# Approaches to Mechanism Design with Boundedly Rational Agents

by

Gabriel D. Carroll

A.B., Mathematics and Linguistics, Harvard College, 2005

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Economics
at the
Massachusetts Institute of Technology
June 2012





Author .....
Department of Economics
May 15, 2012

Certified by .....
Parag A. Pathak
Associate Professor of Economics
Thesis Supervisor

Certified by .....
K. Daron Acemoglu
Elizabeth and James Killian Professor of Economics
Thesis Supervisor

Accepted by .....
Michael Greenstone
3M Professor of Environmental Economics
Chair, Departmental Committee on Graduate Studies

# Approaches to Mechanism Design with Boundedly Rational Agents

by

Gabriel D. Carroll

Submitted to the Department of Economics
on May 15, 2012, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

## Abstract

This dissertation ties together three papers on mechanism design with boundedly rational agents. These papers explore theoretically whether, and to what extent, limitations on agents' ability to strategically misrepresent their preferences can help a mechanism designer achieve outcomes that she could not achieve with perfectly rational agents.

The first chapter investigates whether local incentive constraints are sufficient to logically imply full incentive-compatibility, in a variety of mechanism design settings. This can be motivated by a boundedly rational model in which agents cannot contemplate all possible misrepresentations, but can consider those that are close to their true preferences. This chapter offers a unified approach that covers both continuous and discrete type spaces, showing that in many commonly studied cases, local incentive-compatibility (suitably defined) implies full incentive-compatibility.

The second chapter advances the methodology of looking quantitatively at incentives for strategic behavior, motivated by the premise that agents will be truthful if the incentive to be strategic is small enough. This chapter defines a mechanism's *susceptibility to manipulation* as the maximum amount of expected utility any agent can ever gain from strategic misrepresntation. This measure of susceptibility is then applied to anonymous voting rules. One set of results estimates the susceptibility of specific voting rules; an important finding is that several voting systems previously identified as resistant to manipulation are actually more susceptible than simple plurality rule, by the measure proposed here. A second set of results gives asymptotic lower bounds on susceptibility for any possible voting rule, under various combinations of efficiency, regularity, and informational conditions. These results illustrate how one can quantitatively explore the tradeoffs between susceptibility and other properties of the voting rule.

The third chapter carries the methodology of the second chapter to a market environment: unit-demand, private-value double auction markets. This chapter quanti-

tatively studies the tradeoff between inefficiency and susceptibility to manipulation, among all possible mechanisms for such markets. The main result approximately locates the possibility frontier, pinning it down within a factor that is logarithmic in the size of the market.

Thesis Supervisor: Parag A. Pathak
Title: Associate Professor of Economics

Thesis Supervisor: K. Daron Acemoglu
Title: Elizabeth and James Killian Professor of Economics

# Acknowledgments

Individual chapters of this dissertation thank people whose feedback directly contributed to the content. But a Ph.D. dissertation is also a product of a much longer and broader process, and this is a place to humbly recognize some of the many who have helped me along the way.

My advisors, Parag Pathak, Daron Acemoglu, and Glenn Ellison, have given generously to my professional development, guiding me to ask the right questions, offering practical help with life in the economics world, and placing confidence in me despite my youthful ignorance and questionable social skills. I have drawn on help from many others on the MIT economics faculty as well, and I am incredibly lucky to have been part of such a close and supportive community. My classmates made me miss out on the loneliness and frustration that are supposed to feature prominently in the graduate school experience, and I'd especially like to thank model officemate Alex Wolitzky, roommate Nathan Hendren, and occasional travel companion Arun Chandrasekhar, who have inspired me with their knowledge, energy, and good nature. The economics department staff also played an important role in keeping things running smoothly so that I could focus on my studies, and Gary King and Katherine Swan have been especially kind to me.

Of course, the process that has led to where I am now really began long before graduate school, and I owe the most thanks to my parents, Diana and Allen Carroll, who did so much to get me started, and who have provided me with the perfect mix of support, criticism, and laissez-faire over the years. And I thank my younger brother Julian, who continues to offer company and stimulating conversation despite the years of torment I've inflicted. Many guides and mentors also helped me chart my course forward — whether they know it or not — and among them I particularly thank Zvezdelina Stankova, Jim Propp, Holden Karnofsky, Peter Green, Dan Benjamin, and David Laibson.

On a personal level, I want to thank the friends who have given me needed emotional support and who have made the last few years an enjoyable time — especially Liz Zhang, Yan Zhang, Kenny Walden, Inna Zakharevich, Xiao Yu Wang, Ashley Swanson, Mike Powell, Aifang Guo, and of course Amanda Wang, who has meant so much to me during the period of most intense work.

For some reason the acknowledgments page of a dissertation often reads as though the product were a decisive end point. In a way it is; but I regard it more as one landmark on a longer road, and I look forward to thinking of the same people with renewed appreciation — and new ones as well — when the next landmark is reached.

# Contents

# Introduction

This dissertation ties together three papers on mechanism design with boundedly rational agents.

Economic theory is often criticized for relying on the unrealistic assumption that agents are perfectly rational. Mechanism design is an area of theory that should be especially sensitive to this criticism, because it is intimately tied to practical applications — witness the recent surge of interest in theoretically-informed design of complex auctions, and applications of matching theory to areas from school choice to organ transplantation. My hope is that this dissertation will represent the beginning of a much longer project to address this criticism systematically, by importing specific models of bounded rationality into mechanism design problems, and studying what changes and what stays the same, relative to the traditional benchmark of fully rational agents. Thus, the long-term goal is to understand in exactly what ways, and how seriously, imperfect rationality actually makes a difference for mechanism design.

The papers herein focus on direct mechanisms, where agents are asked to report their preferences and a decision is made based on those preferences. The standard design of (truthful) direct mechanisms is governed by incentive constraints, requiring that each type of agent should not be able to benefit by pretending to be any other type. The focus here is on forms of bounded rationality where agents' default behavior is to report their preferences truthfully, and they have only a limited ability to strategically report other preferences. Such limitations relax the incentive constraints. Accordingly, the question is whether, and to what extent, this relaxation of incentive constraints can help a mechanism designer better achieve her goals.

The first chapter of the dissertation studies the mathematical structure of incentive

constraints, in a variety of mechanism design problems. The question is whether local incentive-compatibility of a mechanism, by itself, logically implies full incentive-compatibility. The paper offers several motivations for this question; in the context of this dissertation, the relevant motivation is a simple model of bounded rationality, in which agents may be unable to contemplate all the possible strategic misreports of their preferences, but can at least consider misreports close to the true preferences. If local incentive constraints do not automatically imply full incentive-compatibility, then this raises the possibility that a designer could potentially benefit from agents' bounded rationality, by using mechanisms in which she expects agents will tell the truth, even though fully rational agents would instead find (nonlocal) manipulations.

It turns out that in a wide range of important mechanism design settings, local incentive constraints — once formulated in the proper way — do imply full incentive-compatibility. Thus, a mechanism designer cannot take advantage of agents' bounded rationality in this way. The paper uses a simple supermodularity argument to unify many continuous and discrete settings, showing that local incentive constraints are sufficient for full incentive-compatibility in each case, for probabilistic as well as deterministic mechanisms. Specifically, the results apply to any convex domain of cardinal or ordinal preferences, as well as domains of single-peaked or successive single-crossing ordinal preferences. For the ordinal domains in particular, these results imply computational versions of many previous impossibility theorems: If a designer wishes for her mechanism to satisfy some properties, and these properties cannot be met without creating some opportunity for strategic manipulation, then in fact, there will always be some such manipulations that can be found with a very small amount of computational effort (by only searching for local manipulations).

The second, and main, chapter advances the methodology of looking quantitatively at incentives for strategic behavior in mechanisms. This is motivated by the view — which has been increasingly influential in recent applied market design literature — that agents will not bother to be strategic if the incentive to do so is small enough. The underlying boundedly-rational model is one in which agents face a small, additive utility cost to behaving strategically (which may be interpreted as a computational

or a psychological cost), whereas being truthful is costless. Then, all else being equal, agents are more likely to be truthful if the incentive to be strategic is smaller. The paper argues that we can and should understand just how small the strategic incentives are in particular mechanisms. Such a quantitative understanding allows us to compare different mechanisms in terms of their incentive properties, and helps inform choices among mechanisms that require trading off strategic incentives against other properties.

More specifically, a direct mechanism's *susceptibility to manipulation* is defined as the maximum amount of expected utility that any agent can gain by behaving strategically instead of sincerely. The maximum is taken over all preferences an agent might have and all possible beliefs about other agents' behavior, within some natural restrictions chosen depending on the environment.

As a concrete application, the paper considers anonymous voting rules, in which each voter in a population submits a preference ranking over candidates and these preferences are aggregated to determine a winner. The paper presents two sets of results. The first set gives quantitative bounds on the susceptibility of several voting systems discussed in prior literature. These results illustrate how to perform simple estimates using the measure of susceptibility advanced here. A highlight of this section of the paper is a finding that many voting systems previously identified as relatively resistant to strategic manipulation, using a straightforward profile-counting measure, actually perform worse than a simple plurality vote under this measure of susceptibility, as long as the number of candidates is moderately large. This shows that this measure of susceptibility leads to new insights and does not simply duplicate previously known observations. The second set of results consists of several theorems providing asymptotic lower bounds on the susceptibility of any voting rule satisfying various combinations of efficiency, regularity, and informational properties. These theorems illustrate how one can quantitatively study the tradeoffs between incentives for strategic behavior and other properties that might be desired in a voting rule. The proof technique for these theorems is also a contribution; for many of the results, we simply take proofs of non-quantitative impossibility theorems and keep track of

error terms.

This second chapter studies the problem of designing a voting rule because it is among the oldest and most widely-studied problems in mechanism design, and also one where there has been significant previous literature attempting to measure manipulation, using other approaches. However, the concept of susceptibility can of course be applied to other problems. As another illustration — and one which perhaps fits more familiarly into the domain of economic problems — the third chapter of the dissertation takes the same measure of susceptibility and applies it to large private-value double auction environments, studying the quantitative tradeoff between efficiency and strategic incentives.

In such an environment (which may be thought of as a stylized model of more general market environments), there are sellers who each have one unit of a homogeneous good for sale, and buyers who have money and would like to buy a good. Each seller, however, has his own, privately known value for the good (or cost of procuring it), and each buyer has his own value as well. A (direct) mechanism asks each buyer and seller his value, and as a function of these reported values, determines who trades with whom and how money changes hands. A mechanism designer's aim is to try to get the goods to whichever combination of buyers and sellers values them the most.

It is well established in previous literature that, under participation and weak budget-balance constraints, there is no way to achieve the first-best efficient outcome while also giving all participants perfect incentives to report their values truthfully. On the other hand, it is possible to achieve the first-best assuming agents are truthful, with incentives to be strategic becoming negligible as the market becomes large (using a $k$-double auction, a version of the Walrasian equilibrium mechanism). It is also possible to give perfect incentives while achieving asymptotic efficiency as the market becomes large (using a double auction due to Preston McAfee). However, under the model of this paper, the susceptibility of the $k$-double auction and the inefficiency of the McAfee double auction both converge to zero at the fairly slow rate of $1/\sqrt{N}$ (where $N$ measures the size of the market). A designer might be concerned that, in moderate-sized markets, this susceptibility to manipulation for the $k$-double auction

is still too large to prevent manipulation by agents who have only small costs to being strategic. This leads to the question of whether there is some other mechanism with significantly smaller susceptibility than the $k$-double auction (but not necessarily zero), and also significantly better efficiency performance than the McAfee double auction. The main result of the chapter answers this question in the negative, by showing that no mechanism does better than order $1/(\sqrt{N}\log N)$ on both dimensions simultaneously. The lesson is that, even in a model in which agents have a small cost to behaving strategically, there is no way to do substantially better than the known mechanisms.

The papers presented here are only the start of a long project on bounded rationality in mechanism design. One can imagine many different forms of bounded rationality that might be theoretically modeled. My hope is that this work will inspire more exploration in these related but distinct directions.

# Chapter 1

# When Are Local Incentive Constraints Sufficient?


## Abstract

We study the question of whether local incentive constraints are sufficient to imply full incentive-compatibility, in a variety of mechanism design settings, allowing for probabilistic mechanisms. We give a unified approach that covers both continuous and discrete type spaces. On many common preference domains — including any convex domain of cardinal or ordinal preferences, single-peaked ordinal preferences, and successive single-crossing ordinal preferences — local incentive-compatibility (suitably defined) implies full incentive-compatibility. On domains of cardinal preferences that satisfy a strong nonconvexity condition, local incentive-compatibility is not sufficient. Our sufficiency results hold for dominant-strategy and Bayesian Nash solution concepts and allow for some interdependence in preferences.



Thanks to (in random order) Alex Wolitzky, Ron Lavi, Suehyun Kwon, Vince Conitzer, Anton Tsoy, Alex Frankel, Ariel Procaccia, Parag Pathak, Daron Acemoglu, Tim Roughgarden, Rakesh Vohra, and Glenn Ellison for helpful comments and suggestions. This work was supported by an NSF Graduate Research Fellowship.


## 1 Introduction

In the analysis of mechanism design problems, taking account of all the possible constraints imposed by incentive-compatibility at once can be unwieldy. It can be


This chapter was originally published in *Econometrica*, volume 80, issue 2, March 2012, pages 661-686. Appendix B was originally published online in *Econometrica*'s supplemental material.

easier to focus attention on local incentive constraints, ensuring that agents have no incentive to make "small" misreports of their type, and then check at the end of the analysis whether or not the mechanisms obtained are fully incentive-compatible.

In the present paper, we ask the general question of whether local incentive constraints are sufficient on their own to guarantee full incentive-compatibility, and obtain an affirmative answer in a wide array of settings. We allow for arbitrary probabilistic mechanisms, which specify a distribution over some (exogenously specified) outcome space as a function of an agent's type. (Our analysis is mostly worded in terms of a single agent, but we show how it readily extends to multi-agent mechanisms, including allowing a limited degree of interdependence in preferences.)

To clarify the significance of these results, it is useful to distinguish two major branches of mechanism design literature. We give a simple and unified approach that applies to both branches but has slightly different implications for the two. One branch, with roots in axiomatic social choice theory, studies problems without monetary transfers. These include voting [25, 50, 39, 7, 48], matching [45, 2, 14], queueing [15], and rationing [52, 22], among others. This literature has recently been influential in applied market design as well; see for example [46] and references therein. It is commonly taken as given here that each agent submits a ranking over outcomes (such as candidates, in a voting context, or schools or jobs, in a matching context) to the mechanism. Thus, agents report ordinal preferences. Incentive-compatibility typically means that reporting one's true preferences should be a dominant strategy. We will say that such a mechanism is locally incentive-compatible if no agent type can benefit from misreporting by switching some two consecutive outcomes in his preference ranking. We show below that for many of the most common preference domains considered in this literature, local incentive-compatibility implies full incentive-compatibility. Specifically, we show this for domains of ordinal preferences having convex closure (Proposition 3.2, which actually gives a generalization to *polyhedral* type spaces); single-peaked ordinal preferences (Proposition 3.3); and successive single-crossing ordinal preferences (Proposition 3.4).

The second large branch of literature concerns settings in which monetary trans-

fers are possible and agents have quasilinear preferences, with applications such as monopoly pricing, auctions, and public projects. Seminal works include [40, 28, 41, 42, 37, 34, 35]. It is generally assumed that agents can report a cardinal valuation for each outcome. We show that local incentive-compatibility, suitably formulated, implies full incentive-compatibility whenever the space of agents' cardinal types is convex (Proposition 3.5). (We should note that this result also has some relevance to the no-transfers literature, as some authors there also allow agents to report cardinal preferences, e.g. [5, 31, 53].)

Our results on the sufficiency of local incentive constraints are relevant for several reasons. One is that they provide a technical tool to facilitate the researcher's task of analyzing mechanism design problems. This is particularly relevant to the transfers branch of the literature, where analysis typically begins by using local incentive constraints and the envelope theorem to obtain an integral formula for the utility attained by each type of agent (as in e.g. [41, Lemma 2]); our sufficiency results provide a general tool to help assure researchers that this reduction of the problem has not neglected important nonlocal constraints. They also can streamline proofs of incentive-compatibility for newly designed mechanisms, since it is enough to show local incentive-compatibility and then invoke sufficiency.

Moreover, our analysis casts light on the *form* of local incentive constraints needed. It is *not* sufficient to specify only that each type of agent should be unable to profitably misreport as any nearby type; one must also specify that each type cannot serve as a profitable misreport for any nearby type. (See the discussion in Subsection 3.1.)

A separate reason our results are relevant is that one may have more literal reasons to impose only a subset of incentive constraints. For example, there may be a monitoring technology that makes it possible to detect and punish reports far away from an agent's true type, in which case the mechanism designer does not need to worry about such misreports (as in Green and Laffont [29]). One might hope that this would provide an operational way to circumvent impossibility results such as the Gibbard-Satterthwaite theorem [25, 50], which are pervasive in the no-transfers literature; or, in a setting with transfers, one might hope, say, to obtain higher rev-

enue than would be possible with fully incentive-compatible mechanisms. If agents are to report truthfully, then our sufficiency results show that in many settings, having access to such a monitoring technology does not enlarge the space of effective mechanisms.

Relatedly, when designing a mechanism for boundedly rational agents, one might consider that agents are not capable of contemplating every possible misreport of their preferences, and again ask whether this provides an operational way to improve on fully incentive-compatible mechanisms. If the designer believes that agents are at least rational enough to be capable of imitating any nearby type, then in the settings covered by our sufficiency results, imposing only the relevant subset of incentive constraints actually does not enlarge the space of usable mechanisms at all.

In particular, for ordinal type spaces, this idea leads to "computational" versions of many existing impossibility or characterization results. This gives a very general reply to a literature that seeks ways around the Gibbard-Satterthwaite theorem by devising voting mechanisms that are computationally difficult, but not impossible, to strategically manipulate (e.g. [9, 8]). On the type spaces where our sufficiency results apply, they immediately imply that any such mechanism is easy to manipulate in some instances, as long as the outcome of the mechanism itself is easy to compute. (Here, as in the preceding literature, we take "easy" to mean computable in polynomial time.) Namely, a manipulator can exhaustively consider each local manipulation — switching some two candidates who are consecutive in the ranking — and compute the outcome of the mechanism; this trial-and-error search is easy and will find an advantageous manipulation in some instances. So a computational-complexity constraint, at least of the naive form, cannot prevent agents from manipulating.[1]

More broadly, there is a tradition in social choice theory of looking for the weakest assumptions necessary to obtain a characterization or impossibility result. Our results can be immediately applied to many axiomatic characterizations (such as those cited in the third paragraph), showing that, say, an axiom requiring dominant-

[1]In the Gibbard-Satterthwaite context, stronger results extending this idea are already known (e.g. [32]). But our results lead more generally to computational versions of many other existing characterization results by the same argument.

strategy incentive-compatibility can be replaced by local incentive-compatibility without changing the conclusion.

The aforementioned results show that, for many important type spaces, local incentive-compatibility implies full incentive-compatibility. On the other hand, there are type spaces where the implication does not hold. In particular, we show this for domains of cardinal preferences that fail to be convex in a sufficiently strong way (Proposition 4.1).

Our work connects with several previous papers on mechanism design under a subset of incentive constraints. Green and Laffont [29], mentioned above, consider a general setup in which the space of messages that agents can send equals the space of types, with exogenous restrictions as to which messages each type is capable of sending, and study when the revelation principle applies. Celik [16] and Sher and Vohra [51] consider specific mechanism design problems under subsets of incentive constraints, though their subsets are not local in our sense.

