r \geq c_i$. The number of challenges party i has left in round r is l_i^r, with $l_i^1 = c_i$. In each round r:

(a) The parties can challenge up to $min\{x_i^r, l_i^r\}$ jurors among the jurors in N who have not yet been challenged. Challenges are sequential if the procedure is sequential, and simultaneous if the procedure is simultaneous.

(b) For each party $i \in \{d,p\}$, the number of challenges left is decreased by the number of jurors that the party challenged in (a) (i.e., l_i^{r+1} equals l_i^r minus the number of jurors that the party challenged in (a)).

The procedure terminates when no party has challenges left, when round f is reached, or when both parties abstain from challenging any juror in a single round. The jurors left unchallenged when the procedure terminates are the impaneled jurors. If more than b jurors are left unchallenged when the procedure terminates, the b impaneled jurors are drawn at random from the unchallenged jurors.[17]

[15] Alternating$_Q$ is recommended by law as the preferred method for criminal cases other than first-degree murder in Minnesota (Minnesota Court Rules, Criminal Procedure, Rule 26.02, Subd.4.(3)b)). Simultaneous challenges are used in alternating procedures for civil cases in Tennessee (Tennessee Court Rules, Rules of Civil Procedure, Rule 47.03), although the mandated procedure in these cases is of the *strike and replace* type (footnote 12).

[16] The name "N-struck procedure" emphasizes the fact that, in each round, the parties can challenge any juror in N that has not been challenged yet. This is not the case in every struck procedure (Bermant, 1982, Step. 5, Comments by Judge Atkins).

[17] The distribution is arbitrary as long as any remaining juror has a strictly positive probability of being selected.

One-shot procedures are N-struck procedures with $f = 1$ and $x_i^1 = c_i$ for both parties $i \in \{d, p\}$. Alternating procedures are N-struck procedures with $x_i^r = c_i$ for both $i \in \{d, p\}$ and all $r \in \{1, \dots, f\}$, and $f = 2\max_{i \in \{d,p\}} c_i$. Besides one-shot and alternating procedures, the class of N-struck procedures includes, for example, the two-round procedure documented by Bermant (1982, Step. 5, Comments by Judge Enright) and described above.

From a game theoretic point of view, a **(jury selection) procedure** is an extensive game form $\Gamma\colon \mathscr{S}_d \times \mathscr{S}_p \to \Delta\mathscr{J}$ that associates any pair of strategies (s_d, s_p) in some strategy space $\mathscr{S}_d \times \mathscr{S}_p$ with a lottery on juries in $\mathscr{J}$. In this paper, *I restrict attention to pure strategies* in any extensive game form Γ, although all the results also hold when mixed strategies are allowed.

3.3 Impossibility results

Given preference R_i, a **best response** for party i to some strategy s_{-i} of her opponent is a strategy $t_i(s_{-i})$ such that

$$\Gamma\big(t_i(s_{-i}), s_{-i}\big)\ R_i\ \Gamma\big(s_i', s_{-i}\big) \qquad \text{for all } s_i' \in \mathscr{S}_i. \tag{3.1}$$

When $-i$ plays s_{-i} and i plays $t_i(s_{-i})$, party i **best responds** to $-i$. A strategy $s_i \in \mathscr{S}_i$ is **dominant for i given some model** $S_{-i} \subseteq \mathscr{S}_{-i}$ of her opponent if s_i is a best response to *every* strategy $s_{-i} \in S_{-i}$. A **dominant strategy** is a strategy $s_i^* \in \mathscr{S}_i$ that is a best response for i to *any* strategy $s_{-i} \in \mathscr{S}_{-i}$. In other words, a dominant strategy is a strategy that is dominant for i given *any model* of her opponent.

Given some domain of preferences, a **dominant strategy procedure** is a procedure in which both parties have a dominant strategy for every profile in the domain. Dominant strategy procedures are strategically simple because each party can determine an optimal strategy independently of any guess about the strategy of her opponent. Dominant strategy procedures guarantee a form of equality among equals: Two parties having the same preferences but different abilities to form expectations about their opponent's strategy should be able to secure similar outcomes.

It is useful to relate dominant strategies with level-k thinking (see the survey in Crawford et al., 2013). In the level-k terminology, an L_i^0 party is a non-strategic party who could potentially play any strategy. An L_i^1 party assumes that her opponent is L_{-i}^0, makes a guess about the L_{-i}^0 strategy s_{-i}^0 that her opponent will employ, and best responds to s_{-i}^0.[18] Similarly, an L_i^k party assumes that her opponent is L_{-i}^{k-1}, makes a guess about the L_{-i}^{k-1} strategy s_{-i}^{k-1} that her opponent will employ, and best responds to s_{-i}^{k-1}.

Observe that, because an L_{-i}^0 strategy can be any of $-i$'s strategies, i has a dominant strategy if and only if i has an L_i^1 strategy that is a best response to *every* L_{-i}^0 strategy of her opponent. In the language of level-k thinking, a dominant strategy procedure limits the impact of differences in strategic skills because i can determine an optimal strategy independently of her belief about her opponent's level of rationality k_{-i}, or her guess about which $L_{-i}^{k_{-i}}$ strategy her opponent will employ.

Unfortunately, most reasonable procedures that permit challenges do not have a dominant strategy. Consider one-shot$_M$. In one-shot$_M$, i's only best response to any s_{-i} is to challenge her c_i worst jurors among the jurors that $-i$ does not challenge in s_{-i}. As illustrated in Example 3.2, such a best response is highly dependent on the challenges chosen by $-i$. Hence, one-shot$_M$ is not a dominant strategy procedure.

Example 3.2. Suppose that each juror has four challenges ($c_d = c_p = 4$) and one juror has to be selected ($b = 1$). A set of nine prospective jurors $N = \{1, \dots, 9\}$ will therefore remain after all challenges for cause have been raised. Let d's preference on these nine jurors be $1\ R_d\ 2\ R_d\ \dots\ R_d\ 9$. If p challenges the circled jurors in (3.2), then d's best response is to

[18] Recall that I only consider pure strategies. Hence, the set of i's level-0 strategies is the set of i's pure strategy. Again, all the results in this paper hold when mixed strategies are allowed. In particular, the results hold when a party i's level-1 strategies include i's best responses to probabilistic beliefs about the (pure) level-0 strategy that $-i$ will employ, as in Ho et al. (1998).

challenge the squared jurors in (3.2).

$$R_d: \quad 1 \quad \textcircled{2} \quad \textcircled{3} \quad \boxed{4} \quad \textcircled{5} \quad \textcircled{6} \quad \boxed{7} \quad \boxed{8} \quad \boxed{9} \tag{3.2}$$

On the other hand, if p challenges the circled jurors in (3.3), d's best response is to challenge the squared jurors in (3.3).

$$R_d: \quad 1 \quad \textcircled{2} \quad \textcircled{3} \quad \textcircled{4} \quad \textcircled{5} \quad \boxed{6} \quad \boxed{7} \quad \boxed{8} \quad \boxed{9} \tag{3.3}$$

Clearly, the challenge of the squared jurors in (3.3) is not a best response for d to p challenging the circled jurors in (3.2), which shows that one-shot$_M$ is not a dominant strategy procedure on any domain a profile of which contains R_d.

As shown in Proposition 3.1, the preceding example generalizes to the whole class of N-struck procedures and to any problem. Intuitively, in any N-struck procedure, if $-i$ does not challenge any jurors, then i's best response is to challenge her c_i worst jurors. On the other hand, if $-i$ challenges one of the c_i worst jurors of i, say w, then i is better off not challenging w and challenging her other c_i worst jurors. Recall that a (jury selection) *problem* is a quintuple (R_d, R_p, c_d, c_p, b).

Proposition 3.1. *For any problem, (i) the first party does not have a dominant strategy in one-shot$_Q$ and (ii) neither party has a dominant strategy in any N-struck procedure different from one-shot$_Q$.*

Note that one-shot$_M$ is an N-struck procedure different from one-shot$_Q$. Hence, Proposition 3.1 shows that, for every problem, neither party has a dominant strategy in one-shot$_M$. One-shot$_Q$ is the exception among N-struck procedures: It is the only N-struck procedure in which one of the parties — the second party to challenge — has a dominant strategy, although the other party does not for the reason explained before the proposition (see the proof of the proposition for more detail).

Of course, N-struck procedures are only a small subset of all possible jury selection procedures. Other procedures used in practice include the strike and replace procedures (see footnote 12), as well as other struck procedures in which the parties can only challenge from subsets of N in each round (Bermant, 1982, Step. 5, Comments by Judge Atkins). It is therefore natural to ask whether there exists dominant strategy procedures for jury selection outside of the N-struck class. The next proposition shows that if such procedures exist, then they must either deprive a party of her right to challenge at least one juror in N or be so intricate that they are unlikely to be used in practice.

A procedure satisfies **finiteness** if the set of its decision nodes is finite for both parties and for Nature. A procedure satisfies **minimal challenge** if for every prospective juror $j \in N$, both parties $i \in \{d, p\}$ have a strategy $s_i^j \in \mathscr{S}_i$ such that j is never part of the chosen jury when i plays s_i^j.[19] Every N-struck procedure satisfies both finiteness and minimal challenge (strategy s_i^j can, for example, involve challenging juror j — and only juror j — in the first round).

Proposition 3.2. *On the domain of separable preferences, no dominant strategy procedure satisfies both finiteness and minimal challenge.*

In the Appendix, I show that Proposition 3.2 is, in fact, true for smaller domains of profiles, including the domain of additive profiles.

3.4 A measure of strategic complexity

Propositions 3.1 and 3.2 show that most procedures are not strategically simple in the sense that both parties cannot always follow the simple recommendation of playing a dominant strategy. This does not mean, however, that judges should give up on the idea of using procedures that are *as simple as possible*. This section and the next show that, although procedures generally fail to feature dominant strategies, not all procedures are equal in terms of strategic complexity.

[19] That is, the probability that j is chosen given that i plays s_i^j is zero for all s_{-i}.

3.4.1 Motivating example

Brams and Davis (1978, p. 969) have argued that, when the parties have juror inverse preferences, one-shot procedures raise “no strategic questions of timing: given that each side can determine those veniremen [i.e., potential jurors] it believes least favorably disposed to its cause, it should challenge these up to the limit of its peremptory challenges.” This may come as a surprise given Example 3.2 and Proposition 3.1. Certainly, one-shot$_M$ is not a dominant strategy procedure. How can we then make sense of Brams and Davis’ claim? The next example suggests one possible answer.

Example 3.3. Consider one-shot$_M$ with $c_d = c_p = 2$ and $b = 5$. Let d have preference $1 \; R_d \; \ldots \; R_d \; 9$. Also, suppose that the parties have juror inverse preferences.

If d believes that p is best responding to one of her strategies, d knows that p will challenge two of the circled jurors in (3.4).

$$\begin{array}{lccccccccc} R_d: & \textcircled{1} & \textcircled{2} & \textcircled{3} & \textcircled{4} & 5 & 6 & 7 & \boxed{8} & \boxed{9} \\ R_p: & 9 & 8 & 7 & 6 & 5 & \textcircled{4} & \textcircled{3} & \textcircled{2} & \textcircled{1} \end{array} \tag{3.4}$$

Indeed, a best response by p always involves challenging her two worst jurors among the seven jurors that she believes d will not challenge. Therefore, regardless of the jurors p believes that d will challenge, a best response by p can never include p challenging a juror in $\{5,\ldots,9\}$.

Thus, a best response by d to the minimal belief that p is a best responder always consists in challenging her two worst jurors (squared in (3.4)). By symmetry, the same is true for p.

In Example 3.3, one-shot$_M$ “raises no strategic question” because a party only needs to know that her opponent is a best responder in order to have a dominant strategy. For each party i, challenging her c_i worst jurors is a best response to *any* strategy of party $-i$ that is itself a best response to one of i’s strategies. In this sense, each party i has a dominant strategy given a *minimal* model of the strategic behavior of her opponent: the model $S_{-i} = L^1_{-i}$.

In the rest of this section, I generalize this logic to obtain a measure of strategic complexity. This measure is then applied in the next two sections to compare struck procedures for different assumptions for the problem (R_d, R_p, c_d, c_p, b).

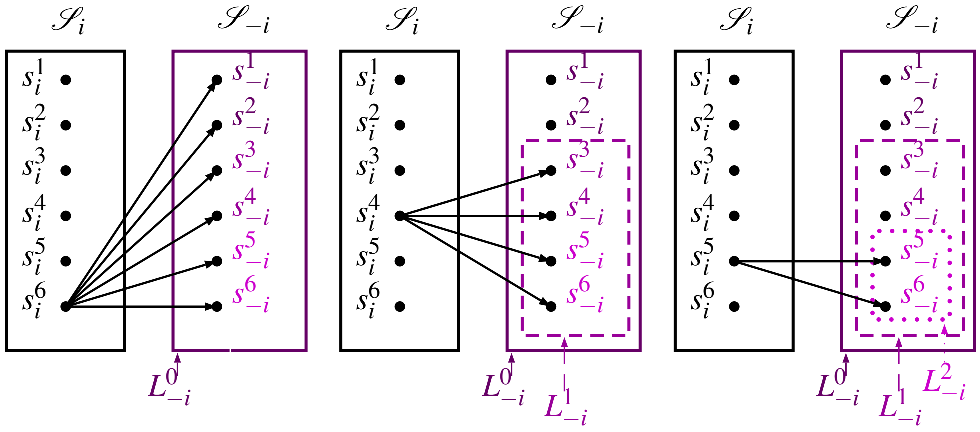
3.4.2 The dominance threshold

As argued above, first-best procedures are procedures in which each party has a dominant strategy *whatever model* she has of her opponent. It is then natural to call a procedure second-best if each party has a dominant strategy *given a minimal model* of her opponent. As suggested in Example 3.3, a meaningful concept of a *minimal model* is for a party to assume that her opponent will play a best response to some of her strategies.

In the language of level-k thinking, a procedure is second-best if each party i has an L^2_i strategy that is a best response to *every* L^1_{-i} strategy of her opponent. Such second-best procedures limit the impact of differences in strategic skills because i's optimal strategy depends *minimally* on her model of $-i$: i only needs to assume that $-i$ is L^1_{-i} to have a dominant strategy.

The difference between first-best and second-best procedures is illustrated in Figure 3.1(a) and (b). In the figure, an arrow from strategy s_i to strategy s_{-i} means that s_i is a best response to s_{-i}. In the first-best procedure represented in (a), party i has a strategy — s^6_i in the figure — that is a best response to every strategy of her opponent (i.e., to every L^0_{-i} strategy). In the second-best procedure represented in (b), party i has a strategy — s^4_i in the figure — that is a best response to every strategy of her opponent that is itself a best response (i.e., to every L^1_{-i} strategy). However, s^4_i does not need to be a best response to *every* L^0_{-i} strategy. For example, in the figure, s^4_i is not a best response to s^1_{-i}.

A second-best procedure guarantees a form of second-best equality among equal parties. Consider two defendants with the same preference who both believe that p is L^1_p and



(a) First-best procedure (b) Second-best procedure (c) Third-best procedure

Figure 3.1: Representation of first-, second-, and third-best procedures.

best responds to one of her strategies. The two defendants might differ in other strategic aspects, such as their ability to guess which of their strategies p best responds to. In a second-best procedure, these differences have no impact: the two defendants play equivalent strategies and secure the same outcome.

Similarly, third-best procedures feature dominant strategies given a model that is *minimally stronger* than in second-best procedures. A natural candidate for such a minimally stronger model is for i to assume that $-i$ is L_{-i}^2 (see Figure 3.1(c)). This logic extends to higher level reasoning.

In procedures with multiple rounds, it is important to ensure that best responses be enforced throughout the game tree. Therefore, the measure of strategic complexity defined below relies on the iterated elimination of strategies that are never best responses *in any subgame* of an extensive game. That is, in each round of elimination, any strategy that fails to be a best response when restricted to *any* subgame of the game is discarded.

Definition 3.1 (Iterated elimination of never best responses)**.** For any procedure Γ and any profile (R_d, R_p), the process of *iterated elimination of never best responses* is defined as follows:

Step 0. For each $i \in \{d, p\}$, the set of $\boldsymbol{L_i^0}$ **(level-0) strategies** is $\mathscr{S}_i$.

Step 1. For each $i \in \{d,p\}$, eliminate from L_i^0 the strategies s_i for which there exists a subgame γ of Γ such that the restriction $s_i|_\gamma$ of s_i to γ is not a best response to any $s_{-i}|\gamma$ in γ.

The remaining set of strategies is denoted by L_i^1. Any $s_i \in L_i^1$ is called a $\boldsymbol{L_i^1}$ **(level-1) strategy** for i.

⋮

Step k. For each $i \in \{d,p\}$, eliminate from L_i^{k-1} the strategies s_i for which there exists a subgame γ of Γ such that the restriction $s_i|_\gamma$ of s_i to γ is not a best response to $s_{-i}|_\gamma$ for any $s_{-i} \in L_{-i}^{k-1}$.

The remaining set of strategies is denoted L_i^k. Any $s_i \in L_i^k$ is called a $\boldsymbol{L_i^k}$ **(level-k) strategy** for i.

Observe that the sets of level-k strategies are nested ($L_i^0 \supseteq L_i^1 \supseteq \dots$). Observe also that, for every procedure Γ that satisfies finiteness, the set of level-k strategies is non-empty for every k.[20]

The argument at the beginning of this section suggests using the following concept of a dominance threshold as a measure of strategic complexity.

Definition 3.2 (Dominance threshold)**.** For any procedure Γ and any profile (R_d, R_p), the *dominance threshold* is the smallest integer r^* such that, for each $i \in \{d,p\}$, there exists an $L_i^{r^*}$ strategy s_i^* that is a best response to *every* $L_{-i}^{r^*-1}$ strategy.

If there exists no such integer, then the dominance threshold of Γ is ∞, i.e., the procedure cannot be solved by iterated elimination of never best responses. Note that if the dominance threshold r^* is finite, then there exists a strategy profile $(s_i, s_{-i}) \in L_i^{r^*} \times L_{-i}^{r^*}$ that is a subgame perfect equilibrium.

[20] When Γ satisfies finiteness, for any $i \in \{d,p\}$ and any strategy s_{-i}, the set $\{\Gamma(s_i,s_{-i}) \mid s_i \in \mathscr{S}_i\}$ is finite because $\mathscr{S}_i$ is finite. Hence, there must exist a strategy $t_i(s_{-i})$ such that $\Gamma\big(t_i(s_{-i}),s_{-i}\big)\ R_i\ \Gamma\big(s_i',s_{-i}\big)$ for all $s_i' \in \mathscr{S}_i$ and L_i^1 is non-empty. The non-emptiness of L_i^k then follows by induction.

Throughout this paper, the parties' knowledge of each others' preferences and levels of rationality is left unspecified. The idea behind the dominance threshold is precisely to measure the "amount" of common knowledge needed for the parties to have dominant strategies. For example, when the dominance threshold of a game is 1, the parties have a dominant strategy regardless of their knowledge of their opponent's preference and level of rationality (the parties only need to know the structure of the game). When the dominance threshold is 2, the parties only need to know their opponent's preference and the fact that their opponent is a best responder in order to have a dominant strategy.

3.5 One-shot procedures

In this section, I show that one-shot$_Q$ is strategically simpler than one-shot$_M$ in the following sense.

Proposition 3.3. *(i) For* every *problem, the dominance threshold of one-shot$_M$ is* no smaller than *the dominance threshold of one-shot$_Q$. (ii) For* some *problems, the dominance threshold of one-shot$_M$ is* larger than *the dominance threshold of one-shot$_Q$.*

In the rest of this section, I prove and illustrate Proposition 3.3.

3.5.1 One-shot$_Q$ is always maximally simple

The next example illustrates how to compute the dominance threshold of one-shot$_Q$ for a particular problem.

Example 3.4. This example is illustrated in Figure 3.2. In the figure, L_T for $T \subseteq N$ represents a lottery in which one juror is drawn at random from T. The labels on the branches of the tree indicate the juror who is challenged in the corresponding action.

Suppose that $c_d = c_p = b = 1$ and d is the first party to challenge. Suppose also that the parties have aligned preferences 1 R_d 2 R_d 3 and 1 R_p 2 R_p 3.