There does not appear to be previously published work asking the broad question of when local and full incentive-compatibility coincide. However, a contemporaneous paper by Sato [49], independent of ours, does address this question. Sato considers only deterministic mechanisms over ordinal type spaces. For such mechanisms, Sato shows that local incentive constraints are sufficient on all of the ordinal type spaces that we consider (type spaces with convex closure, single-peaked, and successive single-crossing preferences), as well as some others.

This paper also bears some formal resemblance to recent work on general settings with cardinal preferences and transfers. In such a setting, a rule mapping types to outcomes is *implementable* if there exists some accompanying payment function (mapping types to transfers) that makes truthful revelation incentive-compatible. There has recently been much interest in simple conditions ensuring that a rule is implementable, e.g. [47, 12, 4]. In particular, our work is somewhat reminiscent of a paper by Archer and Kleinberg [3]. They show that local implementability (suitably defined) implies implementability, on any convex space of cardinal types. However, we show that local implies full incentive-compatibility for a given mechanism, consisting

of an outcome rule and a payment function together, whereas they show that local implementability by *some* payment functions (possibly using different payment functions on different local neighborhoods) implies full implementability. Thus, both their hypothesis and their conclusion are weaker than ours. Moreover, their sets of local incentive constraints are larger than ours, and their theorem would not hold using our constraint sets; this is discussed in detail in Subsection 3.1. Accordingly, our results on cardinal type spaces do not follow from the result of Archer and Kleinberg, nor vice versa.

# 2 Framework

We begin with the general framework. Ensuing sections will give the concrete results.

## 2.1 Definitions

We will focus on incentives in a mechanism for an individual agent. In Subsection 2.2, we will show how the ideas extend straightforwardly to multi-agent mechanisms with private values. We begin by introducing the definitions for the no-transfers setting; in Subsection 3.5 we will allow for transfers, and also for interdependence.

From the agent's point of view, a mechanism takes the agent's preferences as input and determines an outcome, or a probability distribution over outcomes. We must be explicit about the form of preferences that the agent can announce. In some settings, it is standard practice to assume agents announce their cardinal valuation for each of the possible outcomes. In others (specifically in the no-transfers literature), it is assumed that agents only report an ordinal ranking of outcomes.

This latter assumption entails exogenously restricting the message space of the mechanism to consist of the possible ordinal preferences. This restriction is widely accepted, although it does not yet enjoy solid theoretical foundations. It is often made for analytical tractability, and in practical market design applications it can also be justified by the need to make the mechanism accessible to participants who may have difficulty thinking about their preferences over lotteries. Bogomolnaia and

Moulin [14] give a more detailed discussion on this last point.

Finally, in some settings, one might assume that agents report even coarser information than ordinal preferences (for example, they are required to rank only a limited number of outcomes). We will first give a unified treatment that covers all of the different specifications of preferences, then specialize to define local incentive-compatibility in specific settings.

Let $X$, the *outcome space*, be any finite set; $m$ will denote its cardinality. Let $\Delta(X)$ denote the space of lotteries over $X$. The agent is assumed to have expected utility preferences over lotteries. It will be convenient to think of both lotteries over $X$ and utility functions as elements of $\mathbb{R}^m$. If the agent's utility function is $u$, his payoff from a lottery $L$ is given by the inner product $u \cdot L$.

For subsets of $\mathbb{R}^m$, cl denotes the closure and $\partial$ the boundary operator. If $u, v \in \mathbb{R}^m$, we write $[u, v]$ for the line segment $\{(1-\alpha)u + \alpha v \mid \alpha \in [0,1]\}$.

A *type* is a nonempty subset of $\mathbb{R}^m$. A *type space* is a set of pairwise disjoint types. We henceforth use the term *type space* in preference to *domain*: the latter term suggests only an exogenous restriction of the set of utility functions the agent may have, whereas our notion of a type space conveys both which utility functions are possible and which ones the mechanism is required to treat identically.

Given a type space $T$, a *mechanism* is a function $f : T \to \Delta(X)$. Thus, the mechanism chooses a distribution over outcomes, based on the agent's (reported) type.

An *incentive constraint* is an ordered pair of types. The interpretation of the constraint $(t, t')$ is that a type $t$ cannot benefit from misreporting as type $t'$. Accordingly, we say that a mechanism $f$ *satisfies* an incentive constraint $(t, t')$ if, for all $u \in t$, $u \cdot f(t) \geq u \cdot f(t')$; equivalently, $u \cdot (f(t) - f(t')) \geq 0$. A mechanism *satisfies a set* of incentive constraints if it satisfies every constraint in the set.

A mechanism that satisfies the full set of incentive constraints $T \times T$ is *fully incentive-compatible.* This is exactly the usual meaning of incentive-compatibility.

A set $S$ of incentive constraints is *sufficient* if every mechanism that satisfies $S$ is fully incentive-compatible.

We highlight several important kinds of type spaces and define local incentive constraints in each case.

- A type space $T$ is *cardinal* if every type is a singleton. In this case, abusing notation, we will think of types as vectors and $T$ as a subset of $\mathbb{R}^m$. For example, we write $f(u)$ rather than $f(\{u\})$.

  For a cardinal $T$, a set $S$ of incentive constraints will be called *local incentive constraints* if every $u \in T$ has an open neighborhood $N_u$ in $T$ (with the relative topology) such that $(u, u') \in S$ and $(u', u) \in S$ for every $u' \in N_u$.

- A type space is *ordinal* if every type is of the form $t = \{u \mid u(x_1) > u(x_2) > \cdots > u(x_m)\}$ for some strict ordering $x_1 \succ \cdots \succ x_m$ of the elements of $X$. We say that $t$ *represents* this ordering. Note that our definition of an ordinal type space does not require that all possible orderings be represented by types.

  When types are ordinal, $f$ satisfies a constraint $(t, t')$ if and only if the lottery $f(t)$ first-order stochastically dominates $f(t')$ with respect to the ordering on $X$ represented by $t$. (This is easy to show.)

  Call two ordinal types $t, t'$ *adjacent* if the orderings they represent differ only by a switch of two consecutive outcomes. On any ordinal type space $T$, the *local incentive constraints* will refer to the set of all constraints $(t, t')$ such that $t$ and $t'$ are adjacent.

- More generally, we can consider *polyhedral* type spaces. In the space of utility functions, $\mathbb{R}^m$, an *open half-space* is a set of the form $\{u \mid u \cdot \lambda > c\}$ for some nonzero $\lambda \in \mathbb{R}^m$ and some constant $c$. If $H$ is such an open half-space, its closure $\text{cl}(H) = \{u \mid u \cdot \lambda \geq c\}$ is a *closed half-space*, and its boundary $\partial H = \{u \mid u \cdot \lambda = c\}$ is a *hyperplane*. Define an (open) *polyhedron* to be a nonempty set that is the intersection of finitely many open half-spaces. A *polyhedral type space* is a type space consisting of finitely many types that are all polyhedra.

Say that two disjoint polyhedra $t, t'$ are *adjacent* if $\text{cl}(l) \cap \text{cl}(t')$ contains a nonempty, relatively open subset of a hyperplane. In simpler terms, $t$ and $t'$ are polyhedra that border along a face. We then let the *local incentive constraints* on $T$ be the set of constraints $(t, t')$ such that $t$ and $t'$ are adjacent.

Any ordinal type space is polyhedral, and one can check that the definitions of adjacency and local incentive constraints for ordinal types agree with those for polyhedral types. (There exists previous literature in mechanism design also using polyhedra to represent ordinal types, e.g. Duggan [21].) For another example, take the types implied by truncated rankings, i.e. $\{u \mid u(x_1) > \cdots > u(x_p)$ and $u(x_p) > u(y)$ for all $y \neq x_1, \ldots, x_p\}$, for any distinct outcomes $x_1, \ldots, x_p$ with $p < m$ — these are again polyhedral types. Such a type space is natural for studying matching mechanisms in applications such as school choice, where students may be asked to rank, say, 12 favorite schools out of more than 500 available [1].

More generally, any mechanism with a *finite* message space gives rise to a polyhedral type space: for each message, the set of utility functions for which it is optimal forms a polyhedron (ignoring boundary issues). Studying local incentives in these type spaces can be helpful for analyzing such mechanisms. Gibbard [27] gives a fairly complete analysis of dominant-strategy voting mechanisms with arbitrary finite message spaces; much of the analysis focuses on incentives to misreport locally.

We say that a mechanism is *locally incentive-compatible* if it satisfies some set of local incentive constraints (in the cardinal case; or the canonical such set in the polyhedral case).

We are interested in determining whether or not local incentive contraints are sufficient, on various type spaces.

## 2.2 Mechanisms with multiple agents

As already mentioned, while we focus on single-agent mechanisms, our results apply also with multiple agents, under private values. The extension is similar to arguments in previous literature [3, 30], but we spell it out in detail here, as we will further build on it in Subsection 3.5.

Define a mechanism with $n$ agents, type space $T = T_1 \times \cdots \times T_n$, and outcome space $X$ to be a map $f : T \to \Delta(X)$, specifying a (probabilistic) outcome as a function of all the agents' types. Suppose that, for some $i$, a set $S_i$ of incentive constraints is sufficient for $T_i$.

One possible notion of incentive-compatibility is to say that $f$ satisfies the incentive constraint $(t_i, t_i') \in T_i \times T_i$ for agent $i$ if, for all $t_{-i}$ and all $u_i \in t_i$, we have $u_i \cdot (f(t_i, t_{-i}) - f(t_i', t_{-i})) \geq 0$. If $f$ satisfies every incentive constraint in $S_i$ for agent $i$, then holding fixed any profile $t_{-i}$, the single-agent mechanism $t_i \mapsto f(t_i, t_{-i})$ satisfies $S_i$ and so (by sufficiency) is fully incentive-compatible. Thus, $f$ is fully incentive-compatible in dominant strategies (for agent $i$).

One can also consider Bayesian incentive-compatibility. Suppose we are given a probability distribution $\psi_j$ over $T_j$ for each agent $j$, and assume $f(t_i, t_{-i})$ is measurable in $t_{-i}$ for all $t_i$. Then we can say that $f$ satisfies incentive constraint $(t_i, t_i')$ for agent $i$ if, for all $u_i \in t_i$, we have $u_i \cdot (E_i[f(t_i, t_{-i})] - E_i[f(t_i', t_{-i})]) \geq 0$, where the expectation is over $t_{-i}$ with respect to the product distribution $\times_{j \neq i} \psi_j$. Again, if $f$ satisfies each incentive constraint in $S_i$, then the single-agent mechanism $t_i \mapsto E_i[f(t_i, t_{-i})]$ satisfies $S_i$ and so is fully incentive-compatible for agent $i$. This is the standard notion of Bayesian incentive-compatibility for $f$. Notice that this argument depends on the agents' types being independently distributed: the expectation $E_i$ needs to be defined in a way that does not depend on $t_i$.

# 3 Sufficiency

In this section we show that local incentive constraints are sufficient on a variety of common type spaces. All proofs absent from the main text are in Appendix A.

## 3.1 Cardinal type spaces

Recalling that a cardinal type space is identified with a subset of $\mathbb{R}^m$, we can state our first sufficiency result:

**Proposition 3.1** *On a convex cardinal type space $T$, any set of local incentive constraints is sufficient.*

We present the proof in detail, since the proofs of most of our other sufficiency results (Propositions 3.2, 3.3, 3.5) follow the same model. To prove that an agent of type $u$ never wants to misreport as type $v$, we restrict attention to types along the line segment $[u, v]$, effectively reducing to one dimension; we then break the segment into short pieces for which local incentive constraints apply, and combine these local incentive constraints into the incentive constraint $(u, v)$ using the kind of supermodularity or "revealed-preference" argument that is familiar elsewhere in the mechanism design literature (see e.g. [41, Lemma 2], [44, Theorem 1]).

**Proof:** Let $S$ be a set of local incentive constraints and $f$ a mechanism satisfying $S$. For types $u, v$, write $u \leftrightarrow v$ if $(u, v)$ and $(v, u)$ are both in $S$. By definition, every $u \in T$ has some neighborhood $N_u$ in $T$ such that $u \leftrightarrow v$ for all $v \in N_u$.

Fix arbitrary $u, v \in T$. We want to show that $u \cdot (f(u) - f(v)) \geq 0$.

For any $\alpha \in [0, 1]$, define $u_\alpha = (1 - \alpha)u + \alpha v$. Convexity implies $u_\alpha \in T$. Let

$$A = \{\alpha \mid \text{there exist } 0 = \alpha_0 < \alpha_1 < \cdots < \alpha_r \leq 1 \text{ with } u_{\alpha_0} \leftrightarrow u_{\alpha_1} \leftrightarrow \cdots \leftrightarrow u_{\alpha_r} \text{ and } \alpha_r = \alpha\}.$$

Clearly, if $\alpha \in A, \alpha < \alpha' \leq 1$, and $u_\alpha \leftrightarrow u_{\alpha'}$, then $\alpha' \in A$. Now let $\overline{\alpha} = \sup A \geq 0$. If $\overline{\alpha} = 0$ then $\overline{\alpha} \in A$. If $\overline{\alpha} > 0$, then for $\alpha$ sufficiently close to $\overline{\alpha}$ we have $u_\alpha \leftrightarrow u_{\overline{\alpha}}$; since we can choose $\alpha \in A$ arbitrarily close, we again get $\overline{\alpha} \in A$. Moreover, if $\overline{\alpha} < 1$, then $u_{\overline{\alpha}} \leftrightarrow u_\alpha$ for $\alpha$ just slightly larger than $\overline{\alpha}$; this implies $\alpha \in A$, contradicting $\overline{\alpha} = \sup A$. Therefore, we get $\overline{\alpha} = 1$ and $1 \in A$.

So we have $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_r = 1$ with $u_{\alpha_k} \leftrightarrow u_{\alpha_{k+1}}$ for each $k$. Now write out the local incentive constraints:

$$\begin{aligned} u_{\alpha_k} \cdot (f(u_{\alpha_k}) - f(u_{\alpha_{k+1}})) &\geq 0, \\ u_{\alpha_{k+1}} \cdot (f(u_{\alpha_{k+1}}) - f(u_{\alpha_k})) &\geq 0. \end{aligned}$$

Multiplying by $\alpha_{k+1}$ and $\alpha_k$, respectively, and adding gives

$$[\alpha_{k+1} u_{\alpha_k} - \alpha_k u_{\alpha_{k+1}}] \cdot (f(u_{\alpha_k}) - f(u_{\alpha_{k+1}})) \geq 0.$$

But one directly calculates that $\alpha_{k+1} u_{\alpha_k} - \alpha_k u_{\alpha_{k+1}} = (\alpha_{k+1} - \alpha_k)u$. Since $\alpha_{k+1} - \alpha_k > 0$, we can divide through to obtain

$$u \cdot (f(u_{\alpha_k}) - f(u_{\alpha_{k+1}})) \geq 0.$$

Now we can sum over $k = 0, 1, \ldots, r-1$, and telescoping gives

$$u \cdot (f(u) - f(v)) = u \cdot (f(u_{\alpha_0}) - f(u_{\alpha_r})) \geq 0.$$

□

Proposition 3.1 applies to any convex cardinal type space. This includes, for example, the full space of utility functions on $X$; or the space of utility functions that are increasing with respect to some partial order on $X$; or the space of supermodular or submodular utility functions, given a lattice structure on $X$; or the space of utility functions satisfying some concavity conditions.

The proof of Proposition 3.1 clearly uses both parts of the definition of local incentive constraints — that each $u$ should have a neighborhood $N_u$ with both $(u, u') \in S$ and $(u', u) \in S$ for $u' \in N_u$. A seemingly more natural way to define local incentive constraints would only require $(u, u') \in S$. Under this definition, Proposition 3.1 would no longer hold. For example, suppose $X = \{x, y\}$ and $T$ is the full space of all cardinal types. Consider the mechanism $f$ given by $f(u) = x$ if $u(x) < u(y)$

and $f(u) = y$ otherwise. This $f$ meets the weaker definition of local incentive-compatibility, but is not fully incentive-compatible. (Requiring only $(u', u) \in S$ would also not be enough: with the same $X$ and $T$, consider the mechanism $f(u) = x$ if $u(x) = u(y) - 1$ and $f(u) = y$ otherwise.)[2]

By contrast, the local-to-global result of Archer and Kleinberg [3, Corollary 3.7], on implementability in a quasilinear setting, effectively requires stronger local incentive constraints. They assume implementability throughout each $N_u$ — that is, for each $u$, there should be some payment function $p_u$ (specifying a payment for each agent type) so that the mechanism-with-transfers $(f, p_u)$ satisfies incentive constraints $(u', u'')$ for all $u', u'' \in N_u$. The analogue of our constraints in their setting would be to merely require that $(f, p_u)$ should satisfy constraints $(u, u')$ and $(u', u)$ for all $u' \in N_u$. This requirement is a local form of *weak monotonicity*, which is not enough to imply their implementability conclusion without further restrictions; see [12, Example S1] or [47, Section 7].

Unlike the local constraints of [3], ours can be expressed succinctly in terms of local maxima: $f$ is locally incentive-compatible if every $u \in T$ is a local maximum of both the functions $v \mapsto u \cdot f(v)$ and $v \mapsto v \cdot (f(u) - f(v))$. With this interpretation, local incentive-compatibility can potentially be checked by first- and second-order conditions at points where $f$ is differentiable. This convenience is relevant in making the reduction from global to local incentive-compatibility a useful one: if one wishes to check incentive constraints directly, then even local incentive-compatibility can require checking many constraints when $T$ is high-dimensional, since it is necessary to check constraints in every direction at each $u$.

## 3.2 Polyhedral type spaces

Next we consider polyhedral type spaces. Our main result here is:

[2] A referee points out that a mechanism on a cardinal type space is fully incentive-compatible if and only if the indirect utility function $u \mapsto u \cdot f(u)$ is convex, with $f(u)$ belonging to the subdifferential at each point $u$. Our two local conditions can be viewed loosely as local forms of these requirements: the subdifferential condition at $u$ is equivalent to satisfying $(u', u)$ for all $u' \in T$, so our requiring this for all $u' \in N_u$ gives a local form of the subdifferential condition, and then imposing the additional constraints $(u, u')$ ensures convexity.

**Proposition 3.2** *Let $T$ be a polyhedral type space such that $\cup_{t \in T} \mathrm{cl}(t)$ is convex. Then the set of local incentive constraints is sufficient.*

The argument is essentially the same as for Proposition 3.1. For utility functions $u$ and $v$, we consider the line segment $[u, v]$; this segment passes through various types in succession. By jiggling $v$ a bit if necessary, we can ensure that any two successive types along this line segment are adjacent polyhedra, and then we can just add up the corresponding local incentive constraints as before.

A particular case of Proposition 3.2 is that on the full space of all ordinal types over a given $X$, the local incentive constraints are sufficient.[3] (The union of the closures of all types is simply all of $\mathbb{R}^m$.) Proposition 3.2 also applies when $T$ consists of all ordinal types that respect a given partial ordering on $X$. For example, Bogomolnaia and Moulin [15] consider an allocation problem with real objects and a null object; all types have the same preference ordering on the real objects, but rank the null object differently relative to the real objects.

## 3.3 Single-peaked preferences

The preceding results have focused on essentially convex type spaces. One important nonconvex type space is that of single-peaked preferences.

Fix an ordering $x_1, \ldots, x_m$ of the outcomes in $X$. A strict preference ordering $\succ$ over $X$ is *single-peaked* if there exists some outcome $x_{p^*}$ such that, whenever $q < p \leq p^*$ or $q > p \geq p^*$, we have $x_p \succ x_q$. An ordinal type is *single-peaked* if it represents a single-peaked ordering.

Single-peaked preferences have been popular in voting theory ever since Black's [13] observation that the rule choosing the median of the voters' favorite outcomes is dominant-strategy incentive-compatible. Single-peaked preferences are also important in economic applications because single-peakedness is the same as quasiconcavity of the utility function (aside from issues of indifference). Moulin [39] characterizes

[3] An analogous result also holds if we allow indifferences — so that for each weak order on $X$, the set of utility functions representing it constitutes a type — with an appropriate definition of local incentive constraints. We omit the details here.

dominant-strategy incentive-compatible deterministic voting systems under single-peaked preferences. (Moulin assumes the outcome space is the whole real line, but his proofs carry through almost unchanged for a finite outcome space.) Ehlers, Peters, and Storcken [23] extend this work to probabilistic mechanisms. Sprumont [52], Barberà, Jackson, and Neme [6], and Ehlers and Klaus [22] study rationing problems when consumers have single-peaked preferences over quantities.

The space of single-peaked ordinal types does not meet the convexity condition of Proposition 3.2. However, we still have the result:

**Proposition 3.3** *Fix an ordering $x_1, \ldots, x_m$ of the elements of $X$. On the space of single-peaked ordinal types, the set of local incentive constraints is sufficient.*

The argument is a slight extension of that used for Proposition 3.2. In general, in an ordinal type space, say that a utility function $v$ is *accessible* from another utility function $u$ if the segment $[u, v]$ is contained in the union of the closures of all types. In this case we can apply the argument of adding up local incentive constraints from Propositions 3.1 and 3.2. Now, in the single-peaked ordinal type space, it is no longer true (as it was for Proposition 3.2) that all $v$ in a given type $t'$ are accessible from $u \in t$, but we actually only need to be able to find some such $v$ for each $u$. Lemma A.5 in Appendix A shows that this can be done.

One can also consider the space of single-dipped ordinal types [36], or of single-peaked ordinal preferences on a tree [20, 18]. It is straightforward to extend the proof to cover each of these cases, showing that the local incentive constraints are again sufficient.

## 3.4 Single-crossing preferences

Besides single-peaked preferences, another economically important class of ordinal type spaces is given by single-crossing preferences. These are defined as follows: Fix an ordering $x_1, \ldots, x_m$ of the elements of $X$. A sequence $\succ_1, \ldots, \succ_r$ of distinct strict preference orderings is a *single-crossing preference domain* if the following holds: whenever $p < q$ and $x_q \succ_k x_p$ for some $k$, we also have $x_q \succ_l x_p$ for all $l > k$.

Single-crossing ordinal preferences arise in economic models such as the redistributive taxation models of Roberts [43] and Meltzer and Richard [38] (see Saporiti [48] for references to other applications). Just as with single-peaked preferences, preferences coming from any single-crossing domain satisfy a median voter property — the voting scheme that chooses the outcome most preferred by the voter with the median preference is dominant-strategy incentive-compatible. More generally, Saporiti [48] characterizes dominant-strategy incentive-compatible voting schemes on any maximal single-crossing preference domain.

For any strict preference ordering $\succ$ on $X$, let $V(\succ) = \{(p,q) \mid p < q, x_p \prec x_q\}$. By definition, a sequence of preference orderings $\succ_1, \ldots, \succ_r$ is a single-crossing preference domain if and only if $V(\succ_1) \subseteq \cdots \subseteq V(\succ_r)$. In fact, these inclusions must all be strict, since any ordering $\succ$ can be uniquely reconstructed from $V(\succ)$. Therefore $|V(\succ_1)| < \cdots < |V(\succ_r)|$. Call the domain a *successive single-crossing preference domain* if $|V(\succ_{k+1})| = |V(\succ_k)| + 1$ for each $k = 1, \ldots, r-1$.

This covers the domains considered in [48] — any maximal single-crossing preference domain $\succ_1, \ldots, \succ_r$ is successive. For suppose that $|V(\succ_{k+1})| - |V(\succ_k)| > 1$ for some $k$. There must be some two alternatives $x_p, x_q$ that are ranked consecutively by $\succ_k$, with $x_p \succ_k x_q$ but $x_q \succ_{k+1} x_p$. Single-crossing ensures $p < q$. By switching the positions of $x_p$ and $x_q$ in $\succ_k$, we get a new ordering $\succ'$ with $V(\succ') = V(\succ_k) \cup \{(p,q)\}$, and hence $V(\succ_k) \subset V(\succ') \subset V(\succ_{k+1})$. This means that $\succ_1, \ldots, \succ_k, \succ', \succ_{k+1}, \ldots, \succ_r$ is again a single-crossing preference domain, contradicting maximality.

For any successive single-crossing preference domain $\succ_1, \ldots, \succ_r$, call the corresponding space of ordinal types $T = \{t_1, \ldots, t_r\}$ a *successive single-crossing ordinal type space.* In this case, the local incentive constraints are precisely those of the form $(t_k, t_{k+1})$ or $(t_{k+1}, t_k)$. We shall show that on such a type space, the local incentive constraints are sufficient. This result may be surprising, since these incentive constraints are especially parsimonious — each type is adjacent to just two other types.

**Proposition 3.4** *On any successive single-crossing ordinal type space, the local incentive constraints are sufficient.*

The strategy of proof is a little different from that used for the previous propositions. Instead of breaking a single line segment into short pieces, we find a sequence of parallel line segments, each connecting two consecutive types $t_k, t_{k+1}$, but such that each segment need not begin where the previous one ended. (As pointed out by a referee, this method has some precedent in Gibbard [26, Lemma 2], where a similar argument is applied to the full ordinal type space; and the argument can be applied to the space of single-peaked ordinal types as well.)

**Proof:** Suppose the mechanism $f$ satisfies the local incentive constraints. Fix any two types $t_l, t_{l'}$, and let $u \in t_l$. We wish to show that $u \cdot (f(t_l) - f(t_{l'})) \geq 0$. We will show this for $l' > l$; the proof for $l' < l$ is similar.

In fact it suffices to show that

$$u \cdot (f(t_k) - f(t_{k+1})) \geq 0 \qquad \text{for } k \geq l, \tag{3.1}$$

since then we can sum up (3.1) for $k = l, l+1, \ldots, l'-1$ to obtain $u \cdot (f(t_l) - f(t_{l'})) \geq 0$.

So fix $k \geq l$, and also define $M = \max_x u(x) - \min_x u(x)$. Write $V(\succ_{k+1}) \setminus V(\succ_k) = \{(p, q)\}$ by successiveness; then $p < q$, and $\succ_k$ ranks $x_p$ just above $x_q$. Because $u \in t_l$ with $l \leq k$, single-crossing implies that $u(x_p) > u(x_q)$ also. Let $v$ be any utility function representing $\succ_k$ such that $v(x_p) - v(x_q) < u(x_p) - u(x_q)$, and $|v(x) - v(y)| > M$ for all distinct outcomes $x, y \in X$ other than $x_p$ and $x_q$. Because $\succ_k$ ranks $x_p$ and $x_q$ consecutively, we can do this. Then the utility function $v - u$ ranks every pair of outcomes in the same way as $v$ does, except $\{x_p, x_q\}$. Since $V(\succ_{k+1}) = V(\succ_k) \cup \{(p, q)\}$, this means that $v - u$ represents $\succ_{k+1}$.

So, $v \in t_k$ and $v - u \in t_{k+1}$. The local incentive constraints give

$$\begin{aligned} v \cdot (f(t_k) - f(t_{k+1})) &\geq 0, \\ [v - u] \cdot (f(t_{k+1}) - f(t_k)) &\geq 0. \end{aligned}$$

Adding these two gives exactly (3.1), and this completes the proof. □

The hypothesis of successiveness in Proposition 3.4 cannot be dropped, even if the

set of local incentive constraints is modified in the natural way. That is, it is not the case that, for any single-crossing ordinal type space $\{t_1, \ldots, t_r\}$, the set consisting of the incentive constraints $(t_k, t_{k+1})$ and $(t_{k+1}, t_k)$, for $1 \leq k < r$, is sufficient. For a counterexample, consider the three orderings

$$\begin{array}{lllllll} \succ_1\colon & x_1 & \succ_1 & x_2 & \succ_1 & x_3 & \succ_1 \quad x_4 \\ \succ_2\colon & x_2 & \succ_2 & x_1 & \succ_2 & x_3 & \succ_2 \quad x_4 \\ \succ_3\colon & x_4 & \succ_3 & x_2 & \succ_3 & x_1 & \succ_3 \quad x_3 \end{array}$$

and the corresponding ordinal types $t_1, t_2, t_3$. Let $f$ map the types to lotteries over $(x_1, x_2, x_3, x_4)$ as follows:

$$f(t_1) = (1/4, 1/4, 1/2, 0); \qquad f(t_2) = (0, 1/2, 1/2, 0); \qquad f(t_3) = (1/2, 0, 0, 1/2).$$

Then $f$ satisfies the incentive constraints $(t_1, t_2), (t_2, t_1), (t_2, t_3), (t_3, t_2)$, but not $(t_1, t_3)$, so it is not fully incentive-compatible. (The line of the proof of Proposition 3.4 that fails is the statement $V(\succ_{k+1}) \setminus V(\succ_k) = \{(p, q)\}$ in the third paragraph. More broadly, the approach of the proof fails because if we take, say, the utility function $u$ representing $\succ_1$ with $u(x_1) = 4$, $u(x_2) = 3$, $u(x_3) = 2$, $u(x_4) = 1$, then we cannot find any $v$ such that $v$ represents $\succ_2$ and $v - u$ represents $\succ_3$.)

## 3.5 Transfers and interdependent preferences

We now return to the setting of cardinal preferences. However, we generalize in two new directions. First, we consider the transfers setting, in which agents have quasilinear utility in outcomes and money, and a mechanism specifies both a lottery over outcomes and a transfer for each agent. Second, we allow for the possibility of interdependent preferences, where each agent's utility for each outcome depends on the other agents' types. Numerous recent works prove possibility and impossibility results with transfers and interdependence [34, 33, 35, 11], and it is natural to ask to what extent our methods apply here.

We adopt new notations and terminology, for this subsection only, in order to describe these extensions. For clarity, it will help to explicitly write out the dependence of the mechanism on all $n$ agents' types, as in Subsection 2.2. (Of course, a single agent is a special case.) Each agent $i$'s type space $T_i$ is now assumed to be a subset of an arbitrary finite-dimensional Euclidean space, not necessarily $\mathbb{R}^m$. Write $T = T_1 \times \cdots \times T_n$. Agent $i$'s utility is now represented by a function $u_i : T \to \mathbb{R}^m$ specifying his utility for each outcome as a function of the entire type profile. (The private-values case discussed previously is the special case where $T_i \subseteq \mathbb{R}^m$ and $u_i(t_1, \ldots, t_n) = t_i$.)

To allow for transfers, a *mechanism* is now a *pair* $(f, p)$, where $f : T \to \Delta(X)$ specifies a lottery over outcomes for each type profile, and $p : T \to \mathbb{R}^n$ is a *payment function* specifying the net transfer each agent receives. We write $p_i(t)$ for the $i$th component of $p(t)$, representing agent $i$'s transfer.

If the true type profile is $t$ and the agents report profile $t'$, then agent $i$'s realized utility is $u_i(t) \cdot f(t') + p_i(t')$. An *incentive constraint* for agent $i$ is again a pair $(t_i, t_i') \in T_i \times T_i$. We will emphasize here the Bayesian notion of incentive-compatibility, so assume a distribution $\psi_j$ on each agent's type space $T_j$ is given. The mechanism $(f, p)$ *satisfies* the incentive constraint $(t_i, t_i')$ if

$$E_i[u_i(t_i, t_{-i}) \cdot f(t_i, t_{-i}) + p_i(t_i, t_{-i})] \geq E_i[u_i(t_i, t_{-i}) \cdot f(t_i', t_{-i}) + p_i(t_i', t_{-i})].$$

Here the expectations are with respect to the product distribution $\times_{j \neq i} \psi_j$ on other agents' types; it is presupposed that the expressions inside the expectations are measurable in $t_{-i}$, and both expectations are finite. (As in Subsection 2.2, the assumption of independently distributed types is crucial.)

A set $S_i$ of incentive constraints will again be called *local incentive constraints* if every $t_i \in T_i$ has an open neighborhood $N_{t_i}$ in $T_i$ such that $(t_i, t_i') \in S_i$ and $(t_i', t_i) \in S_i$ for all $t_i' \in N_{t_i}$. $S_i$ is *sufficient* for agent $i$ if every mechanism that satisfies it must satisfy the full set of incentive constraints $T_i \times T_i$.

Dominant-strategy incentive-compatibility has an analogue in the interdependent setting, namely *ex post* incentive-compatibility [17, 35], which demands Bayesian

incentive-compatibility for all probability distributions simultaneously. Our result (Proposition 3.5 below) is expressed in terms of Bayesian incentive-compatibility, but an immediate corollary is that the same result holds using ex post incentive-compatibility instead.

To obtain a sufficiency result, we need to restrict interdependence by assuming that, for each fixed $t_{-i}$, the utility function $u_i(\cdot, t_{-i}) : T_i \to \mathbb{R}^m$ is linear in $t_i$. Under this restriction, we have:

**Proposition 3.5** *In the setting with transfers and interdependent utility linear in own type, if agent $i$ has a convex type space $T_i$, then every set of local incentive constraints is sufficient for agent $i$.*

The proof is a straightforward extension of that for Proposition 3.1.

The linearity assumption warrants some comments. It is satisfied trivially in the private-values case (hence, Proposition 3.1 is a special case of Proposition 3.5). It is also satisfied by many concrete models appearing in the interdependent preferences literature; see for example [19, Examples 2,3,4,5], [34], [24] (under Aassumption A2 of that paper), [11, Example 1], and [10, Section 3]. On the other hand, it is quite restrictive, relative to the space of all well-behaved utility functions $u_i : T \to \mathbb{R}^m$ an agent might have.

The linearity assumption is crucial in our analysis, ensuring that the convexity condition in Proposition 3.5 extends that of Proposition 3.1. To understand this, notice (as observed also in [3]) that we can think of each type of agent $i$ as specifying a utility function $X \times T_{-i} \to \mathbb{R}$, and given the priors $\psi_j$, a mechanism induces a distribution over $X \times T_{-i}$ for each $t_i$. In order to apply the argument from the proof of Proposition 3.1, we essentially need agent $i$'s type space to be a convex subset of the linear space of all functions $X \times T_{-i} \to \mathbb{R}$. This is exactly the combination of linearity and convexity assumptions we have made above.

The preceding paragraph does not show that sufficiency fails when the linearity assumption is violated, only that the method of proof used here (adding up incentive constraints along a line) cannot be used. The question of how much further sufficiency

can be generalized is taken up in more extensively in Appendix B, which suggests that sufficiency results do not exist much beyond what can be proven with the present method; as well as Section 4 below, which shows how sufficiency fails under a condition somewhat stronger than nonconvexity.

We have here extended Proposition 3.1 to allow for transfers and/or interdependence with utility linear in own type. The same extension can be applied to Propositions 3.2 and 3.3. Such results may potentially be useful, for example, in analyzing mechanism design problems in such settings when the message spaces are constrained to be finite.

# 4 Insufficiency

The previous section gave numerous classes of type spaces on which local incentive constraints are sufficient. The discussion is not complete without giving some cases where local incentive constraints are not sufficient. We restrict ourselves here to cardinal type spaces. Proposition 4.1 below identifies a large class of such type spaces — roughly, those which violate convexity in a strong enough way — for which we can construct mechanisms that are locally, but not fully, incentive-compatible. (Sato [49, Proposition 4.2]) gives an analogous result for ordinal type spaces.)

It is unclear just how far Proposition 4.1 can be sharpened. Proposition 3.1 showed that if the type space is convex, any local incentive constraints are sufficient, but the converse is not true. The question of exactly characterizing those type spaces $T$ for which all local incentive constraints are sufficient appears to be subtle. This topic is explored further in Appendix B. Proposition B.4 in that appendix gives a nontrivial example of a nonconvex type space for which all local incentive constraints are sufficient; on the other hand, Proposition B.3 gives a kind of converse to Proposition 3.1 for *finite* cardinal type spaces. The details are somewhat technical, so we refer the reader to the appendix, and for now proceed to give our simpler result.

In the space $\mathbb{R}^m$, let $\Pi$ be the subspace of vectors whose sum of components is zero. Let a *fair open half-space* be a set of the form $H = \{u \mid u \cdot \lambda > 0\}$ for some

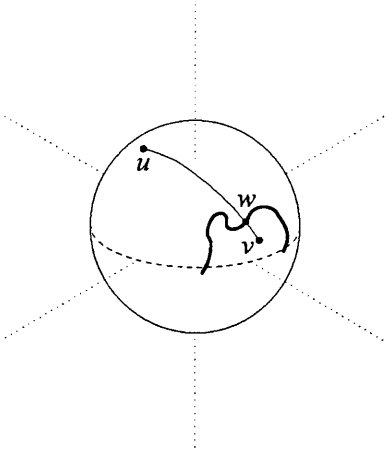


Figure 4.1: Illustration of fair separatedness

nonzero $\lambda \in \Pi$. Say that a cardinal type space $T$ is *fairly separated* if there is some fair open half-space $H$ such that the set $T \cap H$ is not connected.

**Proposition 4.1** *Let $T$ be a cardinal type space that is fairly separated. Then there exists a set of local incentive constraints that is not sufficient.*

Fair separatedness certainly implies nonconvexity. To further indicate the relationship between the two concepts, a little graphical intuition is in order.

For concreteness, suppose $X$ has four elements. By additive and multiplicative renormalization, we can map every utility function either to a point on the unit sphere in the three-dimensional space $\Pi$, or to the origin. This sphere is illustrated in Figure 4.1. The upper hemisphere, whose boundary is shown dashed in the figure, corresponds to a fair open half-space. If $T$ contains the types labeled $u$ and $v$, but does not contain any type along the thick curve (or any type cardinally equivalent to it), then $T$ is fairly separated.

If $T$ were to consist of all possible utility functions except 0 and $w$ (and anything cardinally equivalent to them), then $T$ would be nonconvex. Nonetheless, on this $T$, any local incentive constraints are sufficient; this is just Proposition B.4 in Appendix B. Excluding the whole curve from $T$, rather than just the one point $w$, is enough for insufficiency, by Proposition 4.1.

Fair separatedness might not seem like a natural condition on a type space, so we give one economically important example. Fix an ordering $x_1, \ldots, x_m$ of the outcomes, and let $T$ be the cardinal type space consisting of all quasiconcave utility functions — a cardinal analogue of the single-peaked ordinal type space considered in Section 3.3. Then $T$ is fairly separated. For example, take any $1 \leq p < p' < p'' \leq m$, and let $H = \{u \mid u(x_p) - 2u(x_{p'}) + u(x_{p''}) > 0\}$. If $u \in T \cap H$, then either $u(x_p) > u(x_{p'})$ or $u(x_{p''}) > u(x_{p'})$. So $\{u \in T \cap H \mid u(x_p) > u(x_{p'})\}$ and $\{u \in T \cap H \mid u(x_{p''}) > u(x_{p'})\}$ are two open, nonempty subsets of $T \cap H$, whose union is all of $T \cap H$, and whose intersection is empty (any $u$ satisfying both inequalities would violate quasiconcavity). Hence, $T \cap H$ is not connected. By Proposition 4.1, there are local incentive constraints on $T$ that are insufficient. Note that this result for single-peaked cardinal types contrasts with our sufficiency result for the single-peaked ordinal type space (Proposition 3.3). The accessibility argument underlying that proposition (Lemma A.5 in Appendix A) fails with single-peaked cardinal types.