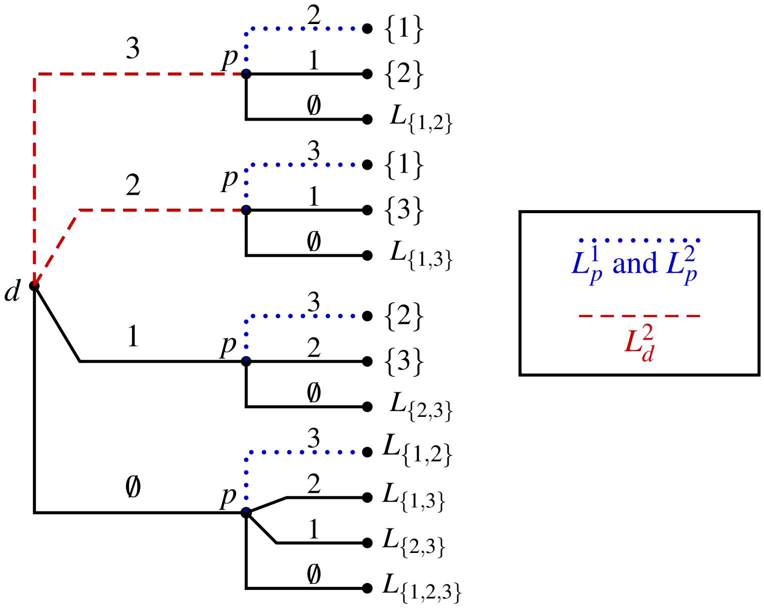


Figure 3.2: Computing the dominance threshold in Example 3.4.

Because preferences for jurors are strict, p has a unique dominant strategy s_p^* which consists in (a) challenging juror 3 if d did *not* challenge 3 and (b) challenging juror 2 if d did challenge juror 3 (dotted blue branches in the figure). Strategy s_p^* is, therefore, the unique L_p^1 strategy. It directly follows from uniqueness that s_p^* is a best response to all L_d^1 strategies.

Because there is a unique L_p^1 strategy s_p^*, any L_d^2 strategy that best responds to s_p^* (either of the red dashed branches in the figure) is a best response to *all* L_p^1 strategies. Hence, the dominance threshold of one-shot$_Q$ is at most 2 for this problem. But by Proposition 3.1, because one-shot$_M$ is an N-struck procedure, the dominance threshold of one-shot$_Q$ is at least 2 for every problem. Thus, the dominance threshold of one-shot$_Q$ is 2 for this problem.

It is not hard to see how the argument in Example 3.4 generalizes to any problem. In general, the party $-i$ who challenges second in one-shot$_Q$ has a unique dominant strategy s_{-i}^*. Then, any best response by i to s_{-i}^* is a best response to every L_{-i}^1 strategy.

Proposition 3.4. *For any problem, the dominance threshold of one-shot$_Q$ is 2.*

Proposition 3.4 does not depend on the separability assumption. Instead, the proof relies on the fact that preferences for the outcomes of the procedure are strict. The proposition also extends to situations in which complete information (which is implicit in the definition of a dominance threshold of 2) is relaxed. Consider Example 3.4. In order to have a dominant strategy, d only needs to know that p will challenge juror 3 if she challenges juror 2. Hence, d only needs to know which juror is p's *worst juror* in order to have a dominant strategy (as opposed to knowing *all* of p's preference for jurors).

By Proposition 3.1, because one-shot$_M$ is an N-struck procedure, the dominance threshold of one-shot$_M$ is at least 2 for every problem. Together with Proposition 3.4, this implies that the dominance threshold of one-shot$_M$ is never smaller than the dominance threshold of one-shot$_Q$, which proves Proposition 3.3(i).

3.5.2 One-shot$_M$ is often complex: one-common profiles

I now show that one-shot$_M$ is more complex than one-shot$_Q$ when the profile is not juror inverse and preferences for jurors satisfy some "commonality at the bottom".

3.5.2.1 Motivating example

Example 3.5. This example is illustrated in Figure 3.3. Suppose that $b = c_d = c_p = 1$. Also suppose that the parties' preferences are $1\ R_d\ 2\ R_d\ 3$ and $2\ R_p\ 1\ R_p\ 3$.

Both challenging juror 3 and challenging juror 1 are L_p^0 strategies.[21] Challenging juror 2 is d's best response to p challenging juror 3, and challenging juror 3 is d's best response to p challenging juror 1. Hence, both challenging juror 2 and challenging juror 3 are L_d^1 strategies. But no strategy of p is a best response to *both* of these L_d^1 strategies. Therefore, the dominance threshold of one-shot$_M$ is at least 3 for this problem.

In Example 3.5, both parties agree that juror 3 is the worst juror. Therefore, any best

[21] Challenging juror 2 and challenging no juror are also L_p^0 strategies. However, it is sufficient to consider the other two L_p^0 strategies.

responding party would challenge juror 3 if her opponent did not. But each party also prefers a situation in which her *opponent* challenges 3, and she challenges her second worst juror. That is, each party would like to make a credible threat *not* to challenge juror 3 and free ride on her opponent's challenge of juror 3. But because the procedure is simultaneous, such a credible threat is impossible. As explained in detail in Example 3.5, the impossibility for the parties to commit to leaving juror 3 unchallenged makes the dominance threshold of one-shot$_M$ larger than 2 for this problem. Together with Proposition 3.4, Example 3.5 therefore proves Proposition 3.3(ii).

In fact, the dominance threshold in Example 3.5 is ∞, which shows just how complex one-shot$_M$ can become when the profile is not juror inverse.

Example 3.5 (Continued)**.** Party p's best responses to these two L_d^1 strategies are to challenge juror 3 (p's best response to d challenging juror 2) and to challenge juror 1 (p's best response to d challenging juror 3), see Figure 3.3. Thus, both challenging juror 3 and challenging juror 1 are L_p^2 strategies. But these two L_p^2 strategies are the two L_p^0 strategies considered at the beginning of the example. The argument therefore extends by induction, which shows that the dominance threshold of one-shot$_M$ is ∞ for this problem.

As this example illustrates, for some profiles that are not juror inverse, the parties' common knowledge of each others' rationality and preferences is not sufficient to provide the parties with dominant strategies and make the game strategically simple. Even for "high levels of common knowledge", the game induced by one-shot$_M$ remains akin to a *game of chicken* in which each party prefers to swerve (i.e., challenge some of her worst jurors) if her opponent stays straight (i.e., does *not* challenge some of her worst jurors), but prefers to stay straight if her opponent swerves.

3.5.2.2 One-common profiles

Profiles for which the dominance threshold of one-shot$_M$ is ∞ are not rare. Given c_d and c_p, a profile is **one-common** if a juror w that is among the c_d worst jurors of d is also

	L_p^0 strategies	L_d^1 strategies	L_p^2 strategies	L_d^3 strategies	...
Juror	3	2	3	2	...
challenged	1	3	1	3	...

Figure 3.3: Iterated best responses based on two of the L_p^0 strategies in the problem of Example 3.5.

among the c_p worst jurors of p. Intuitively, the dominance threshold of one-shot$_M$ is ∞ for one-common profiles because the free rider problem described in Example 3.5 extends to one-common profiles. When the profile is one-common, each party would like to make a credible threat *not* to challenge juror w and free ride on her opponent's challenge of juror w. But if her opponent does not challenge w, each party prefers to challenge w herself than to leave w unchallenged.

Proposition 3.5. *If the profile is one-common, then the dominance threshold of one-shot$_M$ is ∞.*

Although Proposition 3.5 relies more directly than Proposition 3.4 on the separability assumption,[22] the intuition behind Proposition 3.5 applies even when separability is relaxed. Regardless of the assumptions on preferences, if for some juror w, both parties have best responses that include challenging w, then the dominance threshold is larger than 2 in one-shot$_M$. The proposition also extends to situations of incomplete information in which the parties only know that they have a common juror w at the bottom of their ranking of jurors (but do not know each other's complete preferences for jurors).

One-common profiles arise in a number of natural jury selection situations. For example, both parties may dislike a juror who they view as too unpredictable. Both parties may also dislike "devil advocates" or "irresolute" jurors who are likely to induce a hung jury and to force a retrial of the case. Finally, d may dislike juror j's position on some charges, while p dislike juror j's position on different charges.

[22] One-common profiles are not well-defined without the separability assumption.

In the Appendix, I show that one-common profiles were frequent in the selection of an arbitrator between unions and employers by the New Jersey Public Employment Relations Commission from 1985 to 1996 (Bloom and Cavanagh, 1986; de Clippel et al., 2014). I also show that one-common profiles represent a significant proportion of the set of profiles. This is true even when attention is limited to profiles that are close to being juror inverse (in a sense that is made precise in the Appendix).

In the Appendix, the proportion of one-common profiles is shown to be an increasing function of the number of challenges and a decreasing function of the number of jurors b. Based on the objective of reducing strategic complexity, Proposition 3.5 and the results in the Appendix therefore provide a game-theoretic justification for decreasing the number of peremptory challenges, a measure that has some support among those who defend a reform of the peremptory challenge (Henley, 1996). Procedures in which the number of challenges is high relative to b exist in practice. In the United States, the number of challenges tends to increase with the gravity of the charges. For example, in federal cases for which the death penalty is sought by the prosecution, $b = 12$ and $c_d = c_p = 20$. In this case, the dominance threshold is ∞ for more than 97% of the profiles (and more than 15% of the profiles that are close to being juror inverse).

Overall, the results in this section contrast with the judges' beliefs that "blind" (i.e., simultaneous) procedures leave less room for the parties to strategize than sequential ones.[23] Contrary to the judges' beliefs, the dominance thresholds suggest that one-shot$_Q$ is strategically simpler than one-shot$_M$: By making past actions observable, one-shot$_Q$ allows the parties to make credible threats about the jurors they challenge, which reduces the amount of guesswork involved in determining an appropriate strategy. The next section shows that similar results hold for other N-struck procedures.

[23] See the last quotation from Shapard and Johnson (1994) in the Introduction.

3.6 Alternating and other N-struck procedures

In general, it is unclear how alternating$_M$ and alternating$_Q$ compare. However, extending the logic of Proposition 3.5, it is possible to obtain a partial comparison for a significant subset of profiles. For this subset of profiles, the dominance threshold of any *simultaneous* N-struck procedure (including alternating$_M$) is infinite, whereas the dominance threshold of any *sequential* N-struck procedure (including alternating$_Q$) is finite.

If preferences for the outcomes of a sequential N-struck procedure are strict (including preferences on lotteries), then the procedure always has a finite dominance threshold. This follows from the fact that, with no indifferences on outcomes, sequential N-struck procedures induce games of perfect information that can be *uniquely* solved by backward induction.[24] Then, the number of rounds of backward induction required to solve the game is an upper bound for the dominance threshold.

Proposition 3.6. *For any sequential N-struck procedures, if preferences for the outcomes of the procedure are strict, then the dominance threshold is finite and smaller than the depth of the game tree.*[25]

Again, Proposition 3.6 does not depend on the separability assumption, but instead on the assumption that preferences for the outcomes of the procedure are strict.

Recall that one-shot$_M$ has an infinite dominance threshold when the profile is one-common because each party would like to free ride on her opponent's challenge of one of the jurors they both dislike (see Example 3.6). This idea generalizes to the class of simultaneous N-struck procedures as a whole. Below, I identify for each simultaneous N-struck procedure Γ a set of Γ-one-common profiles. In Proposition 3.7, I show that any Γ-one-common profile induces an infinite dominance threshold in Γ.

Informally, given a simultaneous N-struck procedure Γ, a profile is Γ-*one-common* if in

[24] More precisely, multiple strategy profiles can survive backward induction, but each of these profiles must yield the same outcome.

[25] The depth of a game tree is the length of the longest path from the initial node to a terminal node.

one of the final subgames of Γ, the set of jurors that remain unchallenged gives rise to the free rider problem described above. Formally, given Γ, a profile is **Γ-one-common** if there exists a subgame γ of Γ such that (a) both parties can still challenge jurors in γ (i.e., $l_i^\gamma \geq 1$ for both $i \in \{d,p\}$), (b) the first round of γ is the final round of γ in which both parties can challenge jurors,[26] and (c) among the unchallenged jurors, one of the l_d^γ worst jurors according to R_d is also one of the l_p^γ worst jurors according to R_p.

Example 3.6. Consider alternating$_M$ and any problem in which $c_d = c_p = 2$, $b = 1$, and the preferences for jurors are

$$\begin{array}{cccccc} R_d: & 1 & 2 & 3 & 4 & 5 \\ R_p: & 2 & 4 & 5 & 1 & 3 \end{array} \tag{3.5}$$

The profile is not one-common because $\{4,5\} \cap \{1,3\} = \emptyset$.

However, consider the subgame γ^* that follows from d challenging juror 4 and p challenging juror 5 in the first round. Subgame γ^* satisfies (a) and (b) in the definition of an alternating$_M$-one-common profile. Also, both players have the same worst juror among $\{1,2,3\}$, the set of unchallenged jurors at the beginning of γ^*. Hence, condition (c) in the definition of an alternating$_M$-one-common profile is also satisfied and so profile (3.5) is alternating$_M$-one-common.

To see why the dominance threshold is infinite in subgame γ^*, observe that in γ^*, each party wants to free-ride on her opponent's challenge of juror 3. This induces an infinite dominance threshold for the same reasons that the dominance threshold is infinite in Example 3.5.

Consider one-shot$_M$-one-common profiles. The only subgame of one-shot$_M$ is one-shot$_M$ itself. Hence, in the case of one-shot$_M$, (a), (b), and (c) boil down to requiring that among N, one of the c_d worst jurors according to R_d is also one of the c_p worst jurors

[26] This could arise because the first round of γ is the terminal round of Γ or because both parties only have one challenge left in γ.

according to R_p, which is the definition of a one-common profile. Because the sets of one-shot$_M$-one-common and one-common profiles are identical, the next proposition generalizes Proposition 3.5.

Proposition 3.7. *For any simultaneous N-struck procedure Γ, if the profile is Γ-one-common, then the dominance threshold of Γ is ∞.*

Example 3.6 (Continued)**.** To see why the dominance threshold is infinite in subgame γ^*, observe that in γ^*, each party wants to free-ride on her opponent's challenge of juror 3. This induces an infinite dominance threshold for the same reasons that the dominance threshold is infinite in Example 3.5.

Propositions 3.6 and 3.7 jointly imply that, whenever the profile is alternating$_M$-one-common (and preferences on outcomes are strict), the dominance threshold of alternating$_Q$ is smaller than the dominance threshold of alternating$_M$. That is, Proposition 3.3 partially extends to alternating procedures. In the Appendix, I show that the alternating$_M$-one-common profiles are a strict superset of the one-common profiles. Hence, the arguments on the prevalence of one-common profiles in Section 3.5.2 extend to alternating$_M$-one-common profiles.

3.7 Conclusion

This paper shows how jury selection procedures can be compared in terms of their strategic complexity by computing their *dominance thresholds*, i.e., the number of rounds of elimination of strategies that are never responses required for the parties to have a dominant strategy. The results in this paper notably show that procedures in which challenges are made sequentially tend to be strategically simpler than procedures in which challenges are simultaneous.

The dominance threshold offers a new method to compare the strategic complexity of mechanisms. Unlike previous methods in the literature (Pathak and Sönmez, 2013; de

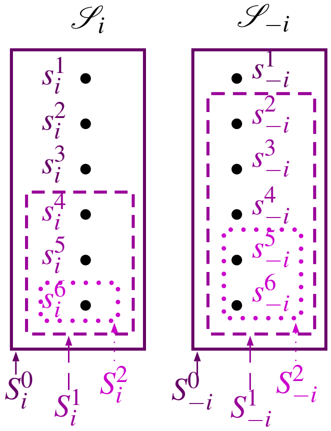


Figure 3.4: Arbitrary hierarchical model for a given preference profile (R_d, R_p).

Clippel et al., 2014; Arribillaga and Massó, 2015), it allows for comparisons even when the mechanisms at stake are indirect or induce games of imperfect information.

More generally, the dominance threshold shows how hierarchical models can be used to compare the strategic complexity of mechanisms. As illustrated in Figure 3.4, for any profile (R_d, R_p), a **hierarchical model** specifies a pair $(\{S_d^0, \dots, S_d^m\}, \{S_p^0, \dots, S_p^m\})$ of collections of nested strategy sets, i.e., $\mathscr{S}_i \subseteq S_i^0 \subseteq \cdots \subseteq S_i^m$ (m could be infinite). As k increases, the sets S_i^k represent increasingly restrictive models of the strategies that i could potentially play.

This paper studies the *level-k* hierarchical model $(\{L_d^0, \dots, L_d^m\}, \{L_p^0, \dots, L_p^m\})$ defined in Section 3.4. Given a profile (R_d, R_p), I define the dominance threshold as the smallest hierarchical level r^* for which each party i has a strategy $s_i^* \in \mathscr{S}_i$ that is a best response to *every* strategy in $L_{-i}^{r^*-1}$. I then use the dominance threshold as a measure of strategy complexity.

Clearly, this logic is not specific to the *level-k* hierarchical model. A natural alternative would be to use the "*undominated*" hierarchical model UD defined by the process of iterated elimination of dominated strategies. One could then define an alternative UD-

dominance threshold,[27] and perform an analysis similar to those of this paper.[28]

In general, there is no logical relation between the UD-dominance threshold and the level-k-dominance threshold of a game. However, some of the results in this paper also apply when the UD-dominance threshold is used instead.

First, it is not hard to see that the UD-dominance threshold of one-shot$_Q$ is 2. Second, it can be shown that for every problem, the UD-dominance threshold of one-shot$_M$ is at least as large as the level-k-dominance threshold of one-shot$_M$.[29] Hence, the results in Section 3.5.2 also apply using the UD-dominance threshold. Specifically, the UD-dominance threshold of one-shot$_M$ is larger than 2 for a significant set of problems. Whether the results of Section 3.6 also extend to the UD-dominance threshold case is left as an open question.

Appendix

Proofs

Proof of Proposition 3.1. Consider any problem. (i). Let $i \in \{d, p\}$ be the party who challenges jurors first and let w be i's worst juror. Let s^w_{-i} be the strategy in which $-i$ challenges only w provided i did not already challenge w, and $-i$ challenges no other juror otherwise. Also, let s^0_{-i} be the strategy in which $-i$ does not challenge any jurors. Because i's preference for jurors is strict, i's unique best response $t_i(s^w_{-i})$ consists in challenging her c_i worst jurors among $N \backslash \{w\}$. In contrast, party i's unique best response $t_i(s^0_{-i})$ consists in

[27] That is, the smallest hierarchical level r^*_{UD} for which each party i has a strategy $s^*_i \in \mathscr{S}_i$ that is a best response to *every* strategy of her opponent that survives $r^*_{UD} - 1$ rounds of iterated elimination of dominated strategies.

[28] Another option is to use a hierarchical model in which the level of sophistication of the parties is fixed, say to level-1, but the parties' *information* about each others' preferences is refined in each iteration of the model. The corresponding I-dominance threshold could, for example, be the smallest r^*_I such that each party i has a strategy $s^*_i \in \mathscr{S}_i$ that is a best response to *every* L^1_{-i} strategy for *any* preference $\widetilde{R}_{-i}$ that has the same r^*_I worst jurors as R_{-i}.

Observe that $r^*_I = 1$ in one-shot$_Q$ for the problem described in Example 3.4.

[29] By the assumption that preferences for jurors are strict, best responses are unique in one-shot$_M$. Thus, every best response is also an undominated strategy, and the set of strategies that survive k rounds of iterated elimination of never best responses is a *subset* of the set of strategies that survive k rounds of iterated undominated strategies.

i challenging her c_i worst jurors among N, which includes challenging juror w. Thus, no strategy of i is a best response to both s^w_{-i} and s^0_{-i}, so i does not have a dominant strategy.

(ii). For one-shot$_M$, the proof is a straightforward adaptation of the argument in (i).[30] Now consider any N-struck procedure Γ that is not a one-shot procedure. By definition of the N-struck class, the one-shot procedures are *essentially* the only N-struck procedure with a single round, i.e., any other N-struck procedures with $f = 1$ must have $x^1_i > c_i$ for both $i \in \{d, p\}$ and is therefore strategically equivalent to a one-shot procedure.