**Proof of Proposition 4.1:** Let $H = \{u \mid u \cdot \lambda > 0\}$ with $\lambda \in \Pi$. There exist lotteries $L, L'$ on $X$ such that $H = \{u \mid u \cdot L > u \cdot L'\}$. Indeed, let $L$ be any lottery with full support, and just let $L' = L - \delta\lambda$, where $\delta > 0$ is chosen small enough so that all components of $L'$ are still positive.

Now write $T \cap H = T_a \cup T_b$, where $T_a, T_b$ are open, disjoint, and nonempty. Consider the mechanism $f$ defined as follows: if $u \in T_a$, then $f(u) = L$; otherwise, $f(u) = L'$.

Let $S = (T \times T) \setminus (T_b \times T_a)$. This is a set of local incentive constraints: If $u \in T_a$, let $N_u = T_a$; if $u \in T_b$, let $N_u = T_b$; and if $u \in T \setminus H$, let $N_u = T$. In each case we have $(u, u'), (u', u) \in S$ for all $u' \in N_u$.

One readily checks that $f$ satisfies the incentive constraints $S$, but does not satisfy any incentive constraint in $T_b \times T_a$ and so is not fully incentive-compatible. Thus, $S$ is not sufficient. □

A similar construction can be applied in the context of Subsection 3.5, to generate many examples with interdependent preferences, nonlinear in own type, for which local incentive constraints are not sufficient.

# 5 Conclusion

This paper has examined the question of whether or not a small set of local incentive constraints is sufficient to ensure that all other incentive constraints are automatically satisfied, allowing for probabilistic mechanisms. We have obtained affirmative answers in many of the most common mechanism design settings. With convex spaces of cardinal types, local incentive constraints are sufficient to imply full incentive-compatibility. This result allows for monetary transfers under quasilinear utility, and for interdependence so long as each agent's utility is linear in his own type. Local incentive constraints are also sufficient on polyhedral (including ordinal) type spaces with convex closure, as well as single-peaked or successive single-crossing ordinal type spaces. Our proofs follow a unified analytical approach based on a simple supermodularity argument that applies across different settings. For cardinal type spaces that are not convex, the argument does not apply, and with a strengthening of nonconvexity we have insufficiency — there are mechanisms that are locally but not fully incentive-compatible.

The sufficiency results provide an immediate strengthening of many existing impossibility and characterization theorems, and a negative answer to a possible line of inquiry as to whether one could obtain new mechanisms by ignoring nonlocal incentive constraints on grounds of bounded rationality or monitoring technology. Most importantly, they facilitate the technical analysis of mechanism design problems in these settings by ensuring that one can focus on local incentive constraints without any loss, avoiding the need for separate verifications of full incentive-compatibility.

Our analysis on cardinal type spaces in particular also sheds some light on the form of local incentive constraints that should be considered in order to ensure full incentive-compatibility. A naive formulation is not sufficient. On the other hand, our local incentive constraints are still substantially weaker than requiring incentive-compatibility throughout a neighborhood of each type (as required for the formally similar result of [3]), and arguably easier to verify in applications.

# A Omitted proofs

Here we present the proofs that were omitted from the main text. We begin with some technical lemmas.

**Lemma A.1** *Let the polyhedron $t$ be written as an intersection of open half-spaces, $t = \cap_{k=1}^{r} H_k$. Then the boundary of $t$ is*

$$\partial t = \mathrm{cl}(t) \setminus t = \bigcup_{\emptyset \neq K \subseteq \{1,\ldots,r\}} \left[ \left( \bigcap_{k \in K} \partial H_k \right) \cap \left( \bigcap_{k \notin K} H_k \right) \right].$$

**Proof:** Write $\mathrm{cl}(t) = \cap_k \mathrm{cl}(H_k) = \cap_k (\partial H_k \cup H_k)$; distribute on the right side; then remove $t = \cap_k H_k$ from both sides. □

In particular, $\partial t \subseteq \cup_k \partial H_k$.

**Lemma A.2** *Suppose $T$ is a polyhedral type space. Fix $u \in \mathbb{R}^m$, and let $t, t' \in T$ be distinct, nonadjacent types. Then there exist finitely many hyperplanes, each one passing through $u$, whose union contains* $\mathrm{cl}(t) \cap \mathrm{cl}(t')$.

**Proof:** Since $t, t'$ are open and disjoint, neither one can intersect the closure of the other, so $\mathrm{cl}(t) \cap \mathrm{cl}(t') = \partial t \cap \partial t'$. Now suppose $t = \cap_{k=1}^{r} H_k$ and $t' = \cap_{k'=1}^{r'} H'_{k'}$. Applying Lemma A.1 to both $t$ and $t'$, and then distributing the intersection operator, we get

$$\partial t \cap \partial t' = \bigcup_{\substack{\emptyset \neq K \subseteq \{1,\ldots,r\} \\ \emptyset \neq K' \subseteq \{1,\ldots,r'\}}} B(K, K'),$$

where

$$B(K, K') = \left( \bigcap_{k \in K} \partial H_k \right) \cap \left( \bigcap_{k \notin K} H_k \right) \cap \left( \bigcap_{k' \in K'} \partial H'_{k'} \right) \cap \left( \bigcap_{k' \notin K'} H'_{k'} \right).$$

It therefore suffices to show that each set $B(K, K')$ is contained in a hyperplane that passes through $u$. We may assume $B(K, K')$ is nonempty.

$B(K, K')$ is a relatively open subset of $P(K, K') = (\cap_{k \in K} \partial H_k) \cap (\cap_{k' \in K'} \partial H'_{k'})$. The set $P(K, K')$ is an affine set, that is, an intersection of hyperplanes. If $P(K, K')$

is itself a hyperplane, then since $B(K, K') \subseteq \text{cl}(t) \cap \text{cl}(t')$, it follows that $t$ and $t'$ are adjacent. This contradicts the hypothesis. Therefore $P(K, K')$ is an affine set of dimension at most $m - 2$, and we can then find a hyperplane containing both $P(K, K')$ and $u$. □

**Lemma A.3** *Let $T$ be any polyhedral type space, and $u \in \mathbb{R}^m$ any utility function. Then there exists a finite collection $Z$ of hyperplanes such that*

- *for any $t \in T$, $\partial t$ is contained in the union of the hyperplanes in $Z$;*
- *if $t, t'$ are distinct, nonadjacent types, and $w \in \text{cl}(t) \cap \text{cl}(t')$, then some hyperplane in $Z$ passes through both $u$ and $w$.*

**Proof:** Immediate from Lemmas A.1 and A.2. □

As in the text, if $T$ is a polyhedral type space and $u, v$ two utility functions, we say that $v$ is *accessible* from $u$ if $[u, v] \subseteq \cup_{t \in T} \text{cl}(t)$.

**Lemma A.4** *Let $T$ be any polyhedral type space, let $u \in \mathbb{R}^m$, and let $Z$ be the set identified by Lemma A.3. Let $v \in \mathbb{R}^m$ be any utility function not lying on any hyperplane in $Z$. Suppose $v$ is accessible from $u$. Let $t_0, t_1, \ldots, t_r$ be the types intersecting the segment $[u, v]$ in order. Then $t_k$ and $t_{k+1}$ are adjacent, for each $k = 0, \ldots, r-1$.*

Note in the statement that the phrase "in order" makes sense, since each type $t_k$ intersects the segment $[u, v]$ in a subsegment, and these subsegments must be disjoint.

**Proof:** As in the proof of Proposition 3.1, define $u_\alpha = (1-\alpha)u + \alpha v$ for $\alpha \in [0, 1]$. Any hyperplane in $Z$ can contain at most one point of $[u, v]$: otherwise it would contain the entire segment and in particular would contain $u_1 = v$, contradicting the choice of $v$.

As noted above, each type $t_k$ intersects $[u, v]$ in a subsegment $\{u_\alpha \mid \alpha \in J_k\}$, where $J_k$ is an open subinterval of $[0, 1]$ — that is, $J_k$ is of the form $(\gamma, \delta)$, $[0, \delta)$, or $(\gamma, 1]$. Write $\gamma_k = \inf J_k$, $\delta_k = \sup J_k$. By the assumption that the $t_k$ are in order, we have $\delta_k \leq \gamma_{k+1}$ for each $k = 0, \ldots, r-1$.

Next we show that in fact $\delta_k = \gamma_{k+1}$ for each $k$. Suppose instead that $\delta_k < \gamma_{k+1}$. By assumption, the union of the closures of all types in $T$ contains all of $[u, v]$. So

for any $\alpha$ with $\delta_k < \alpha < \gamma_{k+1}$, then, $u_\alpha$ is in the closure of some type $\widehat{t}$. Such a point $u_\alpha$ cannot belong to $\widehat{t}$ proper, so it belongs to $\partial\widehat{t}$. Then $u_\alpha$ belongs to one of the hyperplanes in $Z$. But there are infinitely many choices of $u_\alpha$. Since $Z$ is finite and each hyperplane in $Z$ meets $[u, v]$ at most once, we have a contradiction.

This establishes $\delta_k = \gamma_{k+1}$. On the other hand, $u_{\delta_k} \in \text{cl}(t_k)$, and $u_{\gamma_{k+1}} \in \text{cl}(t_{k+1})$. So, $u_{\delta_k} \in \text{cl}(t_k) \cap \text{cl}(t_{k+1})$. If $t_k, t_{k+1}$ are not adjacent types, then some hyperplane of $Z$ passes through $u_{\delta_k}$ and $u$. This again contradicts the fact that each hyperplane of $Z$ can intersect $[u, v]$ only once. So $t_k, t_{k+1}$ are adjacent. □

**Proof of Proposition 3.2:** Suppose $u \in t$, for some type $t \in T$, and let $t'$ be any other type. Suppose the mechanism $f$ satisfies the local incentive constraints. We wish to show that $u \cdot (f(t) - f(t')) \geq 0$.

Let $Z$ be given by Lemma A.3. Because $t'$ is an open set, we can choose $v \in t'$ not lying on any of the hyperplanes in $Z$. By convexity, $v$ is accessible from $u$, so Lemma A.4 applies, and the successive types $t_k, t_{k+1}$ identified in that lemma are adjacent for each $k$.

Again define $u_\alpha = (1 - \alpha)u + \alpha v$. For each $k = 0, \ldots, r$, pick any $\alpha_k$ with $u_{\alpha_k} \in t_k \cap [u, v]$. The local incentive constraints $(t_k, t_{k+1})$ and $(t_{k+1}, t_k)$ for $k = 0, \ldots, r-1$ ensure that

$$\begin{aligned} u_{\alpha_k} \cdot (f(t_k) - f(t_{k+1})) &\geq 0, \\ u_{\alpha_{k+1}} \cdot (f(t_{k+1}) - f(t_k)) &\geq 0. \end{aligned}$$

From here we proceed exactly as in the proof of Proposition 3.1 to reach the conclusion

$$u \cdot (f(t) - f(t')) = u \cdot (f(t_0) - f(t_r)) \geq 0.$$

□

**Lemma A.5** *Let $T$ be the space of single-peaked ordinal types. Fix $u \in t$ for some $t \in T$. For any $t' \in T$, there exists a nonempty open set contained in $t'$ such that every $v$ in the open set is accessible from $u$.*

**Proof:** Let $u, t'$ be as in the lemma. We know that $u$ is strict (i.e. gives different values to different outcomes) since it belongs to an ordinal type. Then a sufficient condition for $v$ to be accessible from $u$ is that $(1-\alpha)u+\alpha v$ be single-peaked whenever it is strict: the set $\{\alpha \in [0,1] \mid (1-\alpha)u+\alpha v \text{ is not strict}\}$ is finite, hence $(1-\alpha)u+\alpha v$ will be in the closure of some single-peaked ordinal type for each $\alpha$.

We first construct some $v$ that is accessible from $u$. Let $x_p$ be the outcome ranked highest by $u$, and let $x_{p'}$ be the outcome ranked highest by $t'$. If $p' = p$, then any $v \in t'$ is accessible from $u$: since $u(x_q)$ and $v(x_q)$ are both increasing in $q$ for $q \leq p$ and decreasing for $q \geq p$, the same is true of any weighted average $(1-\alpha)u + \alpha v$, so each such weighted average is single-peaked (as long as it is strict).

Now suppose $p' > p$ (the case $p' < p$ is similar). So $u(x_q)$ is decreasing and $v(x_q)$ must be increasing in $q$ for $p \leq q \leq p'$. Choose $v(x_{p'})$ and $v(x_{p'-1})$ arbitrarily, with $v(x_{p'-1}) < v(x_{p'})$. If $p < p' - 1$ then successively choose $v(x_q)$ for $q = p' - 2, p' - 3, \ldots, p'$, such that

$$\frac{v(x_{q+2}) - v(x_{q+1})}{u(x_{q+1}) - u(x_{q+2})} < \frac{v(x_{q+1}) - v(x_q)}{u(x_q) - u(x_{q+1})}. \tag{A.1}$$

This can be done by choosing $v(x_q)$ low enough at each step. Finally, we can choose $v(x_q)$ for $q > p'$ or $q < p$ so that $v$ represents the ordering given by $t'$.

Now we will show that, for $\alpha \in [0,1]$, $(1-\alpha)u + \alpha v$ is single-peaked whenever it is strict. That is, we claim that $(1-\alpha)u(x_q) + \alpha v(x_q)$ is increasing in $q$ for $q$ up to some peak, and decreasing after that. Both $u(x_q)$ and $v(x_q)$ are increasing in $q$ for $q \leq p$, and decreasing in $q$ for $q \geq p'$, so we focus on the range $p \leq q \leq p'$. We will show that there cannot exist any $q \in \{p, \ldots, p' - 2\}$ such that

$$(1-\alpha)u(x_q) + \alpha v(x_q) > (1-\alpha)u(x_{q+1}) + \alpha v(x_{q+1}) \tag{A.2}$$

and

$$(1-\alpha)u(x_{q+1}) + \alpha v(x_{q+1}) < (1-\alpha)u(x_{q+2}) + \alpha v(x_{q+2}) \tag{A.3}$$

simultaneously hold; this will prove the claim.

Suppose (A.2) and (A.3) do both hold, for some $q$. (A.2) implies

$$\frac{1-\alpha}{\alpha} > \frac{v(x_{q+1}) - v(x_q)}{u(x_q) - u(x_{q+1})}$$

while (A.3) implies

$$\frac{1-\alpha}{\alpha} < \frac{v(x_{q+2}) - v(x_{q+1})}{u(x_{q+1}) - u(x_{q+2})}.$$

(Note we used the fact that $u(x_q) > u(x_{q+1}) > u(x_{q+2})$ to make sure the signs don't switch when we divide. We know $\alpha > 0$ since (A.3) is violated at $\alpha = 0$.) Combining these two inequalities gives a contradiction of (A.1), completing the proof of the claim.

At this point we have shown that any $v \in t'$ satisfying the inequalities (A.1) is accessible from $u$. Since these inequalities carve out a nonempty open subset of $t'$, the lemma is proven. □

**Proof of Proposition 3.3:** Suppose that $T$ is the space of single-peaked ordinal types. Let $u \in t$, for some $t \in T$, and let $t'$ be any other type. We wish to show that $u \cdot (f(t) - f(t')) \geq 0$, for any $f$ satisfying the local incentive constraints.

Let $Z$ again be the set of hyperplanes promised to us by Lemma A.3 (with respect to $T$ and $u$). By Lemma A.5, we can choose a $v \in t'$ that is accessible from $u$ and does not lie on any of the hyperplanes in $Z$. Accessibility ensures that Lemma A.4 applies. From here onward we just repeat the argument used to prove Proposition 3.2. □

**Proof of Proposition 3.5:** Suppose that the mechanism $(f, p)$ satisfies the set $S_i$ of local incentive constraints for agent $i$. Consider any two types $t_i, t_i' \in T_i$. Write $t_\alpha = (1-\alpha)t_i + \alpha t_i'$. As in the proof of Proposition 3.1, we can find $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_r = 1$ with $(t_{\alpha_k}, t_{\alpha_{k+1}}), (t_{\alpha_{k+1}}, t_{\alpha_k}) \in S_i$ for each $k = 0, \ldots, r-1$. These local incentive constraints give

$$\begin{aligned} E_i[u_i(t_{\alpha_k}, t_{-i}) \cdot (f(t_{\alpha_k}, t_{-i}) - f(t_{\alpha_{k+1}}, t_{-i})) + (p_i(t_{\alpha_k}, t_{-i}) - p_i(t_{\alpha_{k+1}}, t_{-i}))] &\geq 0, \\ E_i[u_i(t_{\alpha_{k+1}}, t_{-i}) \cdot (f(t_{\alpha_{k+1}}, t_{-i}) - f(t_{\alpha_k}, t_{-i})) + (p_i(t_{\alpha_{k+1}}, t_{-i}) - p_i(t_{\alpha_k}, t_{-i}))] &\geq 0. \end{aligned}$$

Multiply by $\alpha_{k+1}$ and $\alpha_k$, respectively, and add:

$$E_i\left[[\alpha_{k+1}u_i(t_{\alpha_k}, t_{-i}) - \alpha_k u_i(t_{\alpha_{k+1}}, t_{-i})] \cdot (f(t_{\alpha_k}, t_{-i}) - f(t_{\alpha_{k+1}}, t_{-i})) + [\alpha_{k+1} - \alpha_k] \cdot (p_i(t_{\alpha_k}, t_{-i}) - p_i(t_{\alpha_{k+1}}, t_{-i}))\right] \geq 0. \quad \text{(A.4)}$$

Because utility is linear in own type, and $t_{\alpha_k}$ is equal to the weighted average $(\alpha_k/\alpha_{k+1})t_{\alpha_{k+1}} + (1 - \alpha_k/\alpha_{k+1})t_i$, we know that for each realization of $t_{-i}$,

$$u_i(t_{\alpha_k}, t_{-i}) = \frac{\alpha_k}{\alpha_{k+1}} u_i(t_{\alpha_{k+1}}, t_{-i}) + \left(1 - \frac{\alpha_k}{\alpha_{k+1}}\right) u_i(t_i, t_{-i}).$$

Rearranging gives

$$\alpha_{k+1}u_i(t_{\alpha_k}, t_{-i}) - \alpha_k u_i(t_{\alpha_{k+1}}, t_{-i}) = (\alpha_{k+1} - \alpha_k)u_i(t_i, t_{-i}).$$

Applying this identity and dividing through (A.4) by the constant $\alpha_{k+1} - \alpha_k > 0$ gives

$$E_i[u_i(t_i, t_{-i}) \cdot (f(t_{\alpha_k}, t_{-i}) - f(t_{\alpha_{k+1}}, t_{-i})) + (p_i(t_{\alpha_k}, t_{-i}) - p_i(t_{\alpha_{k+1}}, t_{-i}))] \geq 0.$$

Summing over $k = 0, \ldots, r-1$ gives

$$E_i[u_i(t_i, t_{-i}) \cdot (f(t_i, t_{-i}) - f(t'_i, t_{-i})) + (p_i(t_i, t_{-i}) - p_i(t'_i, t_{-i}))] \geq 0$$

which shows that the incentive constraint $(t_i, t'_i)$ is satisfied.

□

# B On proofs by adding up

This appendix gives a more detailed study of conditions under which the basic method of proof used for the sufficiency results in the main text can be applied, with an eye to understanding how much the method might potentially be further generalized, and whether the results still hold when the method does not apply. We restrict ourselves

to cardinal type spaces and no transfers, as in Subsection 3.1.

All of the proofs of sufficiency results in the main text follow the general method of showing that the linear inequality corresponding to any desired incentive constraint can be obtained by adding up inequalities corresponding to local incentive constraints. We show here that for *finite* type spaces, whenever a set $S$ of incentive constraints is sufficient, there exists a proof of sufficiency by adding up (Lemma B.1 below). Moreover, with minor exceptions, whenever an incentive constraint $(u, v)$ is provable by adding up, there exists such a proof that uses only types along the line segment $[u, v]$, or types cardinally equivalent to them (Proposition B.3). The conclusion, then, is that for finite type spaces, there exist essentially no sufficiency results beyond those that can be proven using the method of Proposition 3.1.

On the other hand, for *infinite* type spaces, the conclusions are not as tight. We give an example (Proposition B.4) of a type space where local incentive constraints are sufficient, but sufficiency cannot be proven by adding up. In that example, we prove sufficiency by a combination of adding-up arguments and limiting arguments exploiting the compactness of the space $\Delta(X)$.

To begin the investigation, we must first be precise about what it means for an incentive constraint to be provable by adding up other constraints. Let $T$ be a cardinal type space, and let $S$ be a set of incentive constraints. Let $\mathbf{1} \in \mathbb{R}^m$ denote the vector all of whose components are 1, and let $e_p$ denote the $p$th unit vector for $p = 1, \ldots, m$. For any mechanism $f$, we have

$$\mathbf{1} \cdot f(u) = 1 \tag{B.1}$$

for all $u \in T$, and

$$e_p \cdot f(u) \geq 0 \tag{B.2}$$

for $p = 1, \ldots, m$ and all $u \in T$. If $f$ satisfies $S$, then we also have

$$u \cdot (f(u) - f(v)) \geq 0 \tag{B.3}$$

for each $(u, v) \in S$.

We say that an incentive constraint $(u^*, v^*) \in T \times T$ is *provable from $S$ by adding up* if the inequality

$$u^* \cdot (f(u^*) - f(v^*)) \geq 0 \tag{B.4}$$

can be obtained as a finite linear combination of the equations (B.1) and inequalities (B.2), (B.3), with nonnegative coefficients on the inequalities. That is, $(u^*, v^*)$ is provable from $S$ by adding up if there exist real numbers

- $a_u$ for $u \in T$,
- $b_{pu}$ for $p = 1, \ldots, m$, $u \in T$, and
- $c_{uv}$ for $(u, v) \in S$,

such that all but finitely many of these numbers are zero, all the $b_{pu}$ and $c_{uv}$ are nonnegative, and such that adding up $a_u$ times (B.1), $b_{pu}$ times (B.2), and $c_{uv}$ times (B.3) gives (B.4). (For notational convenience, we will assume $c_{uv}$ to be defined for all $u, v \in T$, with $c_{uv} = 0$ whenever $(u, v) \notin S$.)

We can write out the adding-up conditions explicitly, by comparing coefficients of $f(u)$, for each $u \in T$. Assume $u^* \neq v^*$ (otherwise (B.4) just reads $0 = 0$ which is trivially provable by adding up). Then the adding-up condition says that for each $u$, we have

$$a_u \mathbf{1} + \sum_{p=1}^{m} b_{pu} e_p + \sum_{v \in T} c_{uv} u - \sum_{v \in T} c_{vu} v = \begin{cases} u^* & \text{if } u = u^* \\ -u^* & \text{if } u = v^* \\ 0 & \text{otherwise.} \end{cases} \tag{B.5}$$

Also, for the constant terms, the adding-up condition is simply

$$\sum_{u \in T} a_u = 0. \tag{B.6}$$

We say that the set $S$ of incentive constraints *implies* the incentive constraint $(u^*, v^*) \in T \times T$ if every mechanism that satisfies $S$ also satisfies $(u^*, v^*)$.

The present question is: If $S$ implies $(u^*, v^*)$, must the constraint $(u^*, v^*)$ necessarily be provable from $S$ by adding up? When $S$ is finite, the answer is affirmative; this is just a form of the theorem of the alternative.

**Lemma B.1** *If $T$ is a cardinal type space and $S$ a finite set of incentive constraints that implies the incentive constraint $(u^*, v^*)$, then $(u^*, v^*)$ is provable from $S$ by adding up.*

**Proof:** We may as well assume that $T$ consists only of $u^*$, $v^*$, and the types that appear in constraints of $S$. Thus, $T$ is finite. A mechanism $f$ satisfying $S$ then consists simply of a choice of $m \cdot |T|$ real numbers — the components of the $|T|$ vectors $f(u)$ for $u \in T$ — satisfying (B.1), (B.2), and also (B.3) for $(u, v) \in S$. The hypothesis is that any such numbers must also satisfy (B.4).

This can be recast as a linear programming statement: for any choice of $m \cdot |T|$ real numbers satisfying the nonnegativity constraints (B.2) and the linear equations (B.1) and inequalities (B.3), the minimum value of the linear function $u^* \cdot (f(u^*) - f(v^*))$ is 0. (This minimum is attained, for example, by any mechanism such that $f(u)$ is constant across all $u$.) The duality theorem of linear programming then tells us that (B.4) is expressible as a linear combination of (B.1), (B.2), (B.3), with nonnegative coefficients on the inequalities. That is, $(u^*, v^*)$ is provable from $S$ by adding up. □

To proceed further, it will be helpful to have an alternative, cleaner definition of provability by adding up. Let $\Pi \subseteq \mathbb{R}^m$ be the hyperplane orthogonal to $\mathbf{1}$, as in Section 4. For any $u \in \mathbb{R}^m$, let $\overline{u}$ denote its orthogonal projection onto $\Pi$.

**Lemma B.2** *Assume $u^* \neq v^*$. Then $(u^*, v^*)$ is provable from $S$ by adding up if and only if there exist numbers $c_{uv} \geq 0$, finitely many of which are nonzero, such that the equation*

$$\sum_{v \in T} c_{uv}\overline{u} - \sum_{v \in T} c_{vu}\overline{v} = \begin{cases} \overline{u^*} & \text{if } u = u^* \\ -\overline{u^*} & \text{if } u = v^* \\ 0 & \text{otherwise.} \end{cases} \tag{B.7}$$

*holds for each $u \in T$, and $c_{uv} = 0$ unless $(u, v) \in S$.*

**Proof:** First suppose that $(u^*, v^*)$ is provable from $S$ by adding up under the original definition; let $a_u$, $b_{pu}$, $c_{uv}$ be the coefficients satisfying (B.5). By summing (B.5) over all choices of $u$ we get $\sum_u a_u \mathbf{1} + \sum_p \sum_u b_{pu} e_p = 0$. (On the left side, each $c_{uv}$ occurs once multiplied by $u$ and once multiplied by $-u$. On the right side, we get one $u^*$, one $-u^*$, and all zeroes otherwise.) From (B.6), this reduces to $\sum_p \sum_u b_{pu} e_p = 0$. Since the $b_{pu}$ are nonnegative, they must all be zero. Once we know this, then, taking (B.5) and projecting orthogonally onto $\Pi$ gives (B.7).

Conversely, suppose there are coefficients $c_{uv}$ satisfying (B.7). Put $b_{pu} = 0$ for all $p$ and all $u$. Note that (B.7) implies that for each $u$, the expression

$$\begin{cases} \sum_v c_{uv} u - \sum_v c_{vu} v - u^* & \text{if } u = u^* \\ \sum_v c_{uv} u - \sum_v c_{vu} v + u^* & \text{if } u = v^* \\ \sum_v c_{uv} u - \sum_v c_{vu} v & \text{otherwise} \end{cases}$$

must be some multiple of $\mathbf{1}$. Choose $a_u$ so that this expression is equal to $-a_u \mathbf{1}$. Then it is immediate that (B.5) is satisfied for each $u$. Moreover, summing (B.5) across all $u \in T$, the $c_{uv}$ terms cancel as in the previous paragraph, and we are simply left with $\sum_u a_u \mathbf{1} = 0$; hence, with this choice of $a_u$, (B.6) is satisfied as well. Finally, $a_u \neq 0$ only when $u = u^*, v^*$ or when $c_{uv}$ or $c_{vu}$ is nonzero for some $v$; thus, only finitely many of the $a_u$ are nonzero. Thus, the original definition of provability by adding up is satisfied. □

We need just a few more definitions. Say that two types $u, v$ are *equivalent* if $v = \alpha u + \beta \mathbf{1}$ for some $\alpha, \beta \in \mathbb{R}$, $\alpha > 0$, and a type is *indifferent* if it is equivalent to 0. For $u^*, v^* \in T$, let $T_{[u^*, v^*]}$ be the set of all types in $T$ that are equivalent to some type on the segment $[u^*, v^*]$, and let

$$S_{[u^*, v^*]} = \{(u, v) \in S \mid u, v \in T_{[u^*, v^*]}\}.$$

We now arrive at the main result of this appendix.

**Proposition B.3** *Let $T$ be a cardinal type space and $S$ a set of incentive constraints such that $(u^*, v^*)$ is provable from $S$ by adding up. Assume that $v^*$ is not equivalent*

*to $-u^*$. Then $(u^*, v^*)$ is provable from $S_{[u^*,v^*]}$ by adding up.*

This result says that if an incentive constraint $(u^*, v^*)$ can be proved by adding up constraints in $S$, then it can be proved by adding up in a way that only uses types equivalent to convex combinations of $u^*$ and $v^*$. Thus, the method used to prove Proposition 3.1 is (almost) the only possible adding-up argument.

The proof of Proposition B.3 is a bit long, but the main idea is straightforward. It consists of taking the coefficients $c_{uv}$ satisfying (B.7) and successively replacing them by zeroes, checking that (B.7) still holds at each step, until only constraints in $S_{[u^*,v^*]}$ have nonzero coefficients.

**Proof:** We may assume that $u^*$ is not indifferent, since otherwise the conclusion is immediate: (B.7) holds with all $c_{uv}$ equal to 0. We also assume $u^* \neq v^*$; otherwise the conclusion is again trivial.

Let $c_{uv}$ be the coefficients satisfying (B.7), with $c_{uv} > 0$ only if $(u, v) \in S$. We may as well assume that $S$ consists only of the (finitely many) incentive constraints $(u, v)$ for which $c_{uv} > 0$, and $T$ consists only of the types appearing in these constraints.

Now consider any fixed vector $w \in \Pi$ with the following properties:

(i) $w \cdot \overline{u^*} > 0$;

(ii) $w \cdot \overline{v^*} \geq 0$;

(iii) if $u \in T$ and $w \cdot \overline{u} = 0$, then $\overline{u} = 0$.

We claim that if $(u, v) \in S$ such that either

(a) $w \cdot \overline{u} > 0$ and $w \cdot \overline{v} < 0$, or

(b) $w \cdot \overline{u} > 0$ and $w \cdot \overline{v} = 0$ and $v \neq v^*$, or

(c) $w \cdot \overline{u} < 0$ and $w \cdot \overline{v} \geq 0$,

then $c_{uv} = 0$.

Proof: Consider any $u \in T$ such that $w \cdot \overline{u} < 0$. Take the dot product of $w$ with (B.7). We get

$$\sum_{v \in T} c_{uv}(w \cdot \overline{u}) - \sum_{v \in T} c_{vu}(w \cdot \overline{v}) = 0$$

(note that $u \neq u^*, v^*$). Now sum over all $u$ such that $w \cdot \overline{u} < 0$. For each incentive constraint $(u, v) \in S$ such that $w \cdot \overline{u} < 0$ and $w \cdot \overline{v} < 0$, the term $c_{uv}(w \cdot \overline{u})$ appears once with a $+$ sign and once with a $-$ sign, so these cancel out. The remaining terms give us

$$\sum_{w \cdot \overline{u} < 0;\ w \cdot \overline{v} \geq 0} c_{uv}(w \cdot \overline{u}) - \sum_{w \cdot \overline{u} < 0;\ w \cdot \overline{v} \geq 0} c_{vu}(w \cdot \overline{v}) = 0.$$

Since each $c_{uv}$ is nonnegative, every term in the first sum is $\leq 0$ and every term in the second sum is $\geq 0$. Hence, every term must be equal to zero. This implies that whenever $w \cdot \overline{u} < 0$ and $w \cdot \overline{v} \geq 0$, $c_{uv} = 0$, and when moreover $w \cdot \overline{v} > 0$, we also have $c_{vu} = 0$.

This covers (a) and (c). For (b), when $w \cdot \overline{v} = 0$ and $v \neq v^*$, (B.7) for $v$ gives $\sum_u c_{vu}\overline{v} - \sum_u c_{uv}\overline{u} = 0$. Dotting with $w$ gives $\sum_u c_{uv}(w \cdot \overline{u}) = 0$ (after canceling). We have already established that $c_{uv} = 0$ if $w \cdot \overline{u} < 0$, so all the terms on the left are nonnegative, and hence they must all be zero. So $c_{uv} = 0$ whenever $w \cdot \overline{u} > 0$.

This proves the claim.

Next, for each $u, v \in T$, define $c'_{uv} = c_{uv}$ if $w \cdot \overline{u} \geq 0$ and $w \cdot \overline{v} \geq 0$; and $c'_{uv} = 0$ otherwise. Then we again have, for each $u$,

$$\sum_v c'_{uv}\overline{u} - \sum_v c'_{vu}\overline{v} = \begin{cases} \overline{u^*} & \text{if } u = u^* \\ -\overline{u^*} & \text{if } u = v^* \\ 0 & \text{otherwise.} \end{cases} \tag{B.8}$$

Proof: If $u$ is such that $w \cdot \overline{u} < 0$ then (B.8) is trivial since both sides are zero. If $w \cdot \overline{u} > 0$, then the left side of (B.8) differs from the left side of (B.7) by the terms $c_{uv}\overline{u}$ and $-c_{vu}\overline{v}$ for $w \cdot \overline{v} < 0$. These are all zero, by cases (a) and (c) of the claim, respectively; thus (B.8) follows from (B.7). If $w \cdot \overline{u} = 0$ and $u \neq v^*$, then again all the left-hand-side terms of (B.8) are zero:

- all the $c'_{uv}\overline{u}$ are zero because $\overline{u} = 0$, by condition (iii) on $w$;
- $c'_{vu}\overline{v} = 0$ for $w \cdot \overline{v} > 0$ by (b) of the claim;
- $c'_{vu}\overline{v} = 0$ for $w \cdot \overline{v} = 0$ again by (iii) on $w$;

- $c'_{vu}\overline{v} = 0$ for $w \cdot \overline{v} < 0$ by definition of $c'_{vu}$.

So both sides of (B.8) are zero, and it again holds.

Thus, (B.8) is verified for all $u$ except possibly for $u = v^*$. But summing (B.8) over all $u \in T$ gives the identity $0 = 0$; so if it holds for all $u$ except $u = v^*$, it must hold for $u = v^*$ as well.

At this point, we have shown the following: If we start with coefficients $c_{uv}$ for which (B.7) holds, pick any $w \in \Pi$ satisfying (i)-(iii), and replace $c_{uv}$ by 0 whenever $w \cdot \overline{u} < 0$ or $w \cdot \overline{v} < 0$, then (B.7) still holds.

If we find any finite set of vectors $w_1, \ldots, w_q \in \Pi$, each satisfying conditions (i)-(iii), and for each $w_k$ we successively replace $c_{uv}$ by 0 whenever $w_k \cdot \overline{u} < 0$ or $w_k \cdot \overline{v} < 0$, then the resulting coefficients will still satisfy (B.7).

Now let $T_{[u^*,v^*]+}$ consist of the types in $T_{[u^*,v^*]}$ together with all indifferent types (alternatively stated, all types that are equivalent to $\alpha u^* + \beta v^*$ for some $\alpha, \beta \geq 0$); and let $S_{[u^*,v^*]+} = \{(u,v) \in S \mid u, v \in T_{[u^*,v^*]+}\}$. We will show that, for any $u \in T$ that is not in $T_{[u^*,v^*]+}$, there is some $w \in \Pi$ satisfying (i)-(iii) with $w \cdot \overline{u} < 0$. If we consider each such $w$ in turn, and successively replace $c_{uv}$'s by 0 as in the previous paragraph, we will be left with coefficients $c_{uv} \geq 0$ that still satisfy (B.7), and such that $c_{uv} = 0$ unless $u, v \in T_{[u^*,v^*]+}$. Therefore, we will have shown that $(u,v)$ is provable from $S_{[u^*,v^*]+}$ by adding up.

Thus, consider any $u \in T \setminus T_{[u^*,v^*]+}$. We wish to show that there exists $w \in \Pi$ satisfying (i)-(iii) with $w \cdot \overline{u} < 0$. The assumptions that $v^*$ is not equivalent to $-u^*$ and $u^*$ is not indifferent imply that there exists $w' \in \Pi$ with

$$w' \cdot \overline{u^*} > 0, \qquad w' \cdot \overline{v^*} \geq 0$$

and the latter inequality holding strictly unless $\overline{v^*} = 0$. The assumption $u \notin T_{[u^*,v^*]+}$ implies that $\overline{u}$ is not a nonnegative combination of $\overline{u^*}$ and $\overline{v^*}$; hence there is some $w'' \in \Pi$ such that

$$w'' \cdot \overline{u^*} \geq 0, \qquad w'' \cdot \overline{v^*} \geq 0, \qquad w'' \cdot \overline{u} < 0.$$

Taking $w = w' + \kappa w''$ for large $\kappa$ will give (i), (ii), and $w \cdot \overline{u} < 0$. Finally, by perturbing $w$ slightly, we can ensure $w \cdot \overline{v} \neq 0$ for all $v \in T$, $\overline{v} \neq 0$, without breaking any of the strict inequalities; thus we get (iii) as well.

At this point we have finished showing that $(u^*, v^*)$ is provable from $S_{[u^*,v^*]+}$ by adding up.

If $v^*$ is indifferent, then $S_{[u^*,v^*]+} = S_{[u^*,v^*]}$ and so we are done. Otherwise, we have to do just a little more work.

Let $c_{uv}$ now be the coefficients used to prove $(u^*, v^*)$ from $S_{[u^*,v^*]+}$ by adding up (i.e. the coefficients satisfying (B.7)). Whenever $\overline{u} = 0$, we can replace $c_{uv}$ by 0 without affecting the validity of (B.7) (since $c_{uv}$ only ever appears as part of the product $c_{uv}\overline{u}$). So we may assume $c_{uv} = 0$ whenever $u$ is indifferent.

Since $u^*, v^*$ are both non-indifferent and $v^*$ is not equivalent to $-u^*$, we can find $w \in \Pi$ such that $w \cdot \overline{u^*} > 0$ and $w \cdot \overline{v^*} > 0$. Thus, for any element of $T_{[u^*,v^*]+}$ that is not indifferent, its projection has positive dot product with $w$.

Now for any indifferent $u$, taking (B.7) and dotting with $w$ gives $-\sum_v c_{vu}(w \cdot \overline{v}) = 0$. Each term in the sum is nonnegative, so they must all be zero. Hence $c_{vu} = 0$ whenever $\overline{v}$ has positive dot product with $w$; and the remaining $v \in T_{[u^*,v^*]+}$ are indifferent, so $c_{vu} = 0$ for them too by assumption. Thus, if $u$ is indifferent then $c_{uv}, c_{vu} = 0$ for all $v$.