Thus, let Γ have at least two rounds. Let $i \in \{d, p\}$ be either of the parties (not necessarily the first one to challenge if the procedure is sequential). Consider the strategy $\tilde{s}^w_{-i}$ in which (A) $-i$ challenges no juror in the first round and (B) in each of the following rounds, (a) $-i$ challenges as many as possible of i's worst jurors among the jurors that remain unchallenged *provided* i challenged *at least* one juror in the first round and (b) $-i$ never challenges any juror otherwise. Also consider the strategy $\tilde{s}^0_{-i}$ in which $-i$ never challenges any juror.

In any best response $t_i(\tilde{s}^w_{-i})$, (A$'$) i challenges a juror in the first round of Γ, (B$'$) i never challenges any of her c_{-i} worst jurors, and (C$'$) i challenges her c_i other worst jurors over the course of the procedure. In contrast, in any best response $t_i(\tilde{s}^0_{-i})$, (A$''$) i challenges her c_i worst jurors among N over the course of the procedure (and only those jurors by definition of an N-struck procedure), which includes challenging some of her c_{-i} worst jurors.

In particular, the set of jurors that i challenges in the first round of Γ is different under $t_i(\tilde{s}^w_{-i})$ and under $t_i(\tilde{s}^0_{-i})$.[31] Indeed, by (A$'$) i challenges at least one juror j in the first round of Γ and by (B$'$) this juror cannot be one of her c_{-i} jurors. Thus, by (A$''$), j can never be challenged as part of $t_i(\tilde{s}^0_{-i})$. Thus, no strategy of i is a best response to both $\tilde{s}^w_{-i}$ and $\tilde{s}^0_{-i}$, so i does not have a dominant strategy.

Proof of Proposition 3.2. Let $\mathscr{D}$ be the set of separable preferences on $\Delta\mathscr{J}$. In order to

[30] By the symmetry of one-shot$_M$, the argument for party i in (i) applies to both parties in one-shot$_M$.

[31] If i is the second juror to challenge, the set of jurors that i challenges in the first round of Γ *following $-i$ challenging no juror in the first round* is different under $t_i(\tilde{s}^w_{-i})$ and under $t_i(\tilde{s}^0_{-i})$.

derive a contradiction, assume that there exists a challenge procedure Γ satisfying finiteness and minimal challenge such that, for every profile $(R_d, R_p) \in \mathscr{D} \times \mathscr{D}$, both parties have a dominant strategy.

Let $s_i^*(R_i)$ be one of i's dominant strategies when i's preference are R_i. Consider the direct mechanism $M^\Gamma : \mathscr{D} \times \mathscr{D} \to \Delta \mathscr{J}$ constructed from Γ by setting

$$M(R_d, R_p) \coloneqq \Gamma(s_d^*(R_d), s_p^*(R_p)) \qquad \text{for all } (R_d, R_p) \in \mathscr{D} \times \mathscr{D}, \tag{3.6}$$

where $\Gamma(s_d^*(R_d), s^*(R_p))$ is the lottery that obtains when $(s_d^*(R_d), s^*(R_p))$ is played in Γ.

Because $s_i^*(R_i)$ is a dominant strategy given R_i, for all $i \in \{d, p\}$ and all $R_i \in \mathscr{D}$,

$$\Gamma(s_i^*(R_i), s_{-i}^*(R_{-i}))\ R_i\ \Gamma(s_i^*(R_i'), s_{-i}^*(R_{-i})) \qquad \text{for all } R_i' \in \mathscr{D} \text{ and } R_{-i} \in \mathscr{D}.$$

But then, by construction of M, for all $i \in \{d, p\}$ and all $R_i \in \mathscr{D}$,

$$M(R_i, R_{-i})\ R_i\ M(R_i', R_{-i}) \qquad \text{for all } R_i' \in \mathscr{D} \text{ and } R_{-i} \in \mathscr{D}.$$

That is, M is strategy-proof on the domain of separable profiles.

Notice that the domain of additive profiles $\mathscr{D}^{add} \times \mathscr{D}^{add}$ is a subset of the domain of separable profiles.[32] By Van der Linden (2015, Example 3), any domain of profiles that contains the domain of additive profiles is a *negative leximin* domain (see Van der Linden, 2015, Domain Property 3 for a definition of negative leximin domains). But then, M contradicts Van der Linden (2015, Corollary 3) which states that, on a negative leximin domain, no mechanism constructed from a procedure satisfying finiteness and minimal challenge as in (3.6) is strategy-proof, the desired result.

The above proof shows that Proposition 3.2 is, in fact, true on any *negative leximin* domain of profiles, which includes the domain of additive profiles.

[32] Preference R_i is additive if there exists a function $u_i \colon N \to \mathbb{R}$ such that, for all $L, L' \in \mathscr{J}$, $L\ R_i\ L' \Leftrightarrow \sum_{J \in \mathscr{J}} L(J) \sum_{t \in J} u_i(t) \geq \sum_{J \in \mathscr{J}} L'(J) \sum_{t \in J} u_i(t)$.

	L_i^0 strategies	L_{-i}^1 strategies	L_i^2 strategies	L_{-i}^3 strategies
Set of jurors challenged	$t_i(N)$ $\ni w$	$t_{-i}(t_i(N))$ $\not\ni w$	$t_i(t_{-i}(t_i(N)))$ $\ni w$	$\dots$
	$t_i(N \backslash \{w\})$ $\not\ni w$	$t_{-i}(t_i(N \backslash \{w\}))$ $\ni w$	$t_i(t_{-i}(t_i(N \backslash \{w\})))$ $\not\ni w$	$\dots$

Figure 3.5: Iterated best responses based on two of the L_p^0 strategies.

Proof of Proposition 3.4. Consider any problem. Let i be the first party to challenge. For $-i$, the unique best response s_{-i}^* to *any* strategy by i is to challenge her c_{-i} worst jurors among the jurors that i did not challenge (uniqueness follows from preferences for jurors being strict). This strategy is, therefore, the unique L_{-i}^1. It directly follows from uniqueness that s_{-i}^* is a best response to all L_i^0 strategies.

Because there is a unique L_{-i}^1 strategy s_{-i}^*, any L_i^2 strategy that best responds to s_{-i}^* is a best response to *all* L_{-i}^1 strategies. Hence, the dominance threshold of one-shot$_Q$ is at most 2 for this problem. But by Proposition 3.1, because one-shot$_M$ is an N-struck procedure, the dominance threshold of one-shot$_Q$ is at least 2 for every problem. Thus, the dominance threshold of one-shot$_Q$ is 2 for this problem.

Proof of Proposition 3.5. The proof generalizes Example 3.5 and is illustrated in Figure 3.5. Consider any problem in which the profile is one-common. Let $w \in N$ be a juror that is among the c_d worst jurors of d and among the c_p worst jurors of p. For any set $\widetilde{N} \subseteq N$ containing at least c_i jurors, let $t_i(\widetilde{N})$ be the set of the c_i worst jurors in $\widetilde{N}$ according to R_i.

Induction basis. For each $i \in \{d,p\}$, both challenging jurors $t_i(N)$ and challenging jurors $t_i(N \backslash \{w\})$ are L_i^0 strategies of one-shot$_M$. Observe that $w \in t_i(N)$ and $w \notin t_i(N \backslash \{w\})$.

Induction step. For any $k \in \mathbb{N}$, suppose that for both $i \in \{d,p\}$, the set of L_i^k strategies of one-shot$_M$ contains a strategy s_i^k in which w is challenged and a strategy $\tilde{s}_i^k$ in which w is *not* challenged. Then, for each $i \in \{d,p\}$, we have $w \notin t_i(s_{-i}^k)$ and $w \in t_i(\tilde{s}_{-i}^k)$. Hence, for each $i \in \{d,p\}$, the set of L_i^{k+1} strategies of one-shot$_M$ contains a strategy $\tilde{s}_i^{k+1} := t_i(\tilde{s}_{-i}^k)$ in which w is challenged and a strategy $s_i^{k+1} := t_i(s_{-i}^k)$ in which w is *not* challenged.

By the induction step, strategies s_i^k and $\tilde{s}_i^k$ are well-defined for both $i \in \{d,p\}$ and for all $k \in \mathbb{N}$. But observe that, for every $k \in \mathbb{N}$, $w \in t_i(s_{-i}^k)$ and $w \notin t_i(\tilde{s}_{-i}^k)$ which implies that $t_i(s_{-i}^k) \neq t_i(\tilde{s}_{-i}^k)$. Because $t_i(s_{-i}^k)$ and $t_i(\tilde{s}_{-i}^k)$ are unique best responses, no strategy of i is a best response to every L_{-i}^k strategy of one-shot$_M$. Hence, the dominance threshold is at least $k+1$. Because this is true for all $k \in \mathbb{N}$, this concludes the proof.

Proof of Proposition 3.6. Let Γ be any sequential N-struck procedure. Let $\bar{e}$ be the depth of the game tree associated with Γ. It is convenient to describe subgames of Γ in terms of the height of their initial node, where the *height* of a node is the length of the longest path from that node to a terminal node. A subgame of Γ the root node of which has height h is denoted γ^h.

Induction basis. Consider any subgame γ^1 of Γ. Because preferences on outcomes are strict, (a) the outcome of γ^1 is the same under any level-1 profile, and (b) the dominance threshold of γ^1 is 1.

Induction step. For any $h \in \{1,\dots,\bar{e}\}$, suppose that for any subgame γ^{h-1} of Γ with height $h-1$, (a$'$) the outcome of γ^{h-1} is the same under any level-$(h-1)$ profile and (b$'$) the dominance threshold of γ^{h-1} is no larger than $h-1$.

Consider any subgame γ^h of Γ with height h. Let i be the party who moves at the root node of γ^h. By nestedness, any level-h profile is a level-$(h-1)$ profile. Hence, by (a$'$), for any subgame γ^{h-1} directly following γ^h, the outcome of γ^{h-1} is the same under any level-h profile. Thus, if the outcomes of γ^h differ under two level-h profiles, it must be that i's action at the root node of γ^h lead to two subgames $\widehat{\gamma}^{h-1}$ and $\widetilde{\gamma}^{h-1}$ with different level-$(h-1)$ outcomes. But then, because preferences on outcomes are strict, i's action at the root node of γ^h cannot be part of a level-h strategy. Therefore, (a$''$) the outcome of γ^h must be the same under any level-h profile.

It follows from (a$''$) that for each $i \in \{d,p\}$, any best response to an L_{-i}^{h-1} strategy is a best response to all L_{-i}^{h-1} strategies. Hence, (b$''$) the dominance threshold of γ^h is no larger than h.

By induction, (b$''$) is true for all $h \in \{1,\dots,\bar{e}\}$. Hence, (b$''$) is true for $\gamma^{\bar{e}} = \Gamma$, the desired result.

Proof of Proposition 3.7. For any simultaneous N-struck procedure Γ, consider the subgame γ described in the definition of a Γ-one-common profile. Let $w \in N$ be (one of) the unchallenged juror(s) in γ that is among the l_d^γ worst jurors according to R_d and among the l_p^γ worst jurors according to R_p. The proof is then a straightforward adaptation of the proof of Proposition 3.5 with c_i replaced by l_i^γ and the challenges described in the proof of Proposition 3.5 occurring in the first round of γ.

Prevalence of one-common profiles

Proportion of one-common profiles

As shown in Figure 3.6(a), when $c_d = c_p = c$, the proportion of one-common profiles relative to the set of all profiles is high.[33] (*Proportions of profiles* refer to proportions of profiles of preferences *for jurors*.) Even among profiles that are close to being juror inverse, the proportion of one-common profiles can be significant when c is high relative to b. Figure 3.6(b) shows the proportion of one-common profiles among *almost juror inverse* profiles. A profile is **almost juror inverse** if it can be constructed from a juror inverse profiles by changing the ranking of a single juror in the preference of one of the parties.[34] This definition is illustrated in Example 3.7.

Example 3.7. Suppose that $c_d = c_p = 4$ and $b = 1$. The following profile is almost juror inverse because it is constructed from a juror inverse profile by changing the ranking of a

[33] Alternatively, the proportions are lower bounds for $c = \min\{c_d, c_p\}$.

[34] Formally, profile (R_d^*, R_p^*) is *almost juror inverse* if there exists a juror inverse profile (R_d, R_p), a party $i \in \{d, p\}$ and a juror $j^* \in N$ such that (a) for all $j,k \in N$, $j\ R_{-i}^*\ k$ if and only if $j\ R_{-i}\ k$, (b) for all $j,k \in N$ with $j,k \neq j^*$, $j\ R_i^*\ k$ if and only if $j\ R_i\ k$, and (c) for some $j \in N$ with $j \neq j^*$, $j\ R_i^*\ j^*$ and $j^*\ R_i\ j$ or $j^*\ R_i^*\ j$ and $j\ R_i\ j^*$.

single juror — namely juror 7 — in the preference of p.

$$\begin{array}{lccccccccc} R_d: & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ R_p: & 9 & 8 & 6 & 5 & 4 & 3 & 2 & 7 & 1 \end{array} \tag{3.7}$$

The above profile is also one-common because 7 is among the four worst jurors of both d and p, and $c_d = c_p = 4$.

Computing the proportions in Figure 3.6. Because the one-common property is preserved under relabeling of the jurors, let us, without loss of generality, suppose that

$$(b+2c)\ R_{-i}\ (b+2c-1)\ R_{-i}\ \ldots\ R_{-i}\ c\ R_{-i}\ \ldots\ R_{-i}\ 2\ R_{-i}\ 1.$$

Also, for any R_i, let the c worst jurors according to R_i be $w_1^{R_i}, w_2^{R_i}, \ldots, w_c^{R_i}$, with $w_1^{R_i}\ R_i\ \ldots\ R_i\ w_c^{R_i}$.

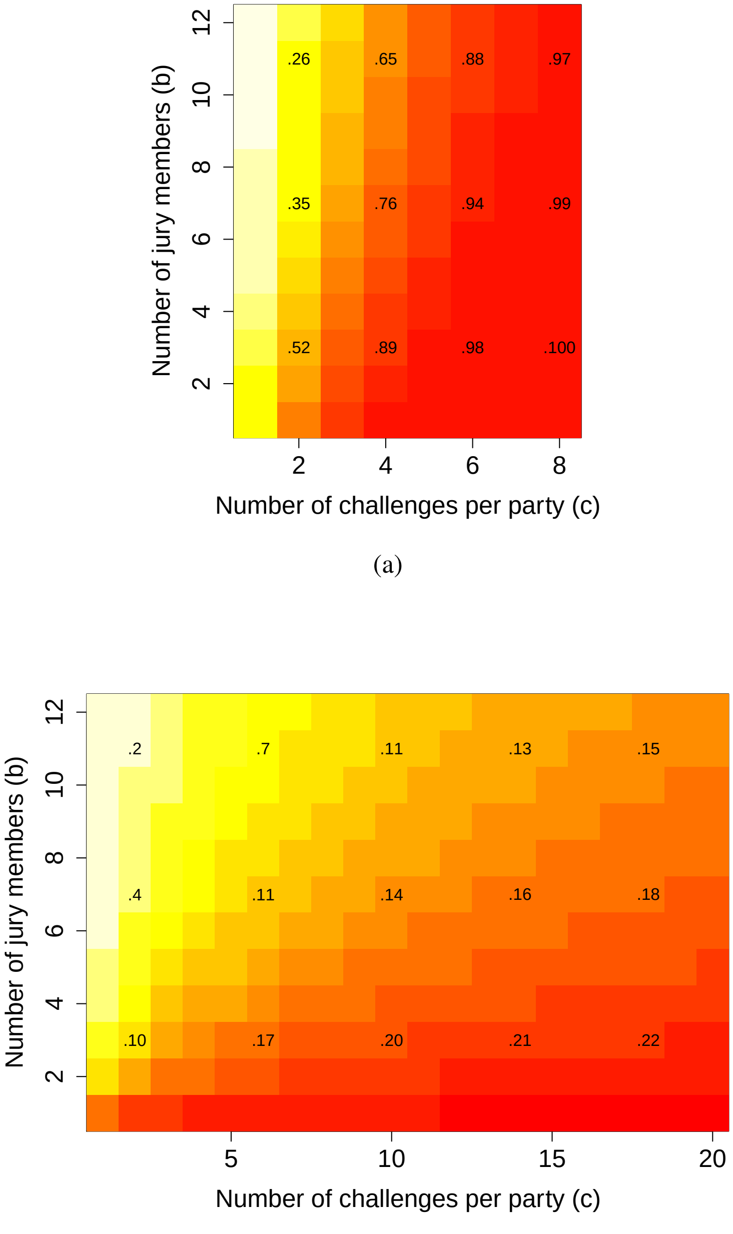
Figure 3.6(a). The proportion of one-common profiles is equal to the proportion of preferences R_i that induce a one-common profile given the above arbitrary choice of R_{-i}. Clearly, this proportion can be treated as a probability, where the set of outcomes is the set of preferences R_i, and the probability of drawing any particular R_i is $\frac{1}{(b+2c)!}$.

We are interested in the probability that one of the c worst jurors of i is among the c worst jurors of $-i$, i.e.,

$$\mathbb{P}\big(w_c^{R_i} \in \{1,\ldots,c\} \cup \cdots \cup w_1^{R_i} \in \{1,\ldots,c\}\big),$$

This probability is equal to

$$\begin{aligned}
&\mathbb{P}\Big(w_c^{R_i} \in \{1,\ldots,c\}\Big) + \\
&\mathbb{P}\Big(w_{c-1}^{R_i} \in \{1,\ldots,c\} \cap \big(w_c^{R_i} \notin \{1,\ldots,c\}\big)\Big) + \\
&\vdots \\
&\mathbb{P}\Big(w_1^{R_i} \in \{1,\ldots,c\} \cap \big(w_c^{R_i} \notin \{1,\ldots,c\} \cap \cdots \cap w_2^{R_i} \notin \{1,\ldots,c\}\big)\Big).
\end{aligned} \tag{3.8}$$



(a)

(b)

Figure 3.6: Proportion of one-common profiles relative to the set of (a) all profiles and (b) almost juror inverse profiles.

The sum in (3.8) can be computed recursively (the recursion has been implemented in `R`). For each $r \in \{1,\dots,c\}$, let σ^r be the sum of the r terms in the first r lines of (3.8) (i.e., the sum of the terms involving $w_c^{R_i} \in \{1,\dots,c\}$ to $w_{c-(r-1)}^{R_i} \in \{1,\dots,c\}$). In step $r=1$, the recursion is initiated by computing σ^1. To do so, observe that

$$\sigma^1 := \mathbb{P}\big(w_c^{R_i} \in \{1,\dots,c\}\big) = \mathbb{P}\big(w_c^{R_i} = 1 \cup \cdots \cup w_c^{R_i} = c\big).$$

Because $w_c^{R_i} = 1, \dots, w_c^{R_i} = c$ are mutually exclusive, we have

$$\mathbb{P}\big(w_c^{R_i} \in \{1,\dots,c\}\big) = \sum_{i=1}^{c} \mathbb{P}\big(w_c^{R_i} = i\big) = \frac{c}{b+2c}.$$

Now, for any $r \in \{2,\dots,c\}$, observe that, by Bayes' rule,

$$\begin{aligned}
&\mathbb{P}\Big(w_{c-(r-1)}^{R_i} \in \{1,\dots,c\} \cap \big(w_c^{R_i} \notin \{1,\dots,c\} \cap \cdots \cap w_{c-(r-2)}^{R_i} \notin \{1,\dots,c\}\big)\Big) \\
&= \mathbb{P}\big(w_{c-(r-1)}^{R_i} \in \{1,\dots,c\} \mid w_c^{R_i} \notin \{1,\dots,c\} \cap \cdots \cap w_{c-(r-2)}^{R_i} \notin \{1,\dots,c\}\big) \\
&\qquad \times \mathbb{P}\big(w_c^{R_i} \notin \{1,\dots,c\} \cap \cdots \cap w_{c-(r-2)}^{R_i} \notin \{1,\dots,c\}\big).
\end{aligned} \tag{3.9}$$

For all $r \in \{2,\dots,c\}$, the first term on the right-hand side of (3.9) is equal to

$$\frac{c}{b+2c-(c-r)}. \tag{3.10}$$

The second term on the right-hand side of (3.9) is equal to

$$1 - \mathbb{P}\big(w_c^{R_i} \in \{1,\dots,c\} \cup \cdots \cup w_{c-(r-2)}^{R_i} \in \{1,\dots,c\}\big) = 1 - \sigma^{r-1}. \tag{3.11}$$

Hence, for all $r \in \{2,\dots,c\}$,

$$\sigma^r = \sigma^{r-1} + \frac{c}{b+2c-(c-r)}(1-\sigma^{r-1}),$$

and $\sigma^c := \mathbb{P}\big(w_c^{R_i} \in \{1,\dots,c\} \cup \cdots \cup w_1^{R_i} \in \{1,\dots,c\}\big)$ can indeed be computed recursively.