But this means that (B.7) holds with $c_{uv}$ zero unless $u, v \in T_{[u^*,v^*]}$, so in fact $(u^*, v^*)$ is provable from $S_{[u^*,v^*]}$ by adding up.

□

Proposition B.3 is stated as a description of the form of proofs by adding up. However, it also provides us with a tool to show show when a particular constraint is *not* provable by adding up. In particular, we can apply it to give an example of an infinite type space and a set of local incentive constraints that are sufficient, but whose sufficiency cannot be proven by adding up, as promised at the beginning of this appendix. In fact, we will give a stronger example: a type space such that *any* set of local incentive constraints is sufficient, yet there exist fairly large such sets whose sufficiency cannot be proven by adding up.

Let $X$ have four elements, and let $w$ be some utility function on $X$ that is not indifferent. Let $T_{w^+}$ be the set of all cardinal types that are either indifferent or equivalent to $w$, and let $T = \mathbb{R}^4 \setminus T_{w^+}$ be the set of cardinal types not in $T_{w^+}$. Say that two types $u, v \in T$ are *$T_{w^+}$-opposed* if $[u, v] \cap T_{w^+} \neq \emptyset$. Let $S$ be any set of local incentive constraints such that if $u$ and $v$ are $T_{w^+}$-opposed, then $(u, v) \notin S$.

This requirement on $S$ can be easily satisfied. Indeed, for each $u \in T$, start with any neighborhood $N_u$, and let $d(u, T_{w^+}) > 0$ be the Euclidean distance from $u$ to $T_{w^+}$. Then the set $N'_u = \{v \in N_u \mid d(u, v) < d(u, T_{w^+})\}$ is again an open neighborhood of $u$, not containing any types $T_{w^+}$-opposed to $u$. So $S = \{(u, v) \mid u \in N'_v \text{ or } v \in N'_u\}$ is a set of local incentive constraints meeting our requirement.

**Proposition B.4** *With $T, S$ as above, $S$ is sufficient. However, for any $u^*, v^* \in T$ that are $T_{w^+}$-opposed, with $u^*$ not equivalent to $-v^*$, the constraint $(u^*, v^*)$ is not provable from $S$ by adding up.*

**Proof:** First we show that $S$ is sufficient. Let $f$ be any mechanism that satisfies $S$. For any possible incentive constraint $(u, v)$, if $u$ and $v$ are not $T_{w^+}$-opposed, then the entire line segment from $u$ to $v$ is contained in $T$. Therefore, the usual argument from Proposition 3.1 of the main text shows that $f$ satisfies $(u, v)$.

So we need only deal with the case where $u, v$ are $T_{w^+}$-opposed. In this case, notice that we can choose $u_k \in T$ arbitrarily close to $(u + v)/2$ such that $u_k$ is not $T_{w^+}$-opposed to either $u$ or $v$. (Any type $T_{w^+}$-opposed to $u$ must lie on the hyperplane $\Pi_{uw}$ generated by $u$, $w$, and $\mathbf{1}$. Similarly, any type $T_{w^+}$-opposed to $v$ must lie on the hyperplane generated by $v, w, \mathbf{1}$, which is again $\Pi_{uw}$. There are types in $T$ arbitrarily close to $(u + v)/2$ not lying on this hyperplane.) For any such $u_k$, then, we have already shown that $f$ satisfies the constraints $(u, u_k)$, $(v, u_k)$, $(u_k, v)$; that is:

$$u \cdot (f(u) - f(u_k)) \geq 0, \tag{B.9}$$

$$v \cdot (f(v) - f(u_k)) \geq 0, \tag{B.10}$$

$$u_k \cdot (f(u_k) - f(v)) \geq 0. \tag{B.11}$$

So we can choose a sequence of types $u_1, u_2, \ldots$ in $T$ with $u_k \to (u+v)/2$, such that (B.9)-(B.11) are satisfied for each $u_k$. Moreover, because the image of $f$ is contained in the compact set $\Delta(X)$, we may assume by passing to a subsequence that $f(u_k)$ converges to some limit $f^*$. Then, taking limits, we get

$$u \cdot (f(u) - f^*) \geq 0, \tag{B.12}$$

$$v \cdot (f(v) - f^*) \geq 0, \tag{B.13}$$

$$\frac{u+v}{2} \cdot (f^* - f(v)) \geq 0. \tag{B.14}$$

Adding (B.12), (B.13), and twice (B.14) gives

$$u \cdot (f(u) - f(v)) \geq 0$$

so $f$ satisfies the constraint $(u, v)$.

This shows that $S$ is sufficient.

It remains to prove that if $u^*, v^* \in T$ are $T_{w^+}$-opposed, and $u^*$ is not equivalent to $-v^*$, then $(u^*, v^*)$ is not provable from $S$ by adding up. By Proposition B.3, if $(u^*, v^*)$ were provable from $S$ by adding up, then it would be provable from $S_{[u^*,v^*]}$ by adding up. So we just need to show that the latter is not the case.

For any $\alpha \in [0,1]$, let $u_\alpha = (1-\alpha)u^* + \alpha v^*$. Let $\alpha^* \in (0,1)$ be such that $u_{\alpha^*} \in T_{w^+}$. Notice that if $u, v$ are equivalent to $u_\alpha, u_\beta$ respectively, and $(u, v) \in S$, then $\alpha, \beta$ are either both less than $\alpha^*$ or both greater than $\alpha^*$: otherwise $u, v$ are $T_{w^+}$-opposed.

Suppose that $(u^*, v^*)$ is provable from $S_{[u^*,v^*]}$ by adding up. Let $c_{uv}$ be the coefficients that satisfy (B.7). Let $T_<$ be the set of types in $T_{[u^*,v^*]}$ that are equivalent to some $u_\alpha$ for $\alpha < \alpha^*$. The observation of the previous paragraph implies that if $c_{uv} > 0$, and one of $u, v$ is in $T_<$, then the other is as well.

Sum up (B.7) over all $u \in T_<$. The $c_{uv}\overline{u}$ terms on the left side appear in pairs of opposite sign, which cancel; thus we are left with $0 = \overline{u^*}$. Since $u^* \in T$ cannot be indifferent, we have a contradiction. $\square$

# Bibliography


[1] Atila Abdulkadiroğlu, A., Parag A. Pathak, and Alvin E. Roth (2005), "The New York City High School Match," *American Economic Review* 95 (2), 364-367.

[2] José Alcalde and Salvador Barberà (1994), "Top Dominance and the Possibility of Strategy-Proof Stable Solutions to Matching Problems," *Economic Theory* 4 (3), 417-435.

[3] Aaron Archer and Robert Kleinberg (2008), "Truthful Germs are Contagious: A Local to Global Characterization of Truthfulness," in *Proceedings of the 9th ACM Conference on Electronic Commerce* (EC-08), 21-30.

[4] Itai Ashlagi, Mark Braverman, Avinatan Hassidim, and Dov Monderer (2010), "Monotonicity and Implementability," *Econometrica* 78 (5), 1749-1772.

[5] Salavdor Barberà, Anna Bogomolnaia, and Hans van der Stel (1998), "Strategy-Proof Probabilistic Rules for Expected Utility Maximizers," *Mathematical Social Sciences* 35 (2), 89-103.

[6] Salavdor Barberà, Matthew O. Jackson, and Alejandro Neme (1997), "Strategy-Proof Allotment Rules," *Games and Economic Behavior* 18 (1), 1-21.

[7] Salvador Barberà, Hugo Sonnenschein, and Lin Zhou (1991), "Voting by Committees," *Econometrica* 59 (3), 595-609.

[8] John J. Bartholdi III and James B. Orlin (1991), "Single Transferable Vote Resists Strategic Voting," *Social Choice and Welfare* 8 (4), 341-354.

[9] J. J. Bartholdi III, C. A. Tovey, and M. A. Trick (1989), "The Computational Difficulty of Manipulating an Election," *Social Choice and Welfare* 6 (3), 227-241.

[10] Dirk Bergemann and Stephen Morris (2009), "Robust Implementation in Direct Mechanisms," *Review of Economic Studies* 76 (4), 1175-1204.

[11] Sushil Bikhchandani (2006), "Ex Post Implementation in Environments with Private Goods," *Theoretical Economics* 1 (3), 369-393.

[12] Sushil Bikhchandani, Shurojit Chatterji, Ron Lavi, Ahuva Mu'alem, Noam Nisan, and Arunava Sen (2006), "Weak Monotonicity Characterizes Deterministic Dominant-Strategy Implementation," *Econometrica* 74 (4), 1109-1132.

[13] Duncan Black (1948), "On the Rationale of Group Decision-Making," *Journal of Political Economy* 56 (1), 23-34.

[14] Anna Bogomolnaia and Hervé Moulin (2001), "A New Solution to the Random Assignment Problem," *Journal of Economic Theory* 100 (2), 295-328.

[15] Anna Bogomolnaia and Hervé Moulin (2002), "A Simple Random Assignment Problem with a Unique Solution," *Economic Theory* 19 (3), 623-635.

[16] Gorken Celik (2006), "Mechanism Design with Weaker Incentive Compatibility Constraints," *Games and Economic Behavior* 56 (1), 37-44.

[17] Kim-Sau Chung and Jeffrey C. Ely (2006), "Ex-Post Incentive Compatible Mechanism Design," unpublished paper, Northwestern University.

[18] Vladimir I. Danilov (1994), "The Structure of Non-Manipulable Social Choice Rules on a Tree," *Mathematical Social Sciences* 27 (2), 123-131.

[19] Partha Dasgupta and Eric Maskin (2000), "Efficient Auctions," *Quarterly Journal of Economics* 115 (2), 341-388.

[20] Gabrielle Demange (1982), "Single-Peaked Orders on a Tree," *Mathematical Social Sciences* 3 (4), 389-396.

[21] John Duggan (1996), "A Geometric Proof of Gibbard's Random Dictatorship Theorem," *Economic Theory* 7 (2), 365-369.

[22] Lars Ehlers and Bettina Klaus (2003), "Probabilistic Assignments of Identical Indivisible Objects and Uniform Probabilistic Rules," *Review of Economic Design* 8 (3), 249-268.

[23] Lars Ehlers, Hans Peters, and Ton Storcken (2002), "Strategy-Proof Probabilistic Decision Schemes for One-Dimensional Single-Peaked Preferences," *Journal of Economic Theory* 105 (2), 408-434.

[24] Karsten Fieseler, Thomas Kittsteiner, and Benny Moldovanu (2003), "Partnerships, Lemons, and Efficient Trade," *Journal of Economic Theory* 113 (2), 223-234.

[25] Allan Gibbard (1973), "Manipulation of Voting Schemes: A General Result," *Econometrica* 41 (4), 587-601.

[26] Allan Gibbard (1977), "Manipulation of Schemes that Mix Voting with Chance," *Econometrica* 45 (3), 665-681.

[27] Allan Gibbard (1978), "Straightforwardness of Game Forms with Lotteries as Outcomes," *Econometrica* 46 (3), 595-614.

[28] Jerry R. Green and Jean-Jacques Laffont (1979), *Incentives in Public Decision-Making*, Amsterdam: North-Holland.

[29] Jerry R. Green and Jean-Jacques Laffont (1986), "Partially Verifiable Information and Mechanism Design," *Review of Economic Studies* 53 (3), 447-456.

[30] Birgit Heydenreich, Rudolf Müller, Marc Uetz, and Rakesh Vohra (2008), "Characterization of Revenue Equivalence," Meteor Research Memorandum RM/08/01, Maastricht University.

[31] Aanund Hylland (1980), "Strategy Proofness of Voting Procedures with Lotteries as Outcomes and Infinite Sets of Strategies," unpublished paper, Harvard University.

[32] Marcus Isaksson, Guy Kindler, and Elchanan Mossel (2010), "The Geometry of Manipulation: A Quantitative Proof of the Gibbard-Satterthwaite Theorem," in *2010 IEEE 51st Annual Symposium on the Foundations of Computer Science* (FOCS 2010), 319-328.

[33] Philippe Jehiel, Moritz Meyer-ter-Vehn, and Benny Moldovanu (2010), "Locally Robust Implementation and Its Limits," unpublished paper, Ecole Nationale des Ponts et Chaussées.

[34] Philippe Jehiel and Benny Moldovanu (2001), "Efficient Design with Interdependent Valuations," *Econometrica* 69 (5), 1237-1259.

[35] Philippe Jehiel, Moritz Meyer-ter-Vehn, Benny Moldovanu, and William R. Zame (2006), "The Limits of Ex Post Implementation," *Econometrica* 74 (3), 585-610.

[36] Bettina Klaus, Hans Peters, and Ton Storcken (1997), "Strategy-Proof Division of a Private Good when Preferences are Single-Dipped," *Economics Letters* 55 (3) 339-346.

[37] Eric Maskin and John Riley (1984), "Monopoly with Incomplete Information," *RAND Journal of Economics* 15 (2, 171-196.

[38] Allan H. Meltzer and Scott F. Richard (1981), "A Rational Theory of the Size of Government," *Journal of Political Economy* 89 (5), 914-927.

[39] Hervé Moulin (1980), "On Strategy-Proofness and Single Peakedness," *Public Choice* 35 (4), 437-455.

[40] Michael Mussa and Sherwin Rosen (1978), "Monopoly and Product Quality," *Journal of Economic Theory* 18 (2), 301-317.

[41] Roger B. Myerson (1981), "Optimal Auction Design," *Mathematics of Operations Research* 6 (1), 58-73.

[42] Roger B. Myerson and Mark A. Satterthwaite (1983), "Efficient Mechanisms for Bilateral Trading," *Journal of Economic Theory* 29 (2), 265-281.

[43] Kevin W. S. Roberts (1977), "Voting over Income Tax Schedules," *Journal of Public Economics* 8 (3), 329-340.

[44] Jean-Charles Rochet (1987), "A Necessary and Sufficient Condition for Rationalizability in a Quasi-Linear Context," *Journal of Mathematical Economics* 16 (2), 191-200.

[45] Alvin E. Roth (1982), "The Economics of Matching: Stability and Incentives," *Mathematics of Operations Research* 7 (4), 617-628.

[46] Alvin E. Roth (2008), "What Have We Learned from Market Design?," *Economic Journal* 118 (527), 285-310.

[47] Michael Saks and Lan Yu (2005), "Weak Monotonicity Suffices for Truthfulness on Convex Domains," in *Proceedings of the 6th ACM Conference on Electronic Commerce* (EC-05), 286-293.

[48] Alejandro Saporiti (2009), "Strategy-Proofness and Single-Crossing," *Theoretical Economics* 4 (2), 127-163.

[49] Shin Sato (2010), "A Sufficient Condition for the Equivalence of Strategy-Proofness and Nonmanipulability by Preferences Adjacent to the Sincere One," unpublished paper, Fukuoka University, October 2010 version.

[50] Mark A. Satterthwaite (1975), "Strategy-proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions," *Journal of Economic Theory* 10 (2), 187-217.

[51] Itai Sher and Rakesh Vohra (2010), "Optimal Selling Mechanisms on Incentive Graphs," unpublished paper, University of Minnesota.

[52] Yves Sprumont (1991), "The Division Problem with Single-Peaked Preferences: A Characterization of the Uniform Allocation Rule," *Econometrica* 59 (2), 509-519.

[53] Lin Zhou (1990), "On a Conjecture by Gale about One-Sided Matching Problems," *Journal of Economic Theory* 52 (1), 123-135.

# Chapter 2

# A Quantitative Approach to Incentives: Application to Voting Rules


## Abstract

We present a general approach to quantifying a mechanism's susceptibility to strategic manipulation, based on the premise that agents report their preferences truthfully if the potential gain from behaving strategically is small. Susceptibility is defined as the maximum amount of expected utility an agent can gain by manipulating. We apply this measure to anonymous voting rules, by making minimal restrictions on voters' utility functions and beliefs about other voters' behavior. We give two sets of results. First, we offer bounds on the susceptibility of several specific voting rules. This includes considering several voting systems that have been previously identified as resistant to manipulation; we find that they are actually more susceptible than simple plurality rule by our measure. Second, we give asymptotic lower bounds on susceptibility for any voting rule, under various combinations of efficiency, regularity, and informational conditions. These results illustrate the tradeoffs between susceptibility and other properties of the voting rule.



Thanks to (in random order) Ben Golub, Elchanan Mossel, Alex Wolitzky, Anton Kolotilin, Mihai Manea, Nathan Hendren, Yusuke Narita, Pablo Querubin, Lirong Xia, Abhijit Banerjee, Jing Chen, Rakesh Vohra, Pablo Azar, Jim Schummer, Iván Werning, Robert Akerlof, Glenn Ellison, Daron Acemoglu, Horacio Larreguy, Nabil Al-Najjar, Jim Snyder, Xiao Yu Wang, Jonathan Weinstein, Parag Pathak, and many seminar audiences for discussions and advice. This work was partially supported by an NSF Graduate Research Fellowship.

# 1 Introduction

## 1.1 Overview

It is standard in mechanism design, as elsewhere in economic theory, to assume that agents perfectly optimize. In particular, for direct revelation mechanisms, which ask agents to report their preferences, conventional theory requires perfect incentives — it should be exactly optimal for agents to report truthfully. In reality, however, decision-makers do not perfectly optimize, or at least do not optimize the material payoffs that are usually modeled. They may not know their environment well enough to be able to do so, and they may prefer to take computational shortcuts. Accordingly, this paper proceeds from an alternative behavioral premise: agents will report truthfully if the potential gains from doing otherwise — that is, from strategically manipulating the mechanism — are sufficiently small.

Under this premise, a mechanism designer may want to mildly relax the incentive constraints, rather than treat them as absolutely rigid, if doing so allows her to improve the performance of the mechanism in other respects. This suggests quantitatively measuring the incentives that a mechanism provides. Armed with such a quantitative measure, the designer can compare different mechanisms in terms of the incentives to manipulate, and consider tradeoffs between these incentives and other properties of the mechanism.

We propose in this paper to measure a mechanism's *susceptibility to manipulation* as the maximum amount of expected utility that an agent can gain by manipulating. That is, in very stylized terms, susceptibility is

$$\sigma = \sup_{u,lie,\phi} \Big( E_\phi[u(lie)] - E_\phi[u(truth)] \Big) \tag{1.1}$$

where the supremum is taken over all true preferences the agent may have (the utility function $u$, represented by the truthful report $truth$); all possible strategic misrepresentations $lie$; and all beliefs $\phi$ that the agent may hold about the behavior of other agents in the mechanism. Of course, the outcomes $u(lie), u(truth)$ depend on the

choice of mechanism, as well as on the behavior of other agents (encapsulated in the belief $\phi$).

The paper's mission is to advocate this approach to quantifying incentives. Issues of motivation and methodology will be taken up in some more detail in Subsection 1.3, but the bulk of the paper is dedicated to demonstrating how our measure can be used to obtain concrete results. For this, we apply the measure to voting rules: Given a population of voters, each with preferences over several candidates, what voting rule should they use to choose a winner as a function of their (reported) preferences?

The problem of choosing among voting rules provides a natural test case for any attempt to quantify manipulation. It is one of the oldest and most widely-studied problems in mechanism design, not to mention its wide range of applications. Moreover, the Gibbard-Satterthwaite theorem [22, 52] shows that no interesting voting rule is immune to strategic manipulation. Since incentives for strategic behavior are unavoidable, the need to quantify such incentives immediately presents itself in this setting.

To operationalize (1.1) for voting rules, we need two restrictions.