Figure 3.6(b). Let us again (without loss of generality) suppose that R_{-i} as described above. Computing the proportion in Figure 3.6b) is equivalent to computing the probability that (R_i, R_{-i}) is one-common when R_i is drawn uniformly at random among the preferences that make (R_i, R_{-i}) almost juror inverse. That is, R_i is drawn uniformly at random among the preferences that differ from R_i^* given by $1\ R_i^*\ \dots\ R_i^*\ b+2c$ by the ranking of a single juror, say $\bar{w}^{R_i}$. For (R_i, R_{-i}) to be one-common, $\bar{w}^{R_i}$ must be one of the c worst jurors in R_{-i} *and* in R_i. The relevant probability is therefore

$$\mathbb{P}\big(\bar{w}^{R_i} \in \{1,\dots,c\} \cap \bar{w}^{R_i} \in \{w_1^{R_i},\dots,w_c^{R_i}\}\big).$$

By Bayes' rule, this is equal to

$$\mathbb{P}\big(\bar{w}^{R_i} \in \{w_1^{R_i},\dots,w_c^{R_i}\} \big| \bar{w}^{R_i} \in \{1,\dots,c\}\big) P\big(\bar{w}^{R_i} \in \{1,\dots,c\}\big).$$

Because R_i is drawn at random among the preferences that differ from R_i^* by the ranking of a single juror, $P(\bar{w}^{R_i} \in \{1,\dots,c\}) = \frac{c}{b+2c}$. Also,

$$\mathbb{P}(\bar{w}^{R_i} \in \{w_1^{R_i},\dots,w_c^{R_i}\} \big| \bar{w}^{R_i} \in \{1,\dots,c\}) = \frac{c}{b+2c-1},$$

where $b+2c-1$ is the number of (equally likely) positions in which $\bar{w}^{R_i}$ can potentially be ranked,[35] and c is the number of these positions for which $\bar{w}^{R_i} \in \{w_1^{R_i},\dots,w_c^{R_i}\}$ given $\bar{w}^{R_i} \in \{1,\dots,c\}$.[36] Hence,

$$\mathbb{P}\big(\bar{w}^{R_i} \in \{1,\dots,c\} \cap \bar{w}^{R_i} \in \{w_1^{R_i},\dots,w_c^{R_i}\}\big) = \frac{c^2}{(b+2c)(b+2c-1)}.$$

[35] By definition of an almost juror inverse profile, $R_i \neq R_i^*$. Because $\bar{w}^{R_i}$ is the only juror the ranking of which is different in R_i and in R_i^*, juror $\bar{w}^{R_i}$ must be ranked differently in R_i and R_i^*. Juror $\bar{w}^{R_i}$ can therefore be ranked in $b+2c-1$ ways in R_i^*.

[36] Indeed, observe that given $\bar{w}^{R_i} \in \{1,\dots,c\}$ and because (R_i^*, R_{-i}) is juror inverse, $\bar{w}^{R_i} \notin \{w_1^{R_i^*},\dots,w_c^{R_i^*}\}$.

One-common profiles in the field

To obtain further evidence of the prevalence of one-common profiles, I consider real-world arbitration cases from the New Jersey Public Employment Relations Commission (Bloom and Cavanagh, 1986; de Clippel et al., 2014). From 1985 to 1996, the Commission used a *veto-rank mechanism* to select an arbitrator in cases involving a union and an employer. In the veto-rank mechanism used by the Commission, the union and the employer are presented with seven potential arbitrators. The union and the employer *simultaneously* challenge three potential arbitrators and rank order the remaining arbitrators. The chosen arbitrator is the unchallenged arbitrator with the lowest combined rank. Except for the way in which an arbitrator is selected when challenges overlap, this procedure is equivalent to one-shot$_M$ with $b = 1$ and $c_d = c_p = 3$.[37]

Out of 750 cases, de Clippel et al. (2014) report that the frequency of overlaps in the challenges was as indicated in Table 3.1.

Number of common challenges	Proportion
0	13%
1	50%
2	34%
3	3%

Table 3.1: Overlap in challenges in the 750 arbitration cases (de Clippel et al., 2014).

Let a party be **truthful** if she challenges her three worst arbitrators (regardless of her reported ranking over the remaining arbitrators). If both parties were truthful in each of the 750 cases, then the data in Table 3.1 would imply that the underlying profile of preferences for arbitrators was one-common in 87% of the 750 cases. However, because truthfulness is *not* a dominant strategy in the veto-rank mechanism, this need not be an accurate estimate of the proportion of one-common profiles in these 750 cases.

To obtain a more realistic estimate, I consider the laboratory experiment on the veto-

[37] de Clippel et al. (2014) report that after 1996, the Commission started selecting the arbitrator at random from the list of unchallenged arbitrators, and so the Commission effectively used one-shot$_M$ after 1996.

rank mechanism described in de Clippel et al. (2014). In the experiment, participants play the veto-rank mechanisms with five arbitrators (i.e., $b = 1$ and $c_d = c_p = 2$). The participants are randomly assigned to four different profiles of preferences, denoted Pf1, Pf2, Pf3, and Pf4 (see de Clippel et al., 2014). For each profile, de Clippel et al. (2014) observe 350 instances of the game being played.

Based on the experimental data from de Clippel et al. (2014), I compute for each profile the proportion of plays in which *both* parties were truthful. I also compute this proportion across the four profiles. These proportions are reported in Table 3.2.

Profile	Proportion of plays in which *both* parties challenged their two worst arbitrators
Pf1	65%
Pf2	45%
Pf3	38%
Pf4	24%
Across the four profiles	43%

Table 3.2: Proportion of experimental plays of the veto-rank mechanism in de Clippel et al. (2014) in which *both* players challenged their two worst arbitrators.

I propose to use the values in Table 3.2 as estimates of the proportion of the 750 New Jersey cases in which *both* parties were truthful. Whichever estimate x from Table 3.2 is used, $x - 13\%$ is a lower bound on the proportion of one-common profiles. This lower bound is obtained by assuming that both parties were truthful in *all* 13% of cases in which the challenges did not overlap.

The obtained lower bounds are illustrated in Figure 3.7 for different values of the truthfulness estimate. Even under the most conservative truthfulness estimate (i.e., 24%), the lower bound on the proportion of one-common profiles is 11%. Using the average truthfulness across the four profiles as an estimate, the lower bound on the proportion of one-common profiles is 30%.

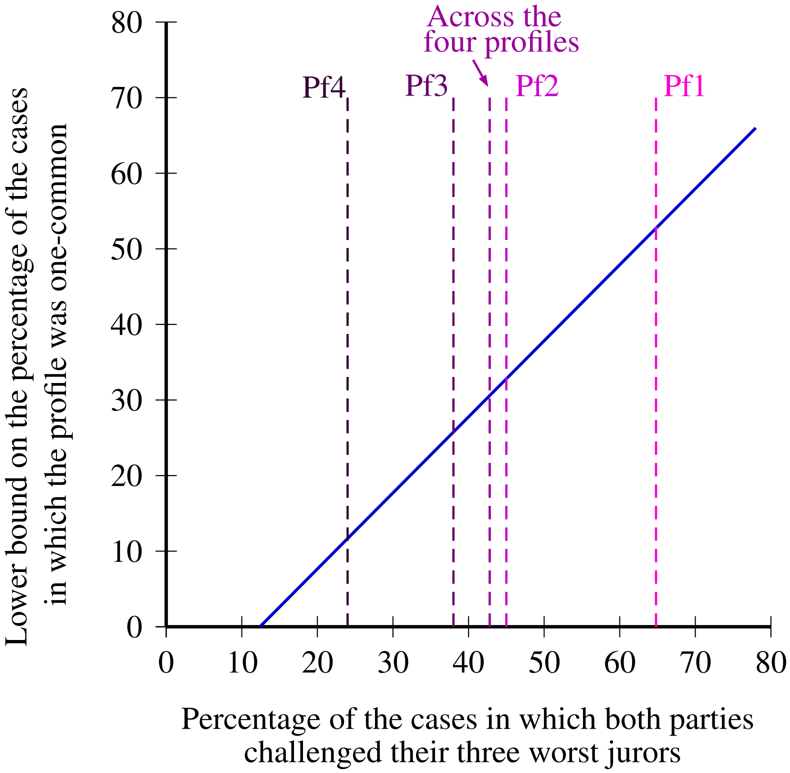


Figure 3.7: Lower bounds on the percentage of one-common profiles in the 750 arbitration cases (blue line) as a function of the percentage of cases in which both parties were truthful. Estimates for the later percentage using experimental data from de Clippel et al. (2014) are shown by the dashed vertical lines (see Table 3.2).

3.8 Prevalence of alternating$_M$-one-common profiles

The procedure used to reach subgame γ^* in Example 3.6 can be generalized. In alternating$_M$, for any set of jurors T of $b+2$ jurors, there exists a subgame γ satisfying (a) and (b) such that T is the set of unchallenged jurors and each party has one challenge left.[38] Hence, a profile is alternating$_M$-one-common if there exists a set T containing $b+2$ jurors such that R_p and R_d have the same worst juror among the jurors in T.

This sufficient condition can be used to prove the following result.

Proposition 3.8. *Every one-common profile is alternating$_M$-one-common.*

Proof. Consider an arbitrary one-common profile (R_d, R_p). Let N_i^1 be the set of jurors in

[38] Such a subgame is reached, for example, after d alone challenges jurors in $N \backslash T$ for $c_d - 1$ rounds, followed by p alone challenging remaining jurors among $N \backslash T$ for $c_p - 1$ rounds.

N that are ranked in positions 1 to $(b+c_{-i})$ according to R_i and N_i^2 the remaining set of jurors ranked in positions $(b+c_{-i}+1)$ to n according to R_i. By assumption, there exits a juror $w \in N_d^2 \cap N_p^2$. Observe that $\#N_i^1 = b+c_{-i}$ and $\#N_i^2 = c_i$, where for any set T, $\#T$ is the cardinality of T.

Because $w \in N_d^2 \cap N_p^2$, we have $w \notin N_d^1 \cap N_p^1$. Hence, if $\#(N_d^1 \cap N_p^1) \geq b+1$, then $\#(N_d^1 \cap N_p^1) \cup \{w\} \geq b+2$ and w is the worst juror for both R_d and R_p among $(N_d^1 \cap N_p^1) \cup \{w\}$, making (R_d, R_p) an alternating$_M$ profile by the sufficient condition identified in the text.

Hence, in order to derive a contradiction, suppose that $\#(N_d^1 \cap N_p^1) \leq b$. That is, for some $i \in \{d, p\}$, *at most* b of the jurors in N_i^1 also belong to N_{-i}^1. Because N_{-i}^1 and N_{-i}^2 partition N, all of the jurors in N_i^1 that do not belong to N_{-i}^1 must belong to N_{-i}^2. Hence, because there are $b+c_{-i}$ jurors in N_i^1 at most b of which belong to N_{-i}^1, *at least* c_{-i} of the jurors in N_i^1 must belong to N_{-i}^2. Recall that $w \in N_i^2$ by assumption. Thus, because N_i^1 and N_i^2 partition N, $w \notin N_i^1$, and w cannot be one of the c_{-i} jurors in N_i^1 that belong to N_{-i}^2. But this implies that there are at least $c_{-i}+1$ jurors in N_{-i}^2, a contradiction. ■

Hence, the proportion of alternating$_M$-one-common profiles is no smaller than the proportion of one-common profiles, and the proportions in Figure 3.6 are therefore lower bounds for the proportions of alternating$_M$-one-common profiles as well. In fact, Proposition 3.8 and Example 3.6 jointly imply that the proportion of alternating$_M$-one-common profiles is *strictly* larger than that of one-common profiles. For the same reason, the lower bounds on the proportions of one-common profiles in the arbitration data in Figure 3.7 are also lower bounds on the proportion of alternating$_M$-one-common profiles in the same data.

Chapter 4

Deferred acceptance is minimally manipulable

4.1 Introduction

The deferred-acceptance *algorithm* was originally developed by Gale and Shapley (1962) to prove by construction the existence of stable matchings in a one-to-one matching problem. A matching is stable if it is individually rational and no two individuals prefer each other to the individuals they are matched with. Following Gale and Shapley's breakthrough, the deferred-acceptance *mechanism* (*DA*) that selects the stable matching constructed by the deferred-acceptance *algorithm* for any report of preferences has been the primary stable mechanism used in theory and practice.

However, because multiple stable mechanisms exist, it is unclear whether *DA* should be used. *DA* has limitations. In particular, it does not give all individuals the incentives to report their preferences truthfully. Although this is true of any stable mechanism (Roth, 1982), the question remains: in one-to-one matching, does any stable mechanism have better incentive properties than *DA*? In other words, is any stable mechanism less manipulable than *DA*? In this paper, I answer negatively. I show that *DA* is minimally manipulable in the sense of partial orders developed by Pathak and Sönmez (2013) (PS) and Arribillaga and Massó (2015) (AM). These partial orders compare for every profile (PS) or preference (AM) the set of individuals who can benefit from misreporting their preference.[1]

Unlike *DA*'s incentive properties, *DA*'s efficiency is rarely questioned. *DA* is often deemed "sufficiently efficient" because it is Pareto efficient. That is, as illustrated in Figure 4.1, *DA* is among the maximal elements of a meaningful efficiency partial order, the Pareto partial order. In that sense, *DA* lies on the *efficiency frontier* of the set of mechanisms because *DA* cannot be unambiguously improved upon in terms of efficiency.

[1] Pathak and Sönmez (2013) introduce several manipulability partial orders. I use the partial order defined

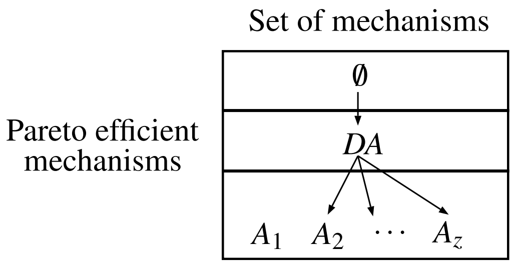


Figure 4.1: Representation of the Pareto partial order. An arrow from mechanism A to mechanism B indicates that A Pareto dominates B. Because DA is Pareto efficient, the set of mechanisms that Pareto dominate DA is the empty set $\emptyset$. Mechanisms $A_1,\ldots,A_z$ are not Pareto efficient.

I show that, in a similar sense, *DA* lies on the *manipulability* frontier of the set of stable mechanisms: No stable mechanism is less manipulable than *DA*, i.e., *DA* is *minimally* manipulable among the stable mechanisms. This result is important because it indicates that *DA* cannot be unambiguously improved upon in terms of the two most common desiderata for matching mechanisms: stability and non-manipulability. There is no point searching for mechanisms that dominate *DA* in both respects. Somewhat surprisingly, a mechanism that was developed to prove the existence of stable matchings turns out to satisfy Sen's demanding requirement of lying on the "desirability" frontier of the set of mechanisms, "just before [impossibility results apply and] all possibilities are eliminated" (Sen, 1999, p. 354)

With manipulability *partial* orders, there is always a risk that some mechanisms lie trivially on the order's frontier because they cannot be compared with any other mechanism.[2] I show that this is not the case of *DA* and the PS- or AM-partial orders. *DA is* less manipulable than other stable mechanisms. Although stable mechanisms that are more manipulable than *DA* are rare with PS, they are abundant with AM. I also show that the desirability frontier is "small". In contrast to *DA*, most stable mechanisms are dominated by another stable mechanism and therefore leave room for manipulability improvements at no stability cost.

in Pathak and Sönmez (2013, Section III).

[2] See, for example, Chen et al. (2016) who show that *no* two stable mechanisms can be compared in the sense of a third comparison partial order also proposed by Pathak and Sönmez (2013).

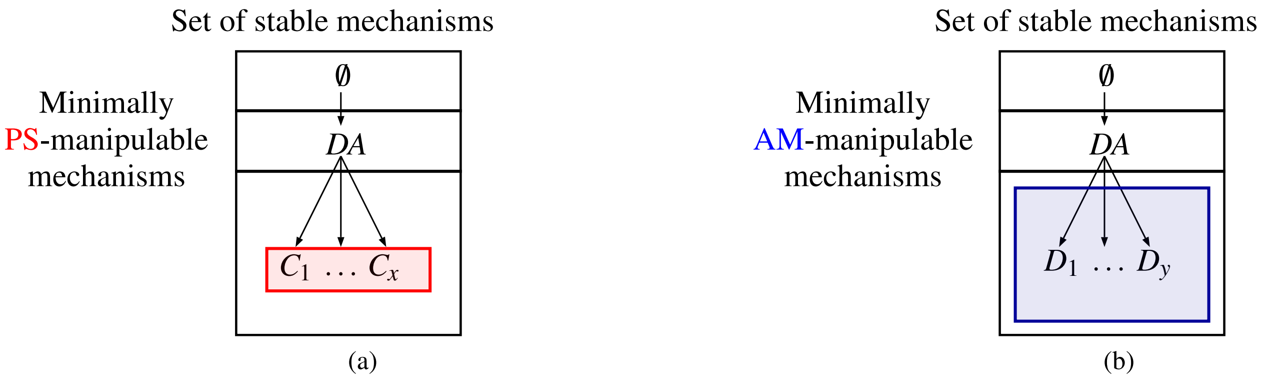


Figure 4.2: Representation of some of the results: An arrow from mechanism A to mechanism B indicates that A is less (AM- or PS-) manipulable than B. A box surrounding a list of mechanisms represent the relative size of that list.

The results described so far are illustrated in Figure 4.2.

In proving the results illustrated in Figure 4.2, I show that with a stable mechanism, individuals cannot benefit from misreporting their preference if and only if they match with their most preferred *achievable* mate.[3] An implication of this observation is that, when some individuals cannot benefit from manipulations, some other individuals must match with their *least* preferred achievable mates (Gale and Shapley, 1962), which points toward a tension between fairness and manipulability.

To achieve minimal manipulability, *DA* always matches one side of the market with their most preferred achievable mate, and the other side with their least preferred achievable mate. This is sometimes viewed as a defect of *DA* in terms of fairness. In response, fairer stable mechanisms have been designed that select "intermediate" stable matchings in which fewer individuals match with their least preferred achievable mate.

I show that such improvements in fairness come at the cost of an increase in manipulability. A fairness criterion that I call *miniworst* requires that the set of individuals who match with their least preferred achievable mate be minimal (with respect to inclusion). As I show, if a stable mechanism A is miniworst, then A is also *maximally* manipulable, i.e.,

[3] An achievable mate is a mate that the individual matches with under *some* stable matching.

no other stable mechanism is more manipulable than A (in the case of the PS-partial order, the miniworst criterion *characterizes* maximal manipulability). All miniworst mechanisms are also dominated by DA in the sense of AM (although this is not true in the sense of PS). A similar trade-off between manipulability and fairness is identified in the case of the median stable mechanisms introduced by Teo and Sethuraman (1998). These results reinforce my finding that DA lies on the desirability frontier because they show that DA cannot be improved upon in terms of fairness without compromising with stability or manipulability.

Related literature. This paper contributes to the literature comparing the manipulability of mechanisms that, like the stable matching mechanisms, fail to have a truthful dominant strategy.[4] It also contributes to the literature on fair stable matchings by identifying a trade-off between manipulability and fairness.[5] Of the results in this paper, only the minimal manipulability of DA in the sense of PS can be obtained as a direct corollary of an existing result (in Pathak and Sönmez, 2013). In particular, the comparisons in the sense of AM and the manipulability analysis of fair stable mechanisms are novel.

In contrast to the previous literature, this paper relies on *two* different partial orders. Comparing mechanisms according to different partial orders is important. The partial orders developed by PS and AM each capture interesting aspects of the relative manipulability of mechanisms. However, as I show in Section 4.2, each partial order has limitations. Using both partial orders mitigates some of these limitations. Furthermore, considering both of these partial orders, it is possible to investigate the robustness of the comparisons obtained with each of them.

Chen et al. (2016) show that *no* two stable mechanisms can be compared in the sense of a third partial order that was also proposed in Pathak and Sönmez (2013). In contrast, I show that many stable mechanisms can be compared in the sense of both AM and another

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

[5] For previous papers on fair stable matchings, see Knuth (1997), Irving et al. (1987), Teo and Sethuraman (1998), and Klaus and Klijn (2006).

one of Pathak and Sönmez's partial orders.

The paper is organized as follows. Section 4.2 gives a general definition of the PS and AM partial orders. Beginning with Section 4.3, the focus is on one-to-one matching. Section 4.3 defines the one-to-one matching environment. Section 4.4 introduces preliminary results. Section 4.5 compares the manipulability of *DA* with the manipulability of all stable mechanisms. Section 4.6 compares more closely the manipulability of *DA* with the manipulability of two classes of fair stable mechanisms: the miniworst and the median stable mechanisms. I conclude with some remarks and open questions. Omitted proofs may be found in the Appendix.

4.2 Two partial orders for manipulability comparisons

The set of individuals is N with $n := \#N$ (for any set S, $\#S$ is the cardinality of S). The set of outcomes is T. An individual $i \in N$ has a preference R_i on the set of outcomes T. For any $s,t \in T$, $s\ R_i\ t$ indicates a weak preference for s over t and $s\ P_i\ t$ a strict preference (i.e., $s\ R_i\ t$ but not $t\ R_i\ s$). For any $i \in N$, the domain of i's preferences is $\mathscr{D}_i$.

A preference profile $R := (R_1, \ldots, R_n)$ is a list of the preferences of all the individuals in N. The domain of preference profiles is $\mathscr{D} := \times_{i\in N}\mathscr{D}_i$. The list of preferences in R for everyone but i is $R_{-i} \in \mathscr{D}_{-i} := \times_{j\in N\setminus\{i\}}\mathscr{D}_j$. A pair $(T,\mathscr{D})$ is called an **environment**. A **mechanism** A is a function that associates every preference profile $R \in \mathscr{D}$ with an outcome $A(R) \in T$.

For each profile, the **PS-partial order** compares with respect to inclusion the set of individuals who can manipulate their report. Formally, an individual $\boldsymbol{i \in N}$ **can manipulate** $\boldsymbol{A}$ **given** ***profile*** $\boldsymbol{R} \in \mathscr{D}$ if for some $R_i' \in \mathscr{D}_i$,

$$A(R_i', R_{-i})\ P_i\ A(R_i, R_{-i}). \tag{4.1}$$

That is, i fails to have a truthful best response in A when i's preference is R_i and other individuals report R_{-i}. Mechanism A is **no more PS-manipulable than** mechanism B if for *all* $R \in \mathscr{D}$,

$$\begin{aligned} &\{i \in N \mid i \text{ can manipulate } A \text{ given } R\} \\ \subseteq\ &\{i \in N \mid i \text{ can manipulate } B \text{ given } R\}. \end{aligned} \tag{4.2}$$

That is, if i fails to have a truthful best response in A when i's preference is R_i and other individuals report R_{-i}, the same is true in B. Mechanism A is **less PS-manipulable than** mechanism B if A is no more PS-manipulable than B but the converse is not true, i.e., (4.2) holds and in addition, for some $R^* \in \mathscr{D}$,

$$\begin{aligned} &\{i \in N \mid i \text{ can manipulate } A \text{ given } R^*\} \\ \subset\ &\{i \in N \mid i \text{ can manipulate } B \text{ given } R^*\}. \end{aligned} \tag{4.3}$$

Alternatively, for each *preference*, the **AM-partial order** compares with respect to inclusion the set of individuals who can manipulate their report for *at least one* report of the other individuals' preferences. For any $R_* \in \cup_{i \in N} \mathscr{D}_i$, let N_{R^*} be the set of individuals $i \in N$ for whom $R_* \in \mathscr{D}_i$. Formally, an individual $\boldsymbol{i \in N_{R_*}}$ **can manipulate** $\boldsymbol{A}$ **given** ***preference*** $\boldsymbol{R_* \in \mathscr{D}_i}$ if for some $R'_i \in \mathscr{D}_i$ and *some* $R_{-i} \in \mathscr{D}_{-i}$,

$$A(R'_i, R_{-i})\ P_*\ A(R_*, R_{-i}). \tag{4.4}$$

That is, i does not have a truthful dominant strategy in A given preference R_*. Mechanism A is **no more AM-manipulable than** mechanism B if for *all* $R_* \in \cup_{i \in N} \mathscr{D}_i$,

$$\begin{aligned} &\{i \in N_{R_*} \mid i \text{ can manipulate } A \text{ given } R_*\} \\ \subseteq\ &\{i \in N_{R_*} \mid i \text{ can manipulate } B \text{ given } R_*\}.^{6} \end{aligned} \tag{4.5}$$

That is, if i fails to have a truthful dominant strategy in A given R_*, the same is true in B. Mechanism A is **less AM-manipulable than** mechanism B if A is no more AM-manipulable than B but the converse is not true, i.e., (4.5) holds and in addition, for some $R_{**} \in \cup_{i \in N} \mathscr{D}_i$,

$$\begin{aligned} &\{i \in N_{R_{**}} \mid i \text{ can manipulate } A \text{ given } R_{**}\} \\ \subset \ &\{i \in N_{R_{**}} \mid i \text{ can manipulate } B \text{ given } R_{**}\}. \end{aligned} \tag{4.6}$$

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

Of particular importance for this paper are the concepts of minimal and maximal manipulability. For any class of mechanisms $\mathscr{A}$, mechanism $A \in \mathscr{A}$ is **minimally (PS-, AM-) manipulable in** $\mathscr{A}$ if there exists no $B \in \mathscr{A}$ such that B is *less* (PS-, AM-) manipulable than A. Conversely, mechanism $A \in \mathscr{A}$ is **maximally (PS-, AM-) manipulable** in $\mathscr{A}$ if there exists no $B \in \mathscr{A}$ such that B is *more* (PS-, AM-) manipulable than A.

As AM explain, the PS-partial order is a sub-relation of the AM-partial order: If A is no more PS-manipulable than B, then A is also no more AM-manipulable than B. The converse is not true. In particular, A can be less AM-manipulable than B although A fails to be less PS-manipulable than B, and vice versa.[7] As a consequence, the concepts of minimal PS- and minimal AM-manipulability are logically independent (by symmetry, the same is true of maximal PS- and maximal AM-manipulability).

Both partial orders have advantages and disadvantages over one another. Example 4.1 illustrates situations in which the AM-partial order yields counter-intuitive comparisons,

[6] In the context of a two-sided matching environment in which the individual domains do not intersect, the above sets are either singletons or empty.

[7] See Proposition 4.8 for examples of mechanisms that are more AM-manipulable than *DA* but fail to be more PS-manipulable than *DA*. For the converse scenario, consider a mechanism B that every individual can manipulate given any profile, and a mechanism A that every individual can manipulate given any profile *except* R^*, for which no individual can manipulate. Mechanism A is less PS-manipulable than mechanism B. However, if $\#\mathscr{D}_i \geq 2$ for all $i \in N$, then for any $R_i \in \mathscr{D}_i$, i can manipulate both A and B given any profile (R_i, R_{-i}) with $R_{-i} \neq R^*_{-i}$. Hence, i can manipulate both A and B given any preference R_i and A is therefore *not* less AM-manipulable than B (A is only *no more* AM-manipulable than B).

whereas the PS-partial order refrains from comparing the two mechanisms at stake. Example 4.2 illustrates situations in which the PS-partial order is unduly incomplete, whereas the AM-partial order is not. In general, the PS-partial order is more conservative but less complete, whereas the AM-partial order is more complete but more often yields counter-intuitive comparisons. Each partial order therefore mitigates some of the other partial order's limitations and using both is a useful robustness check.

Example 4.1. Suppose that *no* $i \in N_{R_*}$ can manipulate A given some preference R_* but that given any preference $R_{**} \neq R_*$, every $i \in N_{R_{**}}$ can manipulate A. Also, for any $R_{**} \neq R_*$, every $i \in N_{R_{**}}$ can manipulate A given any profile of the form (R_{**}, R_{-i}), for any $R_{-i} \in \mathscr{D}_{-i}$. Finally, given any preference $R_\circ$, every $i \in N_{R_\circ}$ can manipulate B, but i can only manipulate B given *a single* profile $(R_\circ, R_{-i}^{R_\circ})$.

By construction, A is less AM-manipulable than B. This seems counter-intuitive because A improves upon B in terms of manipulability for a single preference R_* and does much worse for the vast majority of profiles. In this sense, the conclusion that A is less AM-manipulable than B can be viewed as a "false positive". In this case, the PS-partial order is more consistent with the intuition because it refrains from concluding that A is less PS-manipulable than B.

Example 4.2. For some $j \in N$, suppose that the set $\mathscr{D}_{-j}$ can be partitioned into two sets $\mathscr{D}^1_{-j}$ and $\mathscr{D}^2_{-j}$ of equal size ($\#\mathscr{D}^1_{-j} = \#\mathscr{D}^2_{-j}$). Further, j can manipulate A given any profile R with $R_{-j} \in \mathscr{D}^1_{-j}$, but cannot manipulate A given any profile R with $R_{-j} \in \mathscr{D}^2_{-j}$. Also, *no* $i \in N \backslash \{j\}$ can manipulate A given any preference $R_i \in \mathscr{D}_i$.

Conversely, j can manipulate B given any profile R with $R_{-j} \in \mathscr{D}^2_{-j}$, but cannot manipulate B given any profile R with $R_{-j} \in \mathscr{D}^1_{-j}$. Finally, *every* $i \in N \backslash \{j\}$ can manipulate B given any preference $R_i \in \mathscr{D}_i$.

Because the sets of profiles for which j can manipulate A and B are different, A and B cannot be compared using the PS-partial order. This incompleteness is troublesome because in terms of manipulability, A does much better than B for everyone but j, and A

performs similarly to B for j. In this case, the AM-partial order is more complete than the PS-partial order in a way that is consistent with intuition because it concludes that A is less AM-manipulable than B.

4.3 The one-to-one environment

In the one-to-one two-sided environment (henceforth, the *one-to-one environment*), the set N is partitioned into a set of women W and a set of men M. Throughout, $\#W, \#M \geq 3$. A woman $w \in W$ has a preference R_w on the set of men and herself ($M \cup \{w\}$) and a man $m \in M$ has a preference R_m on the set of women and himself ($W \cup \{m\}$). For any $i \in N$, the set of individuals for which i has a preference are i's **mates**. Henceforth, let $\mathscr{D}_i$ be the domain of all strict preferences over i's mates (i.e., preferences R_i for which $t_1 \; P_i \; t_2$ or $t_2 \; P_i \; t_1$ for any pair (t_1, t_2) of i's mates with $t_1 \neq t_2$).

A **matching** is a function $\mu \colon N \to N$ that matches every individual $i \in N$ with one of his or her mates, and such that matchings are reciprocal. Formally, (i) $\mu(w) \in M \cup \{w\}$ for all $w \in W$ and $\mu(m) \in W \cup \{m\}$ for all $m \in M$, and (ii) $\mu(\mu(i)) = i$ for all $i \in N$.

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

Given matching μ and profile R, a **blocking pair** consists of a man and a woman who prefer matching together to being matched according to μ. Formally a blocking pair in μ is any $(w,m) \in W \times M$ for which $m \; P_w \; \mu_w$ and $w \; P_m \; \mu_m$. For any $i \in N$, if $i \; P_i \; j$ for some $j \in N$, then mate j is **unacceptable** to i. A matching is **individually rational** if no individual matches with an unacceptable mate. A **matching** is **stable** if it does not contain any blocking pairs *and* it is individually rational. A **mechanism** is **stable** if it selects a stable matching for every profile.

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

is minimally manipulable *in the class of stable mechanisms*.

As is well-known, the *deferred acceptance* mechanism (DA) comes in two variants: women-proposing (DA^W) and men-proposing (DA^M). For any $i \in N$, the variant of DA in which i's side proposes is denoted DA^i. When a property applies irrespective of the proposing side, a deferred acceptance mechanism is simply referred to as DA. In this context, an individual $i \in N$ is a **proposer** if i is on the side that proposes in DA (e.g., W in DA^W) and an **acceptor** if i is on the side that does not propose (e.g, M in DA^W). Typical proposers and acceptors are denoted by $p \in N$ and $a \in N$, respectively.

For any $i \in N$, any individual $j \in N$ is an **achievable** mate given R if j matches with i in some stable matching when the profile is R. For all $i \in N$ and all $R \in \mathscr{D}$, f_i^R is i's most preferred achievable mate given R. Similarly, l_i^R is i's least preferred achievable mate given R. Observe that, because preferences are strict, f_i^R and l_i^R are unique.

4.4 Preliminary results

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

Lemma 4.1 (Gale and Shapley, 1962)**.** *For any $R \in \mathscr{D}$, (i) $DA(R)$ is stable with respect to R and (ii) for every proposer $p \in N$, $DA_p(R) = f_p^R$ and for every acceptor $a \in N$, $DA_a(R) = l_a^R$.*

The next lemma plays an essential role in most of the results in this paper. Lemma 4.2 shows that, in a stable mechanism, i can benefit from misreporting her or his preference R_i when the other individuals report R_{-i} *if and only if* i does not match with f_i^R.[8]

[8] The proof of Lemma 4.2 is inspired by the fact that every report of a preference is dominated by the report of a *truncation* (see the Appendix for a formal definition), which was first proven by Roth and Vande Vate (1991, Theorem 2). The proof of Lemma 4.2 follows the same proof strategy as the proof of Roth and Vande Vate's theorem. This proof strategy is also used in the proof of Pathak and Sönmez (2013, Lemma 1). A similar result appears in Coles and Shorrer (2014).

Lemma 4.2. *For any stable mechanism A, any $i \in N$, and any $R \in \mathscr{D}$,*

$$A_i(R_i, R_{-i}) \; R_i \; A_i(R_i', R_{-i}) \qquad \textit{for all } R_i' \in \mathscr{D}_i \tag{4.7}$$

if and only if

$$A_i(R_i, R_{-i}) = f_i^R. \tag{4.8}$$

The next section uses Lemmas 4.1 and 4.2 to compare the manipulability properties of *DA* with those of other stable mechanisms.

4.5 Maximal and minimal manipulability among stable mechanisms

The following characterization of the PS- and AM-partial orders is obtained using Lemma 4.2.

Proposition 4.1. *(i) Stable mechanism A is no more PS-manipulable than stable mechanism B* if and only if*, for* all *$R \in \mathscr{D}$,*

$$\{i \in N \mid B_i(R) = f_i^R\} \subseteq \{i \in N \mid A_i(R) = f_i^R\}. \tag{4.9}$$

(ii) Stable mechanism A is no more AM-manipulable than stable mechanism B if and only if*, for* all *$R_* \in \cup_{i \in N} \mathscr{D}_i$,*

$$\begin{aligned} &\{i \in N_{R_*} \mid B_i(R_*, R_{-i}) = f_i^{(R_*, R_{-i})} \textit{ for all } R_{-i} \in \mathscr{D}_{-i}\} \\ \subseteq \; &\{i \in N_{R_*} \mid A_i(R_*, R_{-i}) = f_i^{(R_*, R_{-i})} \textit{ for all } R_{-i} \in \mathscr{D}_{-i}\}. \end{aligned} \tag{4.10}$$

Proposition 4.1 is used to establish the following characterization of minimal and maximal PS-manipulability.[9]

[9] Recall that minimal and maximal manipulability properties are implicitly defined with respect to the class of stable mechanisms.

Proposition 4.2. *A mechanism A is minimally (resp. maximally) PS-manipulable if and only if there does not exist a profile $R \in \mathscr{D}$ and a stable matching μ such that*

$$\{i \in N \mid A_i(R) = f_i^R\} \subset \{i \in N \mid \mu_i = f_i^R\} \tag{4.11}$$

$$(\textit{resp. } \{i \in N \mid A_i(R) = f_i^R\} \supset \{i \in N \mid \mu_i = f_i^R\}). \tag{4.12}$$

I now show that DA is minimally manipulable.[10] In the case of minimal PS-manipulability, this follows straightforwardly from Proposition 4.2 and Lemma 4.1. Enlarging (with respect to inclusion) the set of *acceptors* who match with their most preferred achievable mate implies that the set of *proposers* who match with their most preferred achievable mates shrinks. Thus, condition (4.11) can never be satisfied when $A = DA$.

An informative characterization of minimal and maximal AM-manipulability is harder to obtain. Unlike comparisons in terms of PS that can be performed on a “profile by profile” basis, comparisons in terms of AM cannot be performed “preference by preference”. Enlarging the set of individuals who cannot manipulate given some preference R_* may have an impact on the set of individuals who cannot manipulate given some other preference R_{**}, and no equivalents to (4.11) and (4.12) exist for the AM-partial order. However, the minimal AM-manipulability of DA can be proven directly using Lemmas 4.1 and 4.2.

Proposition 4.3. *(i) DA is minimally manipulable. (ii) There exists stable mechanisms that are more manipulable than DA.*

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

$$\begin{array}{llll} R_{w_1}: & m_3 & m_2 & m_1 \\ R_{w_2}: & m_2 & m_1 & m_3 \\ R_{w_3}: & m_1 & m_3 & m_2 \end{array} \qquad \begin{array}{llll} R_{m_1}: & w_1 & w_2 & w_3 \\ R_{m_2}: & w_3 & w_1 & w_2 \\ R_{m_3}: & w_2 & w_3 & w_1 \end{array} . \tag{4.13}$$

[10] Recall that manipulability properties that do not refer to either PS of AM hold for both partial orders.

When being self-matched is omitted as in (4.13), being self-matched is implicitly the least preferred outcome. Given R, the stable matchings are

$$\begin{array}{rccc} & w_1 & w_2 & w_3 \\ \cline{2-4} \mu^1: & m_1 & m_3 & m_2 \\ \mu^2: & m_2 & m_1 & m_3 \\ \mu^3: & m_3 & m_2 & m_1 \end{array}.$$

For example, in μ^1, w_1 matches with m_1, w_2 matches with m_3, and w_3 matches with m_2. The mechanism DA^R constructed from DA by changing the stable matching selected for R to μ^2 (and changing nothing else) is more manipulable than DA because, when the profile is R, DA^R matches no individual with her or his most preferred achievable mate.

To see that multiple mechanisms are more manipulable than DA, repeat the above argument for a variant of (4.13) in which m_1 appears on the *downward* diagonal of the womens' profile and w_1 appears on the *upward* diagonal of the mens' profile.[11] ■

Proposition 4.3 shows that DA cannot be improved upon in terms of manipulability without compromising on stability.[12] Minimal manipulability is a relatively *un*common property in the class of stable mechanisms because few stable mechanisms satisfy the conditions for minimal manipulability imposed by Proposition 4.1. For example, consider the PS-partial order. It takes a single profile R and a single stable matching μ for which (4.11) holds for a stable mechanism to violate minimal PS-manipulability. As a consequence, most stable mechanisms leave room for improvement in terms of PS-manipulability at no stability cost.[13]

[11] It is easy to see how the above argument applies to profiles similar to (4.13) when $\#M = \#W > 3$ (see (4.27) in the Appendix). If $\#M \neq \#W$, simply let all individuals other than $\{w_1, \ldots, w_{\max\{\#M,\#W\}}, m_1, \ldots, m_{\max\{\#M,\#W\}}\}$ rank being self-matched first.

[12] For the case of minimal PS-manipulability, Proposition 4.3.(i) can be viewed as a consequence of Pathak and Sönmez (2013, Theorem 2).

[13] Nevertheless, minimally AM-manipulable mechanisms different from DA and minimally PS-manipulable mechanisms different from DA can be shown to exist. Whether any of these mechanisms is of interest is left as an open question.

To illustrate this formally, I focus on the case $\#W = \#M$ and on the domain $\bar{\mathscr{D}}$ of all profiles in which individuals rank being self-matched last. For each profile $R \in \bar{\mathscr{D}}$ a stable mechanism selects a stable matching. There are therefore as many stable mechanisms as there are ways to select a stable matching for each profile in $\bar{\mathscr{D}}$. The proportion of (stable) mechanisms satisfying X is the number of stable mechanisms satisfying X divided by the total number of stable mechanism.