- First, we need to restrict the manipulator's utility function: otherwise the utility from a lie could be taken to be arbitrarily larger than the utility from the truth, and hence every (interesting) voting rule would have $\sigma = \infty$. We therefore impose the normalization that utility functions take values in $[0, 1]$.

- Second, we need to restrict the belief $\phi$: otherwise the manipulator could put probability 1 on some one profile of other voters' preferences for which he can manipulate, and hence we would always have $\sigma = 1$. We impose the restriction that, from the manipulator's point of view, the votes of the rest of the population should be independent and identically distributed across voters. In fact, as we elaborate further in Subsection 2.1, it is enough for us to require others' votes to be IID conditionally on some aggregate state; this restriction is still quite permissive. However, it does mean that we will restrict attention to *anonymous* voting rules (those that are invariant under permuting voters): it would not be

appropriate to assume each voter treats the others interchangeably unless the voting rule does so as well.

We will give the precise definition of susceptibility for voting rules in Subsection 2.1, after laying out basic vocabulary.

Our concrete results are of two sorts. First, in Section 3, we give quantitative bounds on the susceptibility of several rules discussed in prior voting literature. We begin by developing intuitions using simple voting systems, such as supermajority with status quo, plurality, and Borda count. We then reconsider several voting systems which previous literature identified as resistant to strategic manipulation: the Black, Copeland, Fishburn, minimax, and single transferable vote systems. It turns out that under our measure, all of these are *more* susceptible than simple plurality rule, unless the number of candidates is very small. Indeed, it is not trivial to supply an interesting example of a voting system that is less susceptible than plurality rule. We give such an example, in the case of three candidates.

Second, in Section 4, we give several theorems providing asymptotic lower bounds on the susceptibility of any voting rule satisfying various conditions, showing how fast the susceptibility of such rules can shrink as the number $N$ of voters grows. These lower bounds illustrate the tradeoffs between susceptibility and other properties of the voting rule. For example, if the voting rule is simply required to be weakly unanimous (a minimal efficiency condition), our lower bound is on the order of $N^{-3/2}$. If the voting rule is required to be monotone, we have a much stronger bound, on the order of $N^{-1/2}$. The latter bound goes to zero more slowly in $N$, and does not hold without the monotonicity restriction. Thus, imposing monotonicity substantially limits the voting rule's ability to resist manipulation, at least for a large number of voters. If we impose that the voting rule be monotone, unanimous, and also tops-only (i.e. the winner depends only on each voter's first choice), then we can solve exactly for the minimum possible susceptibility. This minimum is also on the order of $N^{-1/2}$, and is attained by majority rule with status quo, among others. The finding that majority rule is optimal again contrasts sharply with results on least-manipulable voting rules using a different measure of manipulability [34, 37]. We also give several more results

of this flavor (see Table 4.1 for a summary).

We should emphasize that this paper focuses on voting rules mainly because doing so constitutes a canonical theoretical exercise. Our conclusions are certainly not meant to be read literally as policy prescriptions — in practice, individual strategic manipulation is only one of many considerations that go into choosing a voting rule.

Our measure of susceptibility can be used to compare mechanisms and evaluate tradeoffs in many other mechanism design settings as well. As an example, the third chapter of this dissertation applies the same approach to study the tradeoff between incentives and efficiency in double auction mechanisms.

We believe that the generality of our method, its connection with a positive description of manipulative behavior, its tractability as illustrated by our results here for voting rules, and the contrast of several of our results with earlier findings using other measures of manipulability, taken together, provide a strong case for using this approach as one way to evaluate and compare mechanisms. In the concluding Section 5, aside from summarizing and indicating directions for future research, we also discuss how our approach fits into a broader program of mechanism design.

In order to avoid interrupting the flow of text with computations, most of the proofs are only sketched in the main text. The details of the omitted proofs are in Appendices C through H.

## 1.2 Related literature

The motivating viewpoint behind this paper is that quantifying strategic incentives is important for practical mechanism design. Accordingly, this paper is allied most closely with a literature arguing that the incentives to manipulate in particular mechanisms are small — beginning with the seminal paper of Roberts and Postlewaite on the Walrasian mechanism [50] and including recent work on matching markets [4, 25, 27, 28]. However, we build on the approach of this literature by showing how to quantify incentives explicitly, and by introducing them into the design problem, rather than focusing only on specific mechanisms.

Our evaluation of voting rules in terms of the incentives to manipulate is most

similar in spirit to a paper by Ehlers, Peters, and Storcken [18]. As in the present paper, their notion of susceptibility is defined as the maximum utility gain from manipulation. However, where we consider voting over a finite number of candidates, they consider voters who must collectively choose a point in Euclidean space, and they restrict attention to tops-only voting rules.

Recent independent work by Birrell and Pass [10] considers quantifying incentives in voting rules, using ideas very similar to ours, but they consider probabilistic voting rules and do not impose any restriction on beliefs. Day and Milgrom [16] and Erdil and Klemperer [19] used quantitative measures of strategic incentives to compare mechanisms for combinatorial auctions. Some other theoretical literature has also constructed mechanisms with small incentives to manipulate [5, 29, 32, 33, 39, 53], but without focusing as we do on comparisons between mechanisms or tradeoffs between incentives and other properties.

Finally, our work also naturally brings to mind the extensive prior literature that evaluates and compares voting systems using other measures of manipulation. By far the most common approach is profile-counting — that is, considering all possible profiles of voters' preferences that may occur, and measuring manipulability as the fraction of such profiles at which some voter can benefit by manipulating. This method appears to have been pioneered by Peleg [47] and has been followed by many authors since [2, 20, 26, 34, 35, 36, 37, 40, 45, 55, 56]. Variations include counting profiles in some weighted manner, e.g. weighted by the number of voters who can manipulate, or by the number of different false preferences by which a manipulator can benefit; or partially ordering mechanisms by the set of profiles at which someone can manipulate [21] (see also [46] for this approach applied to matching mechanisms). Some of the literature also considers manipulation by coalitions rather than individual voters [30, 31, 48, 49, 51]. The measure used by Campbell and Kelly [13], like ours, is based on the maximum gain from manipulating, but they define gain in terms of the number of positions in the manipulator's preference ordering by which the outcome improves. Yet another approach involves studying the computational complexity of the manipulation problem [7, 8].

## 1.3 Methodology

We now discuss in more detail the motivation behind our approach to measuring susceptibility. Readers interested in getting to the concrete results quickly can skip this subsection without loss of continuity.

Our measure is grounded in the following simple model of manipulation (again expressed in terms of voting systems just for specificity). Voters face a cost of $\epsilon > 0$ to behaving strategically, while truthful behavior is costless. The $\epsilon$ may be thought of as a computational cost (to computing a strategy, or acquiring information on other voters' preferences that is needed to strategize), or as a psychological cost of dishonesty. Then, if the gain from strategic manipulation is sure to be less than $\epsilon$, the voters will simply vote truthfully.

A planner needs to choose a voting rule for such voters. The planner cannot anticipate the voters' preferences, beliefs, or their exact strategic behavior, and she evaluates voting rules by their worst-case performance. The planner is, however, certain of one thing: if she chooses a voting rule with susceptibility $\sigma < \epsilon$, voters will vote truthfully. Truthful voting will then ensure that the result of the election really does reflect the voters' preferences in the way specified by the voting rule. This motivates the planner to choose a voting rule with low susceptibility, if possible.

This informal story summarizes verbal arguments in recent market design literature [4, 12, 27, 28], which use approximate strategyproofness of certain mechanisms to advocate their use in practice. We develop the model more formally in a game-theoretic framework in Appendix A.

In our model, the planner tries to prevent manipulation altogether. A common critique [9, 14, 61] argues that the planner's real goal should instead be to choose a mechanism that will ensure a good outcome in equilibrium, which may involve some manipulation along the way. However, that criticism, applied to the present paper, would miss the purpose. As discussed at the end of Appendix A (and further elaborated in the third chapter of this dissertation), a similar model could be used when the planner does have some specific theory of manipulative behavior. Our

general point that incentives can be measured quantitatively remains valid.

In view of the long previous literature mentioned in Subsection 1.2 using other approaches to measuring manipulation, we should also explain why we propose a new measure rather than taking an existing one off the shelf. Our approach has the following benefits:

- The measure of susceptibility (1.1) as the utility gain from misreporting is portable across many mechanism design problems.

- Our measure is tied directly to manipulative *behavior* via the simple model of the $\epsilon$ cost of behaving strategically. Consequently, it acknowledges the distinction between when manipulation is possible and when it will actually occur, in ways that a profile-counting measure would miss.

  For example, suppose that there are two candidates $A, B$, and suppose the number of voters is large. Each voter votes for his (reportedly) preferred candidate. Consider the voting rule that chooses $A$ if the number of $A$ votes is even and $B$ if it is odd. This rule is manipulable at almost every profile. But if a manipulator is fairly uncertain about the votes of the rest of the population, then it is not immediately obvious what the strategically optimal vote is; and the benefits from manipulation are low, because $A$ wins with probability close to $1/2$ no matter what the manipulator does. Hence, even a small cost of strategizing can discourage manipulation.

  For another example, consider the voting rule that chooses $A$ as winner if everyone votes for $B$, and $B$ otherwise. This voting rule is manipulable at only $N + 1$ out of the $2^N$ possible vote profiles. But voting truthfully is weakly dominated, and the incentives to vote strategically can be very strong — each voter is pivotal if his belief is that everyone else will vote for $B$ — so we should expect manipulation to be an important issue.

- Our comparison of plurality vote with other voting systems, and our identification of least-susceptible voting rules (Theorem 4.5 in particular), contrast with

previous results using profile-counting measures of manipulation. So even an analyst who prefers to use profile-counting measures should still take our $\sigma$ into consideration, as it gives novel insights.

# 2 Preliminaries

## 2.1 Framework and definitions

We now review standard concepts from voting theory, and subsequently introduce the terminology that will be needed to study our measure of susceptibility.

There is a set of $M$ *candidates*, $\mathcal{C} = \{A_1, \ldots, A_M\}$. We may refer to the candidates also as $A, B, C, \ldots$; we will use whichever notation is most convenient at the moment. There is also a set of $N + 1$ *voters*. (From here onwards we take the number of voters to be $N + 1$ rather than $N$, as this simplifies calculations.) We assume $M \geq 3$ and $N \geq 1$.[1] Some of our results are asymptotic; it will be understood that these asymptotics apply with $M$ held fixed and $N \to \infty$.

Each voter is assumed to have a strict preference (linear order) on the set of candidates. The symbol $\succ$ denotes a generic such preference. Let $\mathcal{L}$ denote the set of all $M!$ such preferences. A preference may be notated as a list of candidates; for example, if $M = 3$, $ACB$ denotes the preference that ranks $A$ first, $C$ second, and $B$ third. We may similarly write $AC\ldots$ to indicate that $A$ is first, $C$ is second, and the rest of the preference is unspecified. A *(preference) profile* is an element of $\mathcal{L}^{N+1}$, specifying each voter's preference. A *voting rule* is a map $f : \mathcal{L}^{N+1} \to \mathcal{C}$, choosing a winning candidate for each possible profile. (Note that some authors use terms such as *social choice function*, reserving *voting rule* for the special case where each voter reports only his top choice, e.g. [18, 34]).

We restrict attention throughout to voting rules that are *anonymous*, meaning that the outcome is unchanged if the voters are permuted. Consequently, we can notate the argument of $f$ as a list specifying the number of voters with each preference

[1] The case $M = 2$ is uninteresting in terms of incentives, e.g. using majority rule to decide between two alternatives gives no incentives to manipulate.

that occurs. For example, $f(3\ ABC, N-2\ BAC)$ refers to the candidate who wins when any 3 voters report preference $ABC$ and the other $N-2$ report $BAC$. This numbered list will also be called a *profile*. When there is potential ambiguity, we will use *nonanonymous profile* for a list specifying each voter's preference and *anonymous profile* if only the number of voters with each preference is to be specified. It will be useful to think of anonymous profiles as the integer points of a simplex in $M!$-dimensional space — those integer points whose coordinates are nonnegative and sum to $N+1$.

More generally, we define a *$K$-profile* (anonymous or nonanonymous) to be a list specifying the preferences of $K$ voters. When such partial profiles are concatenated, we mean that the votes are to be combined in the obvious way. For example, if $\succ$ represents one voter's preference and $P$ an $N$-profile describing preferences for the other $N$ voters, then $f(\succ, P)$ is the candidate chosen when the $N+1$ voters have the specified preferences.

We will also define here a few properties of voting rules which will be useful later. We organize these into three categories:

- *Efficiency properties:* A voting rule $f$ is *Pareto efficient* if, for any two candidates $A_i$, $A_j$ and any profile $P$ such that every voter ranks $A_i$ above $A_j$, $f(P) \neq A_j$.

  The voting rule is *weakly unanimous* if, for every preference $\succ$, $f(N+1\ \succ)$ is the candidate ranked first by $\succ$. That is, if all voters have identical preferences, their first choice wins. It is *strongly unanimous* if, for every profile $P$ such that all $N+1$ voters rank the same candidate $A_i$ first, $f(P) = A_i$. Clearly, Pareto efficiency implies strong unanimity, which in turn implies weak unanimity.

- *Regularity properties:* One regularity condition often viewed as normatively desirable [41] is monotonicity, which says that if the current winner's status improves, she remains the winner. The precise definition is as follows. First, given a preference $\succ$, a preference $\succ'$ is an *$A_i$-lifting* of $\succ$ if the following holds: for all $A_j, A_k \neq A_i$, we have $A_j \succ A_k$ if and only if $A_j \succ' A_k$, and $A_i \succ A_j$

implies $A_i \succ' A_j$. That is, the position of $A_i$ is improved while holding fixed the relative ranking of all other candidates. Then, a voting rule $f$ is *monotone* if it satisfies the following: For every profile $P$, if $P'$ is obtained from $P$ by replacing some voter's preference $\succ$ by an $f(P)$-lifting of $\succ$, then $f(P') = f(P)$.

We will also define here another very weak regularity condition (though not implied by monotonicity). Say that $f$ is *simple* on the pair of candidates $\{A_i, A_j\}$ if the following two conditions are satisfied:

- at any profile $P$ where every voter ranks $A_i, A_j$ first and second in some order, $f(P) \in \{A_i, A_j\}$;
- moreover, there is a value $K^*$ such that at every such profile, $f(P) = A_i$ if the number of voters ranking $A_i$ first is at least $K^*$, and $f(P) = A_j$ otherwise.

Note that the often-invoked property of *Condorcet-consistency* [41] — that, if a Condorcet winner exists (see Subsection 3.2), she should be elected — implies simplicity on every pair of candidates.

- *Informational properties:* We define just one property here. The voting rule $f$ is *tops-only* if the outcome depends only on each voter's first-choice candidate. In this case we can further economize on notation, writing, for example, $f(3\ A, N-2\ B)$.

  Tops-onliness is useful for intuition, because when $M = 3$, tops-only voting rules can be represented graphically. Indeed, since only first choices matter, the vote profiles now form a simplex in $M$-dimensional space rather than in $M!$-dimensional space. With $M = 3$, this simplex is just a triangular grid; the corners represent the all-$A$ profile, the all-$B$ profile, and the all-$C$ profile. We can illustrate a voting rule by coloring each cell of the grid according to the winning candidate. For example, Figure 2.1 illustrates a supermajority rule with $N + 1 = 7$ voters: either $B$ or $C$ is elected if she receives 5 or more votes; otherwise $A$ wins.

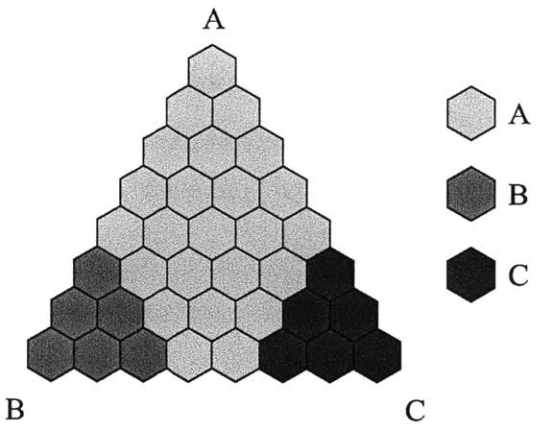


Figure 2.1: A tops-only voting rule

For non-tops-only rules, we can draw such grids, but only for small portions of the vote simplex.

Following [47], we will use the term *voting system* to denote a family of voting rules, one for each value of $N$. (In fact, our examples of voting systems will generally consist of a rule for each $M$ and $N$, but this detail is irrelevant since we think of $M$ as fixed and $N$ as varying.) A voting system is tops-only if the corresponding rule is tops-only for each $N$, and similarly for other properties.

We can now discuss manipulation. We consider one distinguished voter, the *manipulator*. The manipulator has a von Neumann-Morgenstern utility function $u : \mathcal{C} \to [0, 1]$.[2] We say that the utility function $u$ *represents* a preference $\succ$ if, for every two candidates $A_i, A_j$, $A_i \succ A_j$ implies $u(A_i) > u(A_j)$. We say that $u$ *weakly represents* $\succ$ if $A_i \succ A_j$ implies $u(A_i) \geq u(A_j)$.

We will use the term *opponent-profile* to refer to the $N$-profile representing the voters other than the manipulator. Suppose that the manipulator believes that the opponent-profile, $P$, follows the joint probability distribution $\Phi \in \Delta(\mathcal{L}^N)$. ($\Delta(X)$ means the simplex of probability distributions on $X$.) If $\succ$ is his true preference ranking, represented by $u$, then the amount of expected utility he can gain from

[2]Other voters may also have utility functions, but these are irrelevant from the manipulator's point of view because we assume they may only report ordinal preferences.

strategic manipulation is

$$\max_{\succ' \in \mathcal{L}} \Big( E_\Phi[u(f(\succ', P))] - E_\Phi[u(f(\succ, P))] \Big).$$

Here the operator $E_\Phi$ indicates expectation with respect to the distribution $\Phi$ for $P$.

We focus attention on a particular class of beliefs $\Phi$, those for which the other voters' preferences are IID. As argued by McLennan [38], this is a reasonable model of beliefs in a large population, where each member treats the others as interchangeable strangers.[3] For any $\phi \in \Delta(\mathcal{L})$, write $IID(\phi)$ for the distribution over opponent-profiles obtained by drawing each preference independently according to $\phi$.

We can now formally define our measure of susceptibility to manipulation. Let

$$\mathcal{Z} = \{(\succ, \succ', u, \phi) \in \mathcal{L} \times \mathcal{L} \times [0,1]^M \times \Delta(\mathcal{L}) \mid u \text{ represents } \succ\}.$$

The *susceptibility* of the voting rule $f$ is

$$\sigma = \sup_{(\succ, \succ', u, \phi) \in \mathcal{Z}} \Big( E_{IID(\phi)}[u(f(\succ', P))] - E_{IID(\phi)}[u(f(\succ, P))] \Big). \tag{2.1}$$

In words, $\sigma$ is the supremum of the amount the manipulator could gain in expected utility $u$ by reporting a preference other than his true preference $\succ$, given that his belief about $P$ is $IID(\phi)$ for some $\phi$.

The restriction to IID beliefs may seem confining. In fact we can relax it considerably, to conditionally IID beliefs. That is, suppose that instead of requiring the manipulator's belief to be IID, we allow that the manipulator has some uncertainty regarding the aggregate distribution of preferences $\phi$ in the population; but conditional on the realization of $\phi$, the opponent-profile $P$ is drawn $IID(\phi)$. Then, for any such belief, the manipulator still cannot gain more than $\sigma$ expected utility by manipulating. Indeed, suppose he manipulates by reporting $\succ'$ instead of the true preference $\succ$. Conditional on any value of the aggregate preference distribution $\phi$,

[3] It would be easy to extend the model, say, to allow each voter to have separate beliefs about a small number of other voters, representing his friends and family.

the expected gain from manipulating is at most $\sigma$ (by definition). So, by the law of iterated expectations, the *unconditional* expected utility gain from manipulating is again at most $\sigma$.