Proposition 4.4. *Suppose that $\#W = \#M = h$ and the domain of profiles is $\bar{\mathscr{D}}$. (i) The proportion of minimally PS-manipulable mechanisms is at most $\left(\frac{2}{h}\right)^{h!}$. (ii) The proportion of minimally AM-manipulable mechanisms is at most $\left(\frac{h}{(h-1)!} + \frac{1}{h((h-1)!)^2}\right)$.*

A complete proof of Proposition 4.4 appears in the Appendix. In order to gain some insight into this result, I provide a sketch of the proof.

(i) For any h, any individual $i \in N$, and any preference $R_i \in \mathscr{D}_i$, it is possible to construct $R_{-i}^{R_i}$ such that the profile $(R_i, R_{-i}^{R_i})$ mimics the "Latin Square" pattern of profile (4.13). Any of these Latin Square profiles admits h stable matchings. For any Latin Square profile, out of the h stable matchings, minimally PS-manipulable mechanisms must select either the men optimal or the women optimal matching. The upper bound $\left(\frac{2}{h}\right)^{h!}$ is obtained by considering the proportion of stable mechanisms that select one of these two stable matching in every Latin Square profile.

(ii) If a mechanism A is minimally AM-manipulable, then DA cannot be less AM-manipulable than A. By Lemmas 4.1 and 4.2, this implies that there exists an acceptor $a \in N$, a proposer $p \in N$, and a pair of preferences $(R_a, R_p) \in \bar{\mathscr{D}}_a \times \bar{\mathscr{D}}_p$ such that A always matches a and p with their most preferred achievable mate when they report R_a or R_p, respectively.[14] In particular, a and p must match with their most preferred achievable mate given the Latin Square profiles $(R_a, R_{-a}^{R_a})$ and $(R_p, R_{-p}^{R_p})$. Because this only needs to be true for a single acceptor-proposer pair and for a single pair of profiles, this fact alone is

[14] To be precise, only such mechanisms *and DA* can possibly be minimally AM-manipulable. The addition of *DA* is reflected by the second term in the bound of Propositions 4.4.(ii) and 4.5.(i).

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

Table 4.1: Numerical values for the upper and lower-bounds of Propositions 4.4 and 4.5.

not sufficient to prove that the proportion of mechanisms that are more AM-manipulable than *DA* is large. For every $i \in N$ and every $R_i \in \mathscr{D}_i$, it is however possible to construct sufficiently many variants of the Latin Square profiles $(R_i, R_{-i}^{R_i})$ to show that this proportion is in fact bounded below by $1 - \left(\frac{h}{(h-1)!} + \frac{1}{h((h-1)!)^2}\right)$. As a consequence, the proportion of minimally AM-manipulable mechanisms is at most $\left(\frac{h}{(h-1)!} + \frac{1}{h((h-1)!)^2}\right)$.

As illustrated in Table 4.1, the bounds in Proposition 4.4 converge rapidly to zero as h increases.

Although Proposition 4.3.(ii) shows that *DA* is not *trivially* minimally manipulable, the partial orders could still be coarse and rank *DA* above only a few other stable mechanisms.[15] This is only true for the PS-partial order. Most mechanisms are more AM-manipulable than *DA*. However, by the definition of the PS-partial order, *DA* fails to be less PS-manipulable than any stable mechanism that, for at least one profile R^*, selects a stable matching where some acceptor a matches with $f_a^{R^*}$, and such mechanisms abound.

This illustrates the importance of using multiple partial orders when comparing mechanisms. Because the PS-partial order is oversensitive to "*outlier*" profiles, it concludes that few stable mechanisms have worse manipulability properties than *DA*. Considering both the PS- and the AM-partial order paints a more complete picture.

Proposition 4.5. *Suppose that $\#W = \#M$ and the domain of profiles is $\bar{\mathscr{D}}$. (i) DA is less*

[15] A mechanism would be *trivially* minimally manipulable if it cannot be compared with any other stable mechanism (see footnote 2).

AM-manipulable than at least $1-\left(\frac{h}{(h-1)!}+\frac{1}{h((h-1)!)^2}\right)$ *of the stable mechanisms. (ii) DA is less PS-manipulable than* at most $\left(1-\frac{1}{h}\right)^{h!}$ *of the stable mechanisms.*[16]

The proof of Proposition 4.5.(i) is similar to that of Proposition 4.4. To gain some insight into Proposition 4.5.(ii), observe that *DA* fails to be less PS-manipulable than any stable mechanism *A* that, for *at least one* profile R^*, selects a stable matching for which some acceptor *a* matches with f_a^R. Such mechanisms abound. For example, a third of the stable mechanisms select matching μ^1 for the Latin Square profile (4.13). Hence, only $(1-\frac{1}{3})$ of the stable mechanisms select a matching for profile (4.13) that allows them to be more PS-manipulable than DA^W. Considering variants of (4.13) yields the bound in the proposition.

Again, the bounds in Proposition converge rapidly as *h* increases (see Table 4.1).

The results so far clarify the manipulability properties of *DA* when compared with the class of *all* stable mechanisms (Propositions 4.3 to 4.5 are illustrated in Figure 4.2). The class of stable mechanisms contains a number of mechanisms that are "contrived" in the sense that they associate profiles with stable matchings in a very unsystematic way. Rather than comparing *DA* with the whole class of stable mechanisms, it may be useful to compare *DA* with *salient* subsets of this class. An important aspect of stable mechanisms is the fairness with which they match individuals with their mates. The next section compares the manipulability of *DA* with that of stable mechanisms designed to improve upon *DA* in terms of fairness.

4.6 A conflict between fairness and manipulability

When only ordinal information on preferences is available, it is hard to define a comprehensive concept of fairness. Some natural reference points can, however, be used to devise minimal fairness requirements. One such reference point is the situation in which

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

an individual receives her or his least preferred outcome out of the set of admissible outcomes.

4.6.1 A conflict between miniworst and manipulability

In the spirit of the *minimum regret* criterion (Knuth, 1997), a minimal fairness requirement is that the set of individuals who receive their least preferred admissible outcome be minimal (with respect to inclusion). Formally, suppose that the function $C : \mathscr{D} \to T$ identifies the set of admissible outcomes $C(R)$ for any profile $R \in \mathscr{D}$. Mechanism A is **miniworst on** C if A only select admissible outcomes and there exists no $R \in \mathscr{D}$ and no outcome $t \in C(R)$ such that

$$\begin{aligned} &\left\{i \in N \mid \text{for all } t' \in C(R),\ t'_i\ R_i\ t_i\right\} \\ \subset\ &\left\{i \in N \mid \text{for all } t' \in C(R),\ t'_i\ R_i\ A_i(R)\right\}.^{17} \end{aligned} \tag{4.14}$$

In matching problems, a natural set of acceptable outcomes is the set of stable matchings. Henceforth, I focus on mechanisms that are miniworst *on the set of stable matchings* and the reference to this set is suppressed.

It is often argued that *DA* is unfair because proposers match with their most preferred achievable mate, whereas acceptors match with their least preferred achievable mate. The miniworst criterion captures this fairness concern. Indeed, observe that a stable mechanism A is miniworst if and only if there exists no $R \in \mathscr{D}$ and no stable matching μ such that

$$\left\{i \in N \mid \mu_i = l_i^R\right\} \subset \left\{i \in N \mid A_i(R) = l_i^R\right\}. \tag{4.15}$$

Considering Latin Square profiles similar to (4.13), it is easy to see that *DA* is not

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

miniworst. For these profiles, a stable matching in which no individual matches with her or his least preferred achievable mate could be selected instead of the extreme matching selected by *DA*. More generally, minimally manipulable mechanisms cannot be miniworst. Let $\mathscr{D}^3$ be the domain of all profiles in $\mathscr{D}$ in which individuals have at least three acceptable mates. When no reference to a subdomain of $\mathscr{D}$ is made, the results hold for the domain $\mathscr{D}$.

Proposition 4.6. *(i) No miniworst mechanism is minimally PS-manipulable. (ii) When the domain is $\mathscr{D}^3$, no miniworst mechanism is minimally AM-manipulable.*

Miniworst mechanisms not only fail to be minimally manipulable, they are also maximally manipulable. In fact, in the case of the PS-partial order, the miniworst criterion characterizes maximal manipulability.

Proposition 4.7. *(i) Mechanism A is miniworst if and only if A is maximally PS-manipulable. (ii) When the domain is $\mathscr{D}^3$, any miniworst mechanism A is maximally AM-manipulable. (iii) When $\#W, \#M \geq 8$ and the domain is $\mathscr{D}^3$, there exist maximally AM-manipulable mechanisms that are not miniworst.*

As observed earlier, for Latin Square profiles, miniworst mechanisms select matchings in which no individual matches with her or his least preferred achievable mates. It then follows from Lemma 4.1 that no individual matches with her or his most preferred achievable mate either. Because any preference can be included in at least one Latin Square profile, this implies that *DA* is less AM-manipulable than any miniworst mechanism.[18] That is, in terms of the miniworst criterion, any fairness improvement upon *DA* comes at the cost of an increase in AM-manipulability (provided that the improvement does not compromise on stability).

A similar result does not hold for PS-manipulability. The reason is the same as in Proposition 4.5: for *DA* *not* to be less PS-manipulable than some mechanism *A*, it is sufficient that *A* matches an acceptor with her or his most preferred achievable mate for a *single*

[18] This does *not* follow directly from Proposition 4.7. In general, it is possible for a mechanism to be minimally manipulable but fail to be less manipulable than a maximally manipulable mechanism. See the example in the proof of Proposition 4.8 for the case of PS-manipulability.

profile (which is the case of some miniworst mechanisms, see the proof of Proposition 4.8.(ii)).

Proposition 4.8. *When the domain is $\mathscr{D}^3$, (i) DA is less AM-manipulable than any miniworst mechanism, but (ii) some miniworst mechanisms are not more PS-manipulable than DA.*

Proof of (ii). Consider the following profile:

$$\begin{array}{llll} R_{w_1}: & m_1 & m_2 & m_3 \\ R_{w_2}: & m_2 & m_3 & m_1 \\ R_{w_3}: & m_3 & m_1 & m_2 \end{array} \qquad \begin{array}{llll} R_{m_1}: & w_3 & w_1 & w_2 \\ R_{m_2}: & w_1 & w_2 & w_3 \\ R_{m_3}: & w_2 & w_3 & w_1 \end{array} . \qquad (4.16)$$

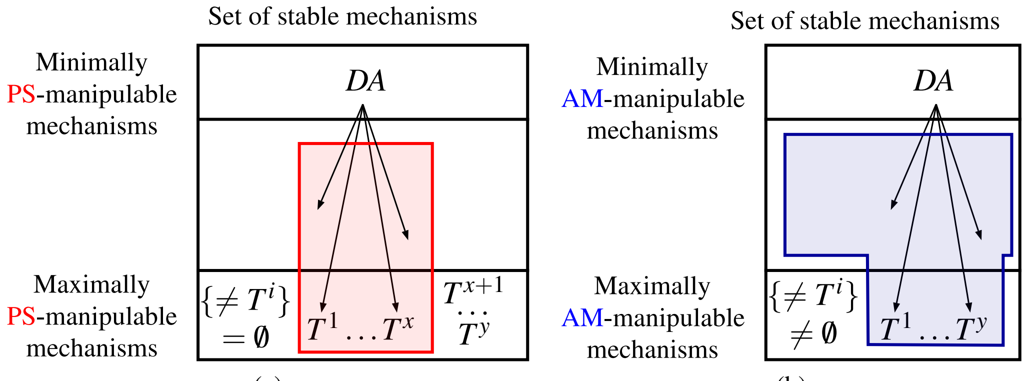
Given R, there are two stable matchings:

$$\begin{array}{c|ccc} & w_1 & w_2 & w_3 \\ \hline \mu^1: & m_1 & m_2 & m_3 \\ \mu^2: & m_2 & m_3 & m_1 \end{array} ,$$

where μ^1 is the women optimal stable matching and μ^2 the men optimal stable matching. Both stable matchings can be selected by a miniworst mechanism. DA^W (DA^M) is not less PS-manipulable than the miniworst mechanism that selects the men (women) optimal matching given the above profile.[19] ■

For the domain $\mathscr{D}^3$, the results in this section are illustrated in Figure 4.3. In the figure, the miniworst mechanisms are denoted $T^1, \ldots, T^y$.

[19] For $\#M = \#W = h > 3$, consider any profile with $R_{w_1}: m_1\ m_2 \ldots$, $R_{w_2}: m_2\ m_3 \ldots$, $\ldots$, $R_{w_h}: m_h\ m_1 \ldots$ and $R_{m_1}: w_h\ w_1 \ldots$, $R_{m_2}: w_1\ w_2 \ldots$, $\ldots$, $R_{m_h}: m_{h-1}\ m_h \ldots$. If $\#M \neq \#W$, simply let all individuals other than $\{w_1, \ldots, w_{\max\{\#M,\#W\}}, m_1, \ldots, m_{\max\{\#M,\#W\}}\}$ rank being self-matched first.



(a) (b)

Figure 4.3: Representation of the results in Section 4.6.1.

4.6.2 A conflict between median stable mechanisms and manipulability

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

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

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

results in a well-defined stable matching.

For each R, Teo and Sethuraman (1998) suggest selecting the stable matching obtained from the above procedure with l equal to one of the medians of $\{1,\ldots,k\}$. This defines the **median stable** (MS) **mechanisms**.

Like mechanisms that satisfy the miniworst criterion, *MS* mechanisms select a fair stable matching at the cost of an increase in manipulability. It is easy to find profiles for which *MS* mechanisms select a stable matching in which *not a single* individual matches with her or his most preferred achievable mate (examples include Latin Square profiles), which yields the following proposition.

Proposition 4.9. *(i) No MS mechanism is minimally PS-manipulable. (ii) When the domain is $\mathscr{D}^3$, no MS mechanism is minimally AM-manipulable.*

Because of the behavior of *MS* mechanisms on Latin Square profiles, *MS* mechanisms are also maximally AM-manipulable on $\mathscr{D}^3$. The same is not true for PS-manipuability. As much as *MS* mechanisms strive to select compromise matchings, these compromise matchings do not always make the set of individuals who match with their least preferred achievable mate minimal. That is, these mechanisms do *not* satisfy the miniworst criterion.

Proposition 4.10. *(i) When the domain is $\mathscr{D}^3$, MS mechanisms are maximally AM-manipulable. (ii) For $\#W, \#M \geq 8$, MS mechanisms are* not *maximally PS-manipulable (even on $\mathscr{D}^3$).*

For Latin Square profiles, *MS* mechanisms select stable matchings in which no individual matches with her or his most preferred achievable mate. Because any preference can be included in at least one Latin Square profile, *DA* is therefore less AM-manipulable than any *MS* mechanism. Again, this is not the case for PS-manipulability for the same reason as in Propositions 4.5 and 4.8. A further reason in the case of *MS* mechanisms is that *MS* mechanisms sometimes select stable matchings in which both a woman *and* a man match with their most preferred achievable mates, even though both individuals have multiple achievable mates (see (4.17)).

Proposition 4.11. *(i) When the domain is $\mathscr{D}^3$, DA is less AM-manipulable than any MS mechanism. (ii) No MS mechanism is more PS-manipulable than DA.*

Proof of (ii). The proof of Proposition 4.8.(ii) applies to Proposition 4.11.(ii). Proposition 4.11.(ii) can also be established by considering the following kind of profile:

$$\begin{array}{lllll} R_{w_1}: & m_2 & m_1 & m_3 & m_4 \\ R_{w_2}: & m_4 & m_2 & m_1 & m_3 \\ R_{w_3}: & m_3 & m_1 & m_2 & m_4 \\ R_{w_4}: & m_2 & m_3 & m_1 & m_4 \end{array} \qquad \begin{array}{lllll} R_{m_1}: & w_4 & w_3 & w_2 & w_1 \\ R_{m_2}: & w_3 & w_2 & w_1 & w_4 \\ R_{m_3}: & w_2 & w_1 & w_4 & w_3 \\ R_{m_4}: & w_1 & w_4 & w_3 & w_2 \end{array} . \qquad (4.17)$$

Given R, the stable matchings are

$$\begin{array}{l|llll} & w_1 & w_2 & w_3 & w_4 \\ \hline \mu^1: & m_2 & m_4 & m_1 & m_3 \\ \mu^2: & m_3 & m_4 & m_2 & m_1 \\ \mu^3: & m_4 & m_3 & m_2 & m_1 \end{array} .$$

Any MS mechanism selects matching μ^2 when the profile is R. Note that $\mu^2_{w_2} = m_4 = f^R_{w_2}$ and $\mu^2_{m_1} = w_4 = f^R_{m_1}$. Thus, for any MS mechanism, the set of individuals who cannot manipulate given profile R contains $\{w_2, m_1\}$ (Lemma 4.2). This set is contained in neither W nor M, which are the sets of individuals who cannot manipulate given R in the two variants of DA. Hence, neither variant is less PS-manipulable than MS mechanisms. ■

The results in this section are illustrated in Figure 4.4. The proofs of Propositions 4.8.(ii) and 4.11.(ii) rely on profiles in which at least one individual has exactly two achievable mates. One may conjecture that such profiles are rare and become arbitrarily unlikely as the number of individuals grows. If this is the case, DA could be less PS-manipulable than some miniworst or MS mechanisms *in the large*, i.e., for a proportion of profiles that tends to one as the number of individuals tends to infinity. Pittel et al. (2008) however conjecture that when the domain is $\bar{\mathscr{D}}$, the probability that at least one individual has exactly two achievable matches tends to *one*, rather than zero. If this conjecture is correct, then DA would fail to be less PS-manipulable than any miniworst or MS mechanism even in the

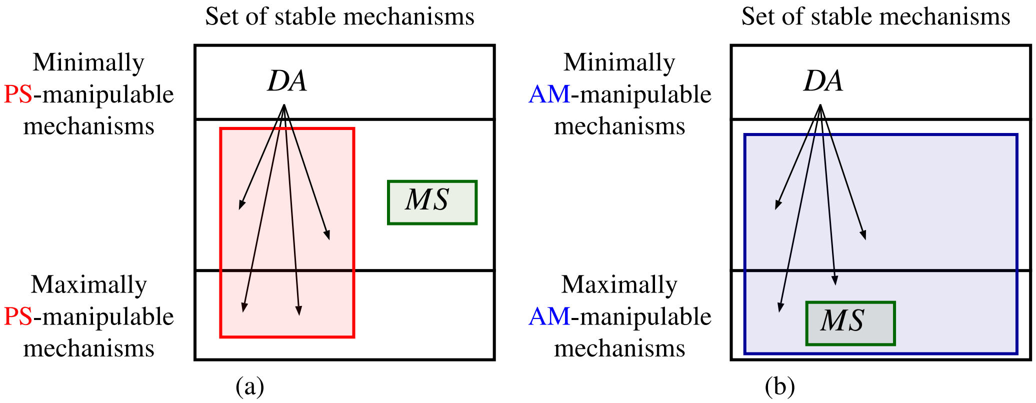

Set of stable mechanisms
Minimally PS-manipulable mechanisms
Maximally PS-manipulable mechanisms
DA
MS
(a)
Set of stable mechanisms
Minimally AM-manipulable mechanisms
Maximally AM-manipulable mechanisms
DA
MS
(b)

Figure 4.4: Representation of the results in Section 4.6.2. The set *MS* is the set of *MS* mechanisms.

large (when the domain is $\bar{\mathscr{D}}$).

4.7 Concluding remarks

I show that, in one-to-one matching, the deferred acceptance mechanism (*DA*) is minimally manipulable among the stable mechanisms in the senses of both Pathak and Sönmez (2013) (PS) and Arribillaga and Massó (2015) (AM). *DA* cannot be unambiguously improved upon in terms of manipulability without compromising with stability. In particular, I show that attempts to construct stable mechanisms that are more fair than *DA* (such as the median stable mechanisms (Teo and Sethuraman, 1998)) come at the cost of an increase in manipulability.

As I also show, providing individuals with incentives to report their preferences truthfully requires individuals to match with their *most* preferred achievable mate. This implies that some other individuals match with their *least* preferred achievable mate (Gale and Shapley, 1962). Fair stable mechanisms underperform in terms of manipulability precisely because they match few individuals with their least preferred achievable mate, and hence few individuals with their most preferred achievable mate.

In contrast, *DA* matches larger sets of individuals (the proposers) with their most preferred achievable mate, and larger sets of individuals (the acceptors) with their least preferred achievable mates. This feature of *DA* is useful when one side of the market is viewed as *weaker* than the other side (e.g., because they can spend less resources determining an appropriate strategy). Then, the weaker side can be made the proposing side in *DA*. This favors individuals on the weaker side because it provides them with simple incentives while matching them with their most preferred stable outcome.

However, because *DA* matches acceptors with their least preferred achievable mates, *DA* also provides acceptors with a larger scope for manipulation than a fair stable mechanism would. This points toward a tension between maximizing the set of individuals who have unambiguous incentives to report truthfully, and minimizing the *scope* for manipulations for individuals who do not. In this respect, fair stable mechanisms may have interesting properties: Instead of providing some individuals with unambiguous incentives to report truthfully and others with a large scope for manipulation, fair stable mechanisms tend to provide every individual with an "average" scope for manipulation.

Observe that fair stable mechanisms do not *uniformly* reduce the scope for manipulation. When compared to *DA*, fair stable mechanisms reduce the scope for manipulation for acceptors in *DA*, but increase it for proposers.[20] Also, the impact on incentives of changes in the scope for manipulation is unclear. This impact likely depends on the players' beliefs and attitudes toward risk, and how this impact plays out remains an open question.

Appendix

Proof of Lemma 4.2. In order to prove Lemma 4.2, I introduce two additional lemmas.

Lemma 4.3. *For any $i \in N$, any $R \in \mathscr{D}$, and any $R'_i \in \mathscr{D}_i$, $f_i^R \; R_i \; f_i^{(R'_i, R_{-i})}$.*

Proof. Because proposers have a truthful dominant strategy in *DA* (Dubins and Freedman,

[20] As a consequence, fair stable mechanisms cannot be compared with *DA* in terms of the "intensity of manipulation" partial order developed in Pathak and Sönmez (2013, Section IV). This last result can be viewed as an application of Chen et al. (2016, Corollary 1).

1981), $DA_i^i(R_i, R_{-i})\ R_i\ DA_i^i(R_i', R_{-i})$ for all $R_i' \in \mathscr{D}_i$. By Lemma 4.1, this is equivalent to $f_i^R\ R_i\ f_i^{(R_i', R_{-i})}$ for all $R_i' \in \mathscr{D}_i$. ■

Individual i is **single** in μ if $\mu_i = i$ and i is **married** if $\mu_i \neq i$.

Lemma 4.4 (Roth and Sotomayor, 1992)**.** *For a given $R \in \mathscr{D}$, the set of single individuals is the same in every stable matching.*

For any $i \in N$, any $R_i \in \mathscr{D}_i$, and any acceptable mate x, let $R_i|_x$ be the **truncation of** R_i **after** x, i.e., $R_i|_x$ is the preference constructed from R_i by moving i up in the ranking to the point where i is ranked right after x, but not changing any other rankings.[21]

Sufficiency. By Lemma 4.3,

$$f_i^R\ R_i\ f_i^{(R_i', R_{-i})}\ R_i\ A_i(R_i', R_{-i}) \qquad \text{for all } R_i' \in \mathscr{D}_i, \tag{4.18}$$

where the second part of (4.18) follows from A being stable and the definition of a most preferred achievable mate. Thus, (4.7) follows directly from (4.8).

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

Case 1: $A_i(R_i|_{f_i^R}, R_{-i})\ R_i\ f_i^R$. Then we have $A_i(R_i|_{f_i^R}, R_{-i})\ R_i\ f_i^R\ P_i\ A_i(R)$, contradicting (4.7).

Case 2: $f_i^R\ P_i\ A_i(R_i|_{f_i^R}, R_{-i})$. Because A is individually rational and by the construction of $R_i|_{f_i^R}$, it follows that $A_i(R_i|_{f_i^R}, R_{-i}) = i$. Also, $DA_i^i(R) = f_i^R$ by Lemma 4.1, which implies $DA_i^i(R)\ P_i\ A_i(R_i|_{f_i^R}, R_{-i})$ by the case assumption, and hence, $DA_i^i(R) \neq i$. By the construction of $R_i|_{f_i^R}$, because $DA^i(R)$ is stable when the profile is R, $DA^i(R)$ is also stable when the profile is $(R_i|_{f_i^R}, R_{-i})$. Indeed, because $DA^i(R)$ is stable when the profile is R, i is not part of

[21] Formally, (i) $x\ P_i\ i$, (ii) for all $y, z \neq i$, $y\ P_i|_x\ z$ if and only if $y\ P_i\ z$, and (iii) for all $z \notin \{i, x\}$, $z\ P_i|_x\ i$ if and only if $z\ P_i\ x$ and $i\ P_i|_x\ z$ if and only if $x\ P_i\ z$. .

a blocking pair with any mate that i ranks above $DA^i_i(R)$ according to R_i. But then because $R_i|_{f_i^R}$ and R_i have the same ranking of mates up to $DA^i_i(R)$, i is not part of a blocking pair in $DA^i(R)$ according to $R_i|_{f_i^R}$ either. Thus, when the profile is $(R_i|_{f_i^R}, R_{-i})$, there exists a stable matching in which i is married ($DA^i(R)$) and another in which i is single ($A(R_i|_{f_i^R}, R_{-i})$), contradicting Lemma 4.4. ■

Proof of Proposition 4.1. In the definition of the PS-partial order, (4.2) is equivalent to

$$\begin{aligned}&\{i \in N \mid A \text{ is } \textit{not} \text{ PS-manipulable for } i \text{ given } R\}\\ \supseteq\ &\{i \in N \mid B \text{ is } \textit{not} \text{ PS-manipulable for } i \text{ given } R\}.\end{aligned}$$

Similarly, in the definition of the AM-partial order, (4.5) is equivalent to

$$\begin{aligned}&\{i \in N_{R_*} \mid A \text{ is } \textit{not} \text{ manipulable for } i \text{ given } R_*\}\\ \supseteq\ &\{i \in N_{R_*} \mid B \text{ is } \textit{not} \text{ manipulable for } i \text{ given } R_*\}.\end{aligned}$$

The proposition then follows directly from Lemma 4.2. ■

Proof of Proposition 4.2. I provide a proof for minimal PS-manipulability. The proof for maximal PS-manipulability is analogous.

Necessity. In order to derive a contradiction, suppose that some stable mechanism B is less manipulable than A. By Proposition 4.1, this implies that for some $R^* \in \mathscr{D}$,

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

But because B is stable, $B_i(R^*)$ is stable with respect to R^*. Thus, $B_i(R^*)$ and R^* satisfy (4.11), a contradiction.

Sufficiency. In order to derive a contradiction, assume that μ^* is a stable matching satisfying (4.11) for some profile R^*. Consider mechanism B constructed from A by setting

$B(R) = A(R)$ for all $R \in \mathscr{D}$ with $R \neq R^*$ and $B(R^*) = \mu^*$. Clearly, for all $R \in \mathscr{D}$ with $R \neq R^*$,

$$\{i \in N \mid A_i(R) = f_i^R\} = \{i \in N \mid B_i(R) = f_i^R\}. \tag{4.19}$$

Also, by (4.11) and because $B_i(R^*) = \mu_i^*$,

$$\{i \in N \mid A_i(R^*) = f_i^{R^*}\} \subset \{i \in N \mid B_i(R^*) = f_i^{R^*}\}. \tag{4.20}$$

By Proposition 4.1, (4.19) and (4.20) imply that B is no more PS-manipulable than A, but that the converse is not true. Hence, by definition, B is less PS-manipulable than A and so A is not minimally PS-manipulable, a contradiction. ■

Proof of Proposition 4.3.(i). I provide a proof for DA^W and minimal manipulability only. The proofs for DA^M and maximal manipulability are analogous.

PS-partial order. By Proposition 4.2, the PS part of (i) holds provided that there does not exist a profile R^* and a stable matching μ^* such that

$$\left\{i \in N \mid DA_i^W(R^*) = f_i^{R^*}\right\} \subset \left\{i \in N \mid \mu_i^* = f_i^{R^*}\right\}. \tag{4.21}$$

In order to derive a contradiction, suppose that there exists a stable matching μ^* and a preference profile R^* satisfying (4.21). By Lemma 4.1,

$$W \subseteq \left\{i \in N \mid DA_i^W(R^*) = f_i^{R^*}\right\}. \tag{4.22}$$

Together, (4.21) and (4.22) imply $W \subset \left\{i \in N \mid \mu_i^* = f_i^{R^*}\right\}$. But (4.21) implies $\mu^* \neq DA^W(R^*)$, which in turn implies that there exists a woman $w^* \in W$ for whom $\mu_{w^*}^* \neq DA_{w^*}^W(R^*) = f_{w^*}^{R^*}$. Hence $W \not\subset \{i \in N \mid \mu_i^* = f_i^{R^*}\}$, a contradiction.

AM-partial order. In order to derive a contradiction, suppose that some stable mechanism A is less AM-manipulable than DA^W. By Lemmas 4.1 and 4.2, for all $R_* \in \cup_{i \in W} \mathscr{D}_i$,

$$\{w \in N_{R_*} \mid w \text{ can manipulate } DA^W \text{ given } R_*\} = \emptyset. \tag{4.23}$$

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

$$\{w \in N_{R_*} \mid w \text{ can manipulate } A \text{ given } R_*\} = \emptyset. \tag{4.24}$$

But by Lemma 4.2, (4.23) and (4.24) imply that for all $R \in \mathscr{D}$ and for all $w \in W$,

$$A_w(R) = DA_w^W(R) = f_w^R. \tag{4.25}$$

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

Proof of Proposition 4.4. Consider any $i \in N$ and any $R_i \in \bar{\mathscr{D}}_i$. Let $i = w_1$ without loss of generality. Without loss of generality again, let the individuals in M be labeled in such a way that

$$R_{w_1}: \quad m_h \quad m_{(h-1)} \quad \dots \quad m_2 \quad m_1 \tag{4.26}$$

A Latin Square profile $(R_i, R_{-i}^{R_i})$ generalizing profile (4.13) can then be constructed in the following way:

$$\begin{array}{cccccccc}
R_{w_1}: & m_h & m_{h-1} & & \cdots & & m_2 & \textcolor{red}{m_1} \\
R^{R_i}_{w_2}: & m_{h-1} & & & & & & m_h \\
 & & & & m_2 & \iddots & & \\
\vdots & \vdots & & m_2 & \textcolor{red}{m_1} & m_h & & \vdots \\
 & & & \iddots & m_h & & & \\
R^{R_i}_{w_{h-1}}: & m_2 & & & & & & m_3 \\
R^{R_i}_{w_h}: & \textcolor{red}{m_1} & m_h & & \cdots & & m_3 & m_2
\end{array} \tag{4.27}$$

The preferences of the men in $(R_i, R^{R_i}_{-i})$ are constructed symmetrically to (4.27) with woman w_1 appearing on the downward diagonal as in (4.13).

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

(ii). By Lemmas 4.1 and 4.2, for any stable mechanism A, if there exists no $R_* \in \cup_{i\in N} \bar{\mathscr{D}}_i$ and no acceptor $a \in N_{R_*}$ such that $A_a(R_*, R_{-a}) = f_a^{(R_*, R_{-a})}$ for all $R_{-a} \in \bar{\mathscr{D}}_{-a}$, then either $A = DA$ or A is more AM-manipulable than DA. We are interested in the proportion of these mechanisms relative to the set of stable mechanisms.

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

$$\begin{aligned} 1 - \mathbb{P}\big(&\cup_{\{i \in N | i \text{ is an acceptor}\}} \cup_{k \in \{1,\dots,h!\}} \\ &[A_i(R_i^k, R_{-i}) = f_i^{(R_i^k, R_{-i})} \text{ for all } R_{-i} \in \bar{\mathscr{D}}_{-i}]\big). \end{aligned} \tag{4.28}$$

The expression in (4.28) is at least

$$1 - \sum_{\{i \in N | i \text{ is an acceptor}\}} \sum_{k \in \{1,\dots,h!\}} \mathbb{P}\big(A_i(R_i^k, R_{-i}) = f_i^{(R_i^k, R_{-i})} \text{ for all } R_{-i} \in \bar{\mathscr{D}}_{-i}\big). \tag{4.29}$$

We can obtain a bound on (4.29) by bounding the term inside the double summation. For any profile $R \in \bar{\mathscr{D}}$, let $\sigma^R(X)$ denote the proportion of stable matchings μ for which X is true. Observe that

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

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

$$\mathbb{P}\big(A_i(R_i^k, R_{-i}) = f_i^{(R_i^k, R_{-i})} \text{ for all } R_{-i} \in \bar{\mathscr{D}}_{-i}\big) \leq \frac{1}{h}. \tag{4.30}$$

A tighter bound for (4.29) can be obtained by tightening the bound in (4.30). This can be done by considering profiles different from the Latin Square profile (4.27). Specifically, I consider variations of (4.27) for which (a) the number of stable matchings and (b) the proportion of stable matchings that match i with her or his most preferred achievable mate are easy to compute.

In what follows, I use the relabeling introduced at the beginning of the proof of Proposition 4.4, with $R_i^k = R_{w_1}$. The first variation of the Latin Square profile that is considered has

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

$$\begin{array}{rcccccccc}
R_{w_1}: & m_h & m_{h-1} & m_{h-2} & & \dots & & m_2 & m_1 \\
R_{w_2}^{R_{w_1}}(h-1,1): & m_h & m_{h-2} & & & & & & m_{h-1} \\
 & & & & & m_2 & \iddots & & \\
\vdots & \vdots & & & m_2 & m_1 & m_{h-1} & \vdots & \vdots \\
 & & & & \iddots & m_{h-1} & & & \\
R_{w_{h-2}}^{R_{w_1}}(h-1,1): & m_h & m_2 & & & & & & m_3 \\
R_{w_{h-1}}^{R_{w_1}}(h-1,1): & m_h & m_1 & m_{h-1} & & \dots & & m_3 & m_2 \\
R_{w_h}^{R_{w_1}}(h-1,1): & m_h & & & & & & & \\
R_{m_h}^{R_{w_1}}(h-1,1): & w_h & & & & & & &
\end{array} \tag{4.31}$$

In (4.31), every woman ranks m_h first. Among the first $h-1$ women, the sub-profile excluding m_h has a Latin Square structure of dimension $h-1$ similar to (4.27). For w_h and m_h, only the most preferred mate is specified.

In $(R_{w_1}, R_{-w_1}^{R_{w_1}}(h-1,1))$, the preferences of men other than m_h are constructed symmetrically to the preferences of the women other than w_h in (4.31) with w_h ranked last and woman w_1 appearing on the downward diagonal as in (4.13).

Observe that m_h and w_h match together in every stable matching given $(R_{w_1}, R_{-w_1}^{R_{w_1}}(h-1,1))$, and m_h and w_h are therefore not achievable for any other man or woman. By analogy with (4.27), there are $(h-1)$ stable matching among the remaining individuals $\{m_1, \dots, m_{h-1}, w_1, \dots, w_{h-1}\}$ due to the Latin Square structure of the profile once m_h and w_h are removed. There are therefore $(h-1)$ stable matchings given $(R_{w_1}, R_{-w_1}^{R_{w_1}}(h-1,1))$, only one of which matches w_1 with her most preferred achievable mate.

A natural variant of (4.31), denoted $(R_{w_1}, R^{R_{w_1}}_{-w_1}(h-1,2))$, also has $(h-1)$ stable matchings. In $(R_{w_1}, R^{R_{w_1}}_{-w_1}(h-1,2))$, the preferences of the women and of man m_1 are as follows:

$$
\begin{array}{rcccccccc}
R_{w_1}: & m_h & m_{h-1} & & \dots & & m_3 & m_2 & m_1 \\
R^{R_{w_1}}_{w_2}(h-1,2): & m_{h-1} & & & & & & m_h & m_1 \\
 & & & & m_3 & \iddots & & & \\
\vdots & \vdots & & m_3 & m_2 & m_h & & \vdots & \vdots \\
 & & & \iddots & m_h & & & & \\
R^{R_{w_1}}_{w_{h-2}}(h-1,2): & m_3 & & & & & & m_4 & m_1 \\
R^{R_{w_1}}_{w_{h-1}}(h-1,2): & m_2 & m_h & & \dots & & m_4 & m_3 & m_1 \\
R^{R_{w_1}}_{w_h}(h-1,2): & m_1 & & & & & & & \\
R^{R_{w_1}}_{m_1}(h-1,2): & w_h & & & & & & &
\end{array}
\tag{4.32}
$$

In (4.32), the first $h-1$ women rank m_1 last. Among the first $h-1$ women, the sub-profile excluding m_1 has a Latin Square structure of dimension $h-1$ similar to (4.27). For w_h and m_1, only the most preferred mate is specified.

In $(R_{w_1}, R^{R_{w_1}}_{-w_1}(h-1,2))$, the preferences of men other than m_1 are constructed symmetrically to the preferences of the women other than w_h in (4.31) with w_h ranked last and woman w_1 appearing on the downward diagonal as in (4.13).

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

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

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

There are $h((h-1)!)^2$ ways to select stable matchings for these $1+2(h-1)$ profiles. For example, a stable mechanism could select the *first* of the h stable matchings in the full Latin Square profile, the *first* of the $(h-1)$ stable matchings in $(R_{w_1}, R^{R_{w_1}}_{-w_1}(h-1,1))$, the *first* of the $(h-2)$ stable matchings in $(R_{w_1}, R^{R_{w_1}}_{-w_1}(h-2,1))$, and so on. A stable mechanism could also select the *second* of the h stable matchings in the full Latin Square profile, the *first* of the $(h-1)$ stable matchings in $(R_{w_1}, R^{R_{w_1}}_{-w_1}(h-1,1))$, the *first* of the $(h-2)$ stable matchings in $(R_{w_1}, R^{R_{w_1}}_{-w_1}(h-2,1))$, and so on. Of the $h((h-1)!)^2$ possible selections for these $1+2(h-1)$ profiles, only one always matches i with her or his most preferred achievable mate. Hence,

$$\begin{aligned}\frac{1}{h((h-1)!)^2} &= \prod_{R_{-i}\in\{R^1_{-i},\ldots,R^{1+2(h-1)}_{-i}\}} \sigma^{(R^k_i,R_{-i})}\big(\mu_i = f_i^{(R^k_i,R_{-i})}\big)\\ &\geq \prod_{R_{-i}\in\bar{\mathscr{D}}_{-i}} \sigma^{(R^k_i,R_{-i})}\big(\mu_i = f_i^{(R^k_i,R_{-i})}\big).\end{aligned} \tag{4.33}$$

Using (4.33) in (4.29) shows that (4.28) is at least

$$1 - \sum_{\{i\in N \mid i \text{ is acceptor}\}} \sum_{k\in\{1,\ldots,h!\}} \frac{1}{h((h-1)!)^2}. \tag{4.34}$$

Because the fraction in (4.34) is independent of the indices used in the summations, (4.34) is equal to $1-\frac{h(h!)}{h((h-1)!)^2} = 1-\frac{h}{(h-1)!}$.

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

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

Proof of Proposition 4.5. **(i)**. As shown in the proof of Proposition 4.4.(ii), the proportion of stable mechanisms that are more AM-manipulable than DA is at least $1-\left(\frac{h}{(h-1)!}+\frac{1}{h((h-1)!)^2}\right)$. Conversely, DA is less manipulable than at least $1-\left(\frac{h}{(h-1)!}+\frac{1}{h((h-1)!)^2}\right)$.

(ii). If DA is less PS-manipulable than stable mechanism A, then A can never select the optimal stable matching of the accepting side whenever any acceptor has more than one achievable mate. This implies that for any acceptor a and any of the $h!$ preferences $R_a \in \bar{\mathscr{D}}_a$, we have $A_a(R_a,R_{-a}^{R_a}) \neq f_a^{(R_a,R_{-a}^{R_a})}$, where the construction of $R_{-a}^{R_a}$ is described in (4.27). Each $(R_a,R_{-a}^{R_a})$ has h stable matchings only one of which matches a with $f_a^{(R_a,R_{-a}^{R_a})}$. Hence, for each $(R_a,R_{-a}^{R_a})$, there are $h-1$ ways to select a stable matching $A(R_a,R_{-a}^{R_a})$ with $A_a(R_a,R_{-a}^{R_a}) \neq f_a^{(R_a,R_{-a}^{R_a})}$. This implies that, of all the $h^{h!}$ possible ways in which a stable mechanism A can select a stable matching for the $h!$ profiles $(R_a,R_{-a}^{R_a})$, only $(h-1)^{h!}$ make it possible for DA to be less PS-manipulable than A. Therefore, at most $\left(\frac{h-1}{h}\right)^{h!}$ of the stable mechanisms A are more manipulable than DA.

Proof of Proposition 4.6. **(i)**. Consider any Latin Square profile R^{LS} as defined in (4.27) (if $\#W \neq \#M$, see the argument in footnote 11). For any miniworst mechanism A, $A_i(R^{LS}) \neq f_i^{R^{LS}}$ for all $i \in N$. The mechanism B constructed from A by changing the stable matching selected for R^{LS} to the men optimal or women optimal stable matching (and selecting the same matching as A otherwise) is less PS-manipulable than A. Hence A is not minimally PS-manipulable.

(ii). See the proof of Proposition 4.8. ■

Proof of Proposition 4.7. **(i)**. **Sufficiency**. By Proposition 4.2, it is sufficient to show that for any miniworst mechanism, any profile R, and any stable matching μ, (4.12) does not hold. Because A is miniworst, (4.15) is false. That is, either

$$\{i \in N \mid A_i(R) = l_i^R\} = \{i \in N \mid \mu_i = l_i^R\}, \tag{4.35}$$

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

$$\mu_{i^*} = l_{i^*}^R \text{ and } A_{i^*}(R) \neq l_{i^*}^R. \tag{4.36}$$

By Lemma 4.1, individuals match with their least preferred achievable mates if and only if they are their mates' most preferred achievable mate. Thus, if (4.35) holds,

$$\{i \in N \mid A_i(R) = f_i^R\} = \{i \in N \mid \mu_i = f_i^R\}. \tag{4.37}$$

On the other hand, if (4.36) holds, we have

$$A_{l_{i^*}^R}(R) \neq f_{l_{i^*}^R}^R = i^* \text{ and } \mu_{l_{i^*}^R} = f_{l_{i^*}^R}^R = i^*. \tag{4.38}$$

If (4.37) holds, then the set of individuals who match with their most preferred achievable mate is the same in $A(R)$ and μ. On the other hand, if (4.38) holds, then there is an individual $l_{i^*}^R$ who matches with $f_{l_{i^*}^R}^R$ in μ, but not in $A(R)$. In both cases, the set of individuals who match with their most preferred achievable mates in $A(R)$ is not a superset of the set of individuals who match with their most preferred achievable mates in μ. That is,

$$\{i \in N \mid A_i(R) = f_i^R\} \not\supset \{i \in N \mid \mu_i = f_i^R\}. \tag{4.39}$$

and (4.12) does not hold.

Necessity. In order to derive a contradiction, assume that A is maximally PS-manipulable but A is not miniworst, i.e., there exists a profile R^* and a matching μ^* such that (4.15) holds. By Lemma 4.1, (4.15) implies

$$\{i \in N \mid A_i(R^*) = f_i^{R^*}\} \supset \{i \in N \mid \mu_i^* = f_i^{R^*}\}. \tag{4.40}$$

Now, construct mechanism B from A by setting $B(R) = A(R)$ for all $R \in \mathscr{D}$ with $R \neq R^*$, and $B(R^*) = \mu^*$. By Lemma 4.2, because $B(R) = A(R)$ for all $R \neq R^*$, we have that for all $R \neq R^*$,

$$\begin{aligned} & \{i \in N \mid i \text{ can manipulate } A \text{ given } R\} \\ = \; & \{i \in N \mid i \text{ can manipulate } B \text{ given } R\}. \end{aligned} \tag{4.41}$$

Also, by (4.40) and Lemma 4.2,

$$\begin{aligned} & \{i \in N \mid i \text{ can manipulate } A \text{ given } R^*\} \\ \subset \; & \{i \in N \mid i \text{ can manipulate } B \text{ given } R^*\}. \end{aligned} \tag{4.42}$$

Together, (4.41) and (4.42) imply that A is less PS-manipulable than B and therefore A is not maximally PS-manipulable, a contradiction.

(ii). For any $i \in N$ and any $R_i \in \mathscr{D}_i^3$, it is possible to construct a Latin Square profile similar to (4.27) among the mates that are acceptable according to R_i. Slightly abusing the notation, this profile is also denoted $(R_i, R_{-i}^{R_i})$. For example, if three mates are acceptable according to R_i, let $N^6 \subseteq N$ consist of (a) i, (b) i's three acceptable mates, and (c) two more individuals on i's side of the market. Profile $(R_i, R_{-i}^{R_i})$ then has the same structure as (4.27) among the individuals in N^6. (For any $i \in N^6$, any individual on the other side of the market that does not belong to N^6 is unacceptable).

If mechanism A is miniworst, then for any $i \in N$ and any $R_i \in \mathscr{D}_i^3$, $A_i(R_i, R_{-i}^{R_i}) \neq f_i^{(R_i, R_{-i}^{R_i})}$. Hence, for any $R_* \in \cup_{i \in N} \mathscr{D}_i^3$,

$$\{i \in N_{R_*} \mid i \text{ can manipulate } A \text{ given } R_*\} = N$$

and A is clearly maximally AM-manipulable.

(iii). For any $i \in N$ and any $R_i \in \mathscr{D}_i^3$, it is possible to construct a Latin Square profile similar to (4.27) among the mates that are acceptable according to R_i. Slightly abusing the notation and terminology, this profile is also denoted $(R_i, R_{-i}^{R_i})$ and called a Latin Square profile. Consider any mechanism A such that

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

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

By (a) and Lemma 4.2, for any $R_* \in \cup_{i \in N} \mathscr{D}_i^3$,

$$\{i \in N_{R_*} \mid i \text{ can manipulate } A \text{ given } R_*\} = N,$$

and A clearly is maximally AM-manipulable. However, for many *non* Latin Square profiles R^*, there exists a stable matching μ such that (4.14) holds. This is the case, for example, for the profile presented in the proof of Proposition 4.10. For this profile, DA selects either μ^9 or μ^1 (depending on the variant of DA that is used). Hence, by construction, A also selects either μ^9 or μ^1 although μ^4 satisfies (4.14). Thus, A is maximally AM-manipulable but not miniworst. ■

Proof of Proposition 4.8. As shown in the proof of Proposition 4.7.(ii), if a mechanism A is miniworst, then for any $R_* \in \cup_{i \in N} \mathscr{D}_i^3$,

$$\{i \in N_{R_*} \mid i \text{ can manipulate } A \text{ given } R_*\} = N.$$

Hence, DA is less AM-manipulable than A. ■

Proof of Proposition 4.9. **(i)**. The proof is for $\#W = \#M$. For the case $\#W \neq \#M$, see the argument in footnote 11. Consider any Latin Square profile R^{LS} as defined in (4.27). Given R^{LS}, MS mechanisms select a stable matching in which no individual matches with her or his most preferred achievable mate. The mechanism A constructed from MS mechanisms by changing the stable matching selected for R^{LS} to the men optimal or women optimal stable matching (and selecting the same matching as MS mechanisms otherwise) is less PS-manipulable than MS mechanisms. Hence, MS mechanisms are not minimally PS-manipulable.

(ii). For any $i \in N$ and any $R_i \in \mathscr{D}_i^3$, it is possible to construct a Latin Square profile similar to (4.27) among the mates that are acceptable according to R_i. Slightly abusing the notation and terminology, this profile is also denoted $(R_i, R_{-i}^{R_i})$ and called a Latin Square profile. In $(R_i, R_{-i}^{R_i})$, MS mechanisms select a stable mechanism in which i does not match with her or his most preferred achievable mate. That is, for any $i \in N$ any $R_i \in \mathscr{D}_i^3$, and any MS mechanism MSM, $MSM_i(R_i, R_{-i}^{R_i}) \neq f_i^{(R_i, R_{-i}^{R_i})}$. Thus, by Lemma 4.2, for any $R_* \in \cup_{i \in N} \mathscr{D}_i^3$,

$$\{i \in N_{R_*} \mid i \text{ can manipulate } MSM \text{ given } R_*\} = N, \tag{4.43}$$

and MS mechanisms are clearly not minimally AM-manipulable. ■

Proof of Proposition 4.10. **(i)**. By (4.43) in the proof of Proposition 4.9.(ii) and Lemma

4.1, for any $R_* \in \cup_{i \in N} \mathscr{D}_i^3$ and any MS mechanism MSM,

$$\begin{aligned} & \{i \in N_{R_*} \mid i \text{ can manipulate } MSM \text{ given } R_*\} = N \\ \supset\ & \{i \in N_{R_*} \mid i \text{ can manipulate } DA \text{ given } R_*\}. \end{aligned}$$

(ii). By Proposition 4.7, it is sufficient to show that *MS* mechanisms are not miniworst on the set of stable matchings. Let $N^8 := \{w_1, \ldots, w_8, m_1, \ldots, m_8\}$ and consider any profile R including the following sub-profile for individuals in N^8:

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

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

Observe that because no two women have the same most preferred man, matching every woman with her favorite man yields a stable matching. Thus, because the set of individuals who are married is the same in every stable matching (Lemma 4.4), every individual in N^8 is married in every stable matching. Therefore, because no man in $\{m_1, m_2, m_3\}$ is acceptable to any woman in $\{w_4, w_5, w_6, w_7, w_8\}$, but every man in $\{m_1, m_2, m_3\}$ is acceptable to every woman in $\{w_1, w_2, w_3\}$, all men in $\{m_1, m_2, m_3\}$ must match with women in $\{w_1, w_2, w_3\}$ in any stable matching (by definition, stable matchings are individually rational). As a consequence, all men in $\{m_4, m_5, m_6, m_7, m_8\}$ also match with women in $\{w_4, w_5, w_6, w_7, w_8\}$ in every stable matching.

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

$$\begin{array}{l|lll}
 & w_1 & w_2 & w_3 \\
\hline
\mu^1_{123}: & m_1 & m_3 & m_2 \\
\mu^2_{123}: & m_2 & m_1 & m_3 \\
\mu^3_{123}: & m_3 & m_2 & m_1
\end{array}$$

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

	w_4	w_5	w_6	w_7	w_8
$\mu^1_{45678}:$	m_4	m_8	m_7	m_6	m_5
$\mu^2_{45678}:$	m_5	m_4	m_8	m_7	m_6
$\mu^3_{45678}:$	m_6	m_5	m_4	m_8	m_7
$\mu^4_{45678}:$	m_7	m_6	m_5	m_4	m_8
$\mu^5_{45678}:$	m_8	m_7	m_6	m_5	m_4

Observe that in any stable matching that includes μ^1_{123} or μ^2_{123}, the sub-matching among $\{m_4, m_5, m_6, m_7, m_8\} \cup \{w_4, w_5, w_6, w_7, w_8\}$ must be either μ^1_{45678} or μ^2_{45678}. Indeed, in any other combination that includes μ^1_{123} or μ^2_{123} (e.g., $(\mu^1_{123}, \mu^4_{45678})$), every man in $\{m_4, m_5, m_6, m_7, m_8\}$ forms a blocking pair with every woman in $\{w_1, w_2, w_3\}$. On the other hand, in stable matchings that includes μ^3_{123}, the sub-matching among $\{m_4, m_5, m_6, m_7, m_8\} \cup \{w_4, w_5, w_6, w_7, w_8\}$ can be any of the stable sub-matchings $(\mu^1_{45678}, \ldots, \mu^5_{45678})$. Overall, the stable matchings among individuals in N^8 are

$$\mu^1 := (\mu^1_{123}, \mu^1_{45678})$$

$$\mu^2 := (\mu^1_{123}, \mu^2_{45678})$$

$$\mu^3 := (\mu^2_{123}, \mu^1_{45678})$$

$$\mu^4 := (\mu^2_{123}, \mu^2_{45678})$$

$$\mu^5 := (\mu^3_{123}, \mu^1_{45678})$$

$$\mu^6 := (\mu^3_{123}, \mu^2_{45678})$$

$$\mu^7 := (\mu^3_{123}, \mu^3_{45678})$$

$$\mu^8 := (\mu^3_{123}, \mu^4_{45678})$$

$$\mu^9 := (\mu^3_{123}, \mu^5_{45678})$$

For every woman $w \in \{w_1, w_2, w_3\}$, the stable matchings are ranked in the same way with

$$\mu^9 \; R_w \; \mu^8 \; R_w \; \mu^7 \; R_w \; \mu^6 \; R_w \; \mu^5 \; R_w \; \mu^4 \; R_w \; \mu^3 \; R_w \; \mu^2 \; R_w \; \mu^1.$$

Hence, given profile R, MS mechanisms match every woman in $\{w_1, w_2, w_3\}$ with her match in μ^5.

For every woman $w \in \{w_4, w_5, w_6, w_7, w_8\}$, the stable matchings are also ranked in the same way with

$$\mu^9 \; R_w \; \mu^8 \; R_w \; \mu^7 \; R_w \; \mu^6 \; R_w \; \mu^4 \; R_w \; \mu^2 \; R_w \; \mu^5 \; R_w \; \mu^3 \; R_w \; \mu^1.$$

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

Thus, given profile R, MS mechanisms match individuals in N^8 according to matching μ^6. In μ^6, the set of $i \in N^8$ who match with l_i^R is $\{m_1, m_2, m_3\}$. But note that in μ^4 the set of $i \in N^8$ who match with l_i^R is empty. Hence, MS mechanisms are not miniworst on the set of stable matchings. ■

Proof of Proposition 4.11.(i). See (4.43) in the proof of Proposition 4.9.(ii). ■

BIBLIOGRAPHY

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

Alpern, S., Gal, S., 2009. Analysis and design of selection committees: a game theoretic secretary problem. International Journal of Game Theory 38, 377–394.

Alpern, S., Gal, S., Solan, E., 2010. A sequential selection game with vetoes. Games and Economic Behavior 68, 1–14.

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

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

Arribillaga, R.P., Massó, J., 2017. Comparing voting by committees according to their manipulability. American Economic Journal: Microeconomics, forthcoming.

Barberà, S., Massó, J., Neme, A., 2005. Voting by committees under constraints. Journal of Economic Theory 122, 185–205.

Barberà, S., Peleg, B., 1990. Strategy-proof voting schemes with continuous preferences. Social Choice and Welfare 7, 31–38.

Barberà, S., Sonnenschein, H., Zhou, L., 1991. Voting by committees. Econometrica 59, 595–609.

Basteck, C., Mantovani, M., 2016a. Cognitive ability and games of school choice. University of Milan Bicocca Department of Economics, Management and Statistics Working Paper No. 343.

Basteck, C., Mantovani, M., 2016b. Protecting unsophisticated applicants in school choice through information disclosure. UNU-WIDER Research Paper No. 65.

Benoît, J.P., 2002. Strategic manipulation in voting games when lotteries and ties are permitted. Journal of Economic Theory 102, 421–436.

Bermant, G., 1982. Jury Selection Procedures in United States District Courts. volume 25. Federal Judicial Center, Washington D.C.

Bermant, G., Shapard, J., 1981. The voir dire examination, juror challenges, and adversary advocacy, in: Sales, B.D. (Ed.), The Trial Process. Springer, Berlin, pp. 69–114.

Bloom, D., Cavanagh, C.L., 1986. An analysis of the selection of arbitrators. American Economic Review 76, 408–422.

Brams, S.J., Davis, M.D., 1978. Optimal jury selection: A game-theoretic model for the exercise of peremptory challenges. Operations Research 26, 966–991.

Caditz, D.M., 2015. Selection under veto with limited foresight. Working Paper .

Chatterji, S., Roy, S., Sen, A., 2012. The structure of strategy-proof random social choice functions over product domains and lexicographically separable preferences. Journal of Mathematical Economics 48, 353–366.

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

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

Crawford, V.P., Costa-Gomes, M.A., Iriberri, N., 2013. Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. Journal of Economic Literature 51, 5–62.

Daly, M., 2016. Foster v. Chatman: Clarifying the Batson test for discriminatory peremptory strikes. Duke Journal of Constitutional Law and Public Policy Sidebar 11, 148–162.

de Clippel, G., Eliaz, K., Knight, B., 2014. On the selection of arbitrators. American Economic Review 104, 3434–3458.

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

DeGroot, M.H., Kadane, J.B., 1980. Optimal challenges for selection. Operations Research 28, 952–968.

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

Dur, U., Hammond, R.G., Morrill, T., 2017. Identifying the Harm of Manipulable School-Choice Mechanisms. American Economic Journal: Economic Policy .

Dutta, B., Peters, H., Sen, A., 2006. Strategy-proof cardinal decision schemes. Social Choice and Welfare 28, 163–179.

Flanagan, F.X., 2015. Peremptory challenges and jury selection. Journal of Law and Economics 58, 385–416.

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

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

Gerber, A., Barberà, S., 2016. Sequential voting and agenda manipulation. Theoretical Economics 12, 211–247.

Gibbard, A., 1973. Manipulation of voting schemes: A general result. Econometrica , 587–601.

Henley, P., 1996. Improving the jury system: Peremptory challenges. San Francisco: Hastings College of Law, Public Law Research Institute Report.

Ho, T.H., Camerer, C., Weigelt, K., 1998. Iterated dominance and iterated best response in experimental" p-beauty contests". American Economic Review , 947–969.

Hylland, A., 1980. Strategy proofness of voting procedures with lotteries as outcomes and infinite sets of strategies. Unpublished manuscript, Department of Economics, University of Oslo .

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

Ju, B.G., 2003. A characterization of strategy-proof voting rules for separable weak orderings. Social Choice and Welfare 21, 469–499.

Kadane, J.B., Stone, C.A., Wallstrom, G., 1999. The donation paradox for peremptory challenges. Theory and Decision 47, 139–155.

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

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

LaFave, W., Israel, J., King, N., Kerr, O., 2009. Criminal procedure. West Academic Publishing, St. Paul, MN. 5 edition edition.

Le Breton, M., Weymark, J.A., 1999. Strategy-proof social choice with continuous separable preferences. Journal of Mathematical Economics 32, 47–85.

Marder, N.S., 2012. Batson revisited - Batson symposium. SSRN Scholarly Paper ID 2165561. Social Science Research Network. Rochester, NY.

Maskin, E., 1999. Nash equilibrium and welfare optimality. Review of Economic Studies 66, 23–38.

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

Moulin, H., 1981. Prudence versus sophistication in voting strategy. Journal of Economic Theory 24, 398–412.

Mueller, D.C., 1978. Voting by veto. Journal of Public Economics 10, 57–75.

Nandeibam, S., 2012. The structure of decision schemes with cardinal preferences. Review of Economic Design 17, 205–238.

O'Malley, J.W., 2015. Catholic History for Today's Church: How our Past Illuminates our Present. Rowman and Littlefield, Lanham, MD.

Pathak, P.A., Sönmez, T., 2008. Leveling the playing field : Sincere and sophisticated players in the boston mechanism. American Economic Review 98, 1636–1652.

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

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

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

Roth, A., Kadane, J.B., Degroot, M.H., 1977. Optimal Peremptory Challenges in Trials by Juries: A Bilateral Sequential Process. Operations Research 25, 901–919.

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

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

Rudin, W., 1976. Principles of Mathematical Analysis. McGraw-Hill, New York.

Schummer, J., 1999. Strategy-proofness versus efficiency for small domains of preferences over public goods. Economic Theory 13, 709–722.

Sen, A., 1970. The impossibility of a Paretian liberal. Journal of Political Economy 78, 152–157.

Sen, A., 1999. The possibility of social choice. American Economic Review 89, 349–378.

Shapard, J., Johnson, M., 1994. Memorandom on a survey of active jurdges regarding their voir dire practices. Technical Report. Federal Judicial Center, Research Division.

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

Van der Linden, M., 2015. Impossibilities for strategy-proof veto in selection problems. Working Paper .

Van der Linden, M., 2017. Bounded rationality and the choice of jury selection procedures. SSRN Working Paper No. 2922790.

Yuval, F., 2002. Sophisticated voting under the sequential voting by veto. Theory and Decision 53, 343–369.

Cases

Batson v. Kentucky, 476 U.S. 79 (1986).

Swain v. Alabama, 380 U.S. 202 (1965).

Court rules

Minnesota Court Rules, Criminal Procedure, Rule 26.02, Subd.4.(3)b).

Tennessee Court Rules, Rules of Civil Procedure, Rule 47.03.