Thus, we could have defined susceptibility in (2.1) using conditionally-IID beliefs, rather than pure-IID beliefs; the two definitions would be equivalent. However, the pure-IID definition is easier to work with, so we stick to it, and refer to the conditionally-IID definition only for motivation.

We next introduce a useful alternative formulation of the definition of susceptibility. To work toward this alternative definition, we first use continuity to rewrite the supremum over $\mathcal{Z}$ in (2.1) as a maximum over the closure $\text{cl}(\mathcal{Z})$, and also take the difference inside the expectation:

$$\begin{aligned}\sigma &= \sup_{(\succ,\succ',u,\phi)\in\mathcal{Z}} \Big( E_{(IID(\phi)}[u(f(\succ',P))] - E_{IID(\phi)}[u(f(\succ,P))] \Big) \\ &= \max_{(\succ,\succ',u,\phi)\in\text{cl}(\mathcal{Z})} \Big( E_{IID(\phi)}[u(f(\succ',P)) - u(f(\succ,P))] \Big). \end{aligned} \tag{2.2}$$

Here the maximum is over the set

$$\text{cl}(\mathcal{Z}) = \{(\succ,\succ',u,\phi) \mid u \text{ weakly represents } \succ\}.$$

For given $\succ,\succ',\phi$, the maximand in (2.2) is a linear function of the values of $u$, so the maximum is attained at an extreme point of the simplex of utility functions $u$ weakly representing the given $\succ$. The extreme points are those that take the value 1 for the highest-ranked $L$ candidates, for some $L$, and 0 for the remaining candidates. Hence, we can also write

$$\sigma = \max_{(\succ,\succ',\mathcal{C}^+,\phi)} \Big( E_{IID(\phi)}[\mathbf{I}(f(\succ',P)\in\mathcal{C}^+) - \mathbf{I}(f(\succ,P)\in\mathcal{C}^+)] \Big), \tag{2.3}$$

where $\mathbf{I}(E)$ is the indicator function of event $E$, and the maximum is taken over all $\succ,\succ'\in\mathcal{L}$, $\phi\in\Delta(\mathcal{L})$, and $\mathcal{C}^+\subseteq\mathcal{C}$ such that $\mathcal{C}^+$ consists of the $L$ highest-ranked candidates under $\succ$ for some $L$. This is our alternative definition.

Expression (2.3) can be suggestively interpreted as the probability of being pivotal — that is, the probability (under the critical belief $\phi$) of drawing an opponent-profile $P$ for which the manipulation $\succ'$ changes the outcome from an undesirable one to a desirable one ($f(\succ, P) \notin \mathcal{C}^+$, $f(\succ', P) \in \mathcal{C}^+$). Indeed, many of our results, especially in Section 3, will be built on this interpretation. We stress however that the interpretation is not exactly correct, since for some opponent-profiles $P$ the manipulator is "antipivotal," changing the outcome from desirable to undesirable ($f(\succ, P) \in \mathcal{C}^+, f(\succ', P) \notin \mathcal{C}^+$). Thus, (2.3) can be more accurately described as the *net* probability of being pivotal.[4]

## 2.2 Analytical tools

When each voter's preference is drawn IID, the resulting profile follows a multinomial distribution. Consequently, it will be essential to have a compact notation for such distributions. We will write $\mathbf{M}(K; \alpha_1, \ldots, \alpha_r)$ to denote the multinomial distribution with $K$ trials and per-trial probabilities $\alpha_1, \ldots, \alpha_r$, with $\sum_i \alpha_i = 1$. We likewise write

$$\mathbf{P}(x_1, \ldots, x_r \mid K; \alpha_1, \ldots, \alpha_r) = \frac{K!}{x_1! \cdots x_r!} \alpha_1^{x_1} \cdots \alpha_r^{x_r}, \tag{2.4}$$

the probability that the values $(x_1, \ldots, x_r)$ are realized in such a distribution. (This applies when the $x_i$ are nonnegative integers with $\sum_i x_i = K$. For any other values of the $x_i$, we define $\mathbf{P}(x_1, \ldots, x_r \mid K; \alpha_1, \ldots, \alpha_r) = 0$.)

If $P$ is an (unordered) list of $K$ preferences and $\phi$ a distribution on $\mathcal{L}$, then we will write $\mathbf{P}(P \mid K; \phi)$ with the same meaning.[5] As before, we may notate $P$ by simply writing out each preference with its multiplicity. Similarly $\phi$ may be represented by writing each preference, preceded by its probability. More generally, we

[4]Expression (2.3) is also reminiscent of the notion of *influence* developed by Al-Najjar and Smorodinsky [3]. However, there are some important differences. Influence in [3] is defined with respect to a specific belief $\phi$, whereas we take the max over beliefs. The analysis in [3] imposes a noise assumption on $\phi$ — every voter must report every possible preference with probability bounded away from 0 — whereas we make no such assumption.

[5]We often use the letter $\alpha$ for a vector, or $\alpha_1, \ldots, \alpha_r$ for its components, to denote the parameters of the multinomial distribution thought of as abstract quantities, and $\phi$ for this same vector thought of as a probability distribution on $\mathcal{L}$ or $\mathcal{C}$.

can concatenate probability distributions, preceded by weights, to represent convex combinations: if $\phi, \phi' \in \Delta(\mathcal{L})$ and $\lambda \in [0,1]$, we may write $(\lambda\ \phi, 1-\lambda\ \phi')$ rather than $\lambda\phi + (1-\lambda)\phi'$. These concatenations will sometimes be written vertically rather than horizontally, as in

$$\mathbf{P}\left(\begin{array}{cc|cc} & K_1 & ABC & \\ & K_2 & ACB & N;\ \alpha_2\ ACB \\ & N-K_1-K_2 & BCA & \end{array}\right).$$

If $S$ is a set of profiles, we may write $\mathbf{P}(S \mid K; \phi)$ for $\sum_{P \in S} \mathbf{P}(P \mid K; \phi)$.

Many of our results will concern asymptotics as $N \to \infty$, so we should establish convenient notation accordingly. We are concerned not only with how quickly susceptibility declines to zero as $N \to \infty$, but also with the constant factors involved (when we are able to estimate them). This calls for somewhat nonstandard notation. We will write $F(N) \sim G(N)$ to indicate that $F(N)/G(N) \to 1$ as $N \to \infty$. If $F$ and $G$ depend on both $N$ and $M$, then it is understood that $M$ is held fixed. We will write $F(N) \lesssim G(N)$, or equivalently $G(N) \gtrsim F(N)$, to indicate $\limsup_{N\to\infty} F(N)/G(N) \leq 1$.

Now that we have finished introducing notation, we can lay out the main analytical tools that will be used in the rest of the paper. We present here a conceptual overview and a few of the most important technical results. The proofs of these results, as well as other useful technical computations, are given in Appendix C.

The single most important conceptual tool for our asymptotic analysis is the central limit theorem approximation of multinomial distributions: When $K$ is large, the distribution $\mathbf{M}(K \mid \alpha_1, \alpha_2, \ldots, \alpha_r)$ is approximately normal with mean equal to $(K\alpha_1, K\alpha_2, \ldots, K\alpha_r)$ and variance matrix

$$\begin{pmatrix} \alpha_1(1-\alpha_1)K & -\alpha_1\alpha_2 K & \cdots & -\alpha_1\alpha_r K \\ -\alpha_2\alpha_1 K & \alpha_2(1-\alpha_2)K & \cdots & -\alpha_2\alpha_r K \\ \vdots & \vdots & \ddots & \vdots \\ -\alpha_r\alpha_1 K & -\alpha_r\alpha_2 K & \cdots & \alpha_r(1-\alpha_r)K \end{pmatrix}.$$

This has numerous implications. For example, if $0 < \beta < 1$ and $x$ is an integer with $x \approx \beta N$, then $\mathbf{P}(x, N-x \mid N; \beta, 1-\beta) \approx \sqrt{1/(2\pi N\beta(1-\beta))}$. For a precise statement:

**Lemma 2.1** *Let $0 < \beta < 1$, and let $c$ be a constant. For each positive integer $N$, let $x_N$ be an integer with $|x_N - \beta N| < c$, and let $\beta_N \in [0,1]$ satisfy $|x_N - \beta_N N| < c$. Then*

$$\mathbf{P}\left(\begin{array}{c} x_N \\ N - x_N \end{array} \middle| \, N; \begin{array}{c} \beta_N \\ 1-\beta_N \end{array}\right) \sim \sqrt{\frac{1}{2\pi N\beta(1-\beta)}}.$$

Another set of implications that will be extremely useful for Section 3, where we bound the susceptibility of specific voting rules, is given by the following lemma. Its statement is notationally intense, but the content is intuitive, as we explain momentarily.

**Lemma 2.2** *Let $\mathcal{I}$ be a finite collection of strict linear inequalities in $r$ free variables $\beta_1, \ldots, \beta_r$, each of the form $c_0 + c_1\beta_1 + \cdots + c_r\beta_r > 0$. Let $J$ be a compact set of probability distributions $(\alpha_1, \ldots, \alpha_r)$, satisfying all the inequalities in $\mathcal{I}$. For each positive integer $N$, let $S_N^{\mathcal{I}}$ be the set of all $r$-tuples of nonnegative integers $(x_1, \ldots, x_r)$ summing to $N$, such that the numbers $x_1/N, \ldots, x_r/N$ satisfy the inequalities in $\mathcal{I}$.*

*(a) There is some $\lambda > 0$ such that*

$$1 - \min_{(\alpha_1,\ldots,\alpha_r)\in J} \mathbf{P}(S_N^{\mathcal{I}} \mid N; \alpha_1, \ldots, \alpha_r) \lesssim e^{-\lambda N}.$$

*(b) Fix $(\alpha_1, \ldots, \alpha_r) \in J$, and suppose further $\alpha_i = \alpha_j \in (0, 1/2)$ for some $i, j$; and let $y$ be any (integer) constant. Let $T_{ij,y} = \{(x_1, \ldots, x_r) \mid x_i - x_j = y\}$. Then*

$$\mathbf{P}(S_N^{\mathcal{I}} \cap T_{ij,y} \mid N; \alpha_1, \ldots, \alpha_r) \sim \frac{1}{2}\sqrt{\frac{1}{\pi\alpha_i N}}.$$

Part (a) is just a strengthened form of the law of large numbers. It states that when $(x_1, \ldots, x_r) \sim \mathbf{M}(N; \alpha_1, \ldots, \alpha_r)$, then each $x_i$ is close to $\alpha_i N$, with probability converging exponentially fast to 1 for large $N$. Part (b) estimates the further probability that $x_i - x_j$ takes on a particular constant value. The estimate follows from

the fact that $x_i - x_j$ is approximately normal with mean 0 and variance $2\alpha_i N$, and is approximately independent of all other components of $x$.

In many of the examples we will consider in Section 3, the manipulator is pivotal when the number of other voters reporting some preference order $\succ_i$ is exactly equal to the number of voters reporting another order $\succ_j$. In these cases, Lemma 2.2(b) is useful for estimating the probability of being pivotal.

We note for future reference that the pivotal probability in Lemma 2.2(b) declines in $N$ at rate $1/\sqrt{N}$, but that the constant factor depends on $\alpha_i$. In particular, the smaller $\alpha_i$ is, the higher the probability is. This is because the population shares of $\succ_i$ and $\succ_j$ have smaller variance, so are more likely to differ by exactly the required constant $y$.

We draw attention to one peculiarity: Consider $r = 2$, $\alpha_1 = \alpha_2 = 1/2$. This is a limiting case of Lemma 2.2(b), and so one might expect that the corresponding probability would be $\sim (1/2)\sqrt{1/\pi \cdot (1/2) \cdot N} = \sqrt{1/2\pi N}$. However, the probability is actually 0 if $N$ is the opposite parity from $y$, and $\sim \sqrt{2/\pi N}$ if $N$ is the same parity as $y$ (this follows from Lemma 2.1). The discontinuity occurs because the equality $x_1 + x_2 = N$ constrains the difference $x_1 - x_2$ to be the same parity as $N$, whereas in Lemma 2.2(b), as long as $\alpha_i = \alpha_j < 1/2$, the parity of $x_i - x_j$ is unrestricted.

Finally, in view of our worst-case approach to susceptibility — and particularly interpretation (2.3), the worst-case probability of being pivotal — it is natural to be interested in identifying the critical probability distributions for which some vote profile is most likely.

**Lemma 2.3** *For given nonnegative integers $x_1, \ldots, x_r$ with sum $K$, the maximum value of $\mathbf{P}(x_1, \ldots, x_r \mid K; \alpha_1, \ldots, \alpha_r)$ with respect to the $\alpha_i$ is attained at $\alpha_i = x_i/K$.*

**Lemma 2.4** *The expression*

$$\max_{\alpha \in [0,1]} \mathbf{P}\left( \begin{array}{c|c} K & \alpha \\ N-K & 1-\alpha \end{array} \right)$$

*is strictly decreasing in $K$ for $K \leq N/2$ and strictly increasing for $K \geq N/2$. In*

*particular, it is minimized over $K$ at $K = N/2$ if $N$ is even, and $(N \pm 1)/2$ if $N$ is odd.*

# 3 Susceptibility of specific voting systems

Now that we have developed the basic tools, we can begin applying our measure of susceptibility to manipulation to various voting systems.

We first develop intuitions in Subsection 3.1 by studying the susceptibility of four simple voting systems: (super)majority with status quo, plurality, $Q$-approval voting, and Borda count. Then, in Subsection 3.2, we consider several voting systems that have been identified in previous literature as resistant to manipulation, and find that by our measure, they are all *more* susceptible than simple plurality rule. In the process, we uncover several qualitative properties that make a voting rule relatively susceptible. Finally, the result of Subsection 3.2 raises the question of whether there are well-behaved voting systems that are less susceptible than plurality rule; in Subsection 3.3, we give an example of such a voting system for the case of three candidates.

For each of the voting systems studied in this section, the winner can be identified by checking a fixed set of inequalities (independent of $N$) in the population shares of the various possible preference orderings. In thinking about such systems, the most useful interpretation of susceptibility is (2.3), the probability of being pivotal.

## 3.1 Four simple voting systems

**Supermajority with status quo.** We begin by studying a rule for which we can compute the susceptibility exactly. Let $K$ be an integer with $(N+1)/2 < K \leq N+1$, and choose any fixed candidate, without loss of generality say $A$. The *supermajority rule with status quo* associated to $K$ and $A$ is the tops-only voting system defined as follows: if any candidate other than $A$ receives at least $K$ first-place votes, this candidate is chosen; otherwise $A$ wins. (Recall Figure 2.1.) If $K = \lfloor (N+3)/2 \rfloor$ then we have the *majority rule with status quo*. If $K = N+1$ then we have *unanimity*

*rule.*

**Proposition 3.1** *The supermajority rule with status quo has susceptibility*

$$\sigma_N^{smaj(K)} = \mathbf{P}\left( \begin{array}{c|cc} K-1 & & (K-1)/N \\ N-(K-1) & N; & 1-(K-1)/N \end{array} \right).$$

The basic approach to calculating susceptibility is to identify the profiles where opportunities for manipulation occur, and then identify a particular belief for which such opportunities are especially likely. For supermajority rule, we can actually identify the critical distribution that exactly maximizes the probability of being pivotal. Manipulation is possible only when the manipulator is pivotal between candidate $A$ and some other candidate (say $C$), and his true first choice (say $B$) cannot get elected. The manipulator is pivotal when $C$ has $K-1$ votes among the other voters. This is most likely to occur when each other voter chooses $C$ with probability $(K-1)/N$.

We give the full proof here.

**Proof:** Consider the formulation of susceptibility (2.3), as the probability that the manipulation changes the outcome from an undesirable one to a desirable one. If the manipulator's first choice is $A$, then manipulation cannot have such benefits: for any opponent-profile $P$, either the manipulator can ensure $A$ wins by voting for $A$, or else some other candidate has at least $K$ votes and the manipulator cannot change the outcome. If his first choice is some other candidate, say $B$, then manipulating to $A$ cannot affect whether or not any candidate different from $A$ and $B$ wins, and therefore cannot change the outcome except by adversely switching it from $B$ to $A$.

So the only possible beneficial manipulation is when the true first-choice is some non-$A$ candidate, and the manipulator votes for some other non-$A$ candidate. Without loss of generality, these are $B$ and $C$. The manipulation can be advantageous only if the opponent-profile $P$ is such that the manipulation changes the winner from $A$ to $C$. This in turn happens only if $C$ has exactly $K-1$ first-place votes in $P$. Let $S_C$ be the set of such profiles. Thus, the maximand in (2.3) is bounded above by $\Pr_{IID(\phi)}(P \in S_C) = \mathbf{P}(S_C \mid N; \phi)$. If $P$ is distributed according to $IID(\phi)$, and $\phi_C$ is

the probability (under $\phi$) of ranking $C$ first, then the total probability that $P \in S_C$ is $\mathbf{P}(K-1, N-K+1 \mid N; \phi_C, 1-\phi_C)$. So, combining these observations, we have

$$\sigma \leq \max_{\phi} \Pr_{IID(\phi)}(P \in S_C) = \max_{\phi_C} \mathbf{P}(K-1, N-K+1 \mid N; \phi_C, 1-\phi_C). \tag{3.1}$$

On the other hand, suppose the manipulator's true preferences are $BCA\ldots$ and the opponents' votes are distributed $(\phi_C\ C, (1-\phi_C)\ A)$, with $\phi_C = (K-1)/N$. A manipulation from $B$ to $C$ changes the outcome from $A$ to $C$ if $P \in S_C$, which happens with probability $\mathbf{P}(K-1, N-K+1 \mid N; \phi_C, 1-\phi_C)$, and leaves the outcome unchanged otherwise. By taking $\mathcal{C}^+ = \{B, C\}$ in definition (2.3), then, we get the reverse inequality of (3.1). Thus the inequality must hold as an equality.

From Lemma 2.3, the maximum in (3.1) is attained when $\phi_C = (K-1)/N$, giving the result of the proposition. □

From Lemma 2.4, the susceptibility $\sigma_N^{smaj(K)}$ is increasing in $K$. In particular, it is maximized for unanimity rule. This contrasts with results for (nonanonymous) profile-counting measures, where the number of manipulable profiles is lower for higher $K$ (compare in particular with [34, 36, 37], who identify the least-manipulable voting rules by such measures; they look qualitatively like unanimity rules). Likewise, the value of $K$ that minimizes $\sigma_N^{smaj(K)}$ is $K = (N+1)/2$ (for $N$ odd) or $N/2$ (for $N$ even). The corresponding value will actually come up again several times, so we establish a separate notation for it: The susceptibility of majority rule with status quo is given by

$$\sigma_N^* = \begin{cases} \binom{N}{N/2} \cdot \left(\frac{1}{2}\right)^N & \text{if } N \text{ is even} \\ \binom{N}{(N-1)/2} \cdot \left(\frac{(N-1)/2}{N}\right)^{(N-1)/2} \left(\frac{(N+1)/2}{N}\right)^{(N+1)/2} & \text{if } N \text{ is odd} \end{cases}$$

By Lemma 2.1, $\sigma_N^* \sim \sqrt{2/\pi N}$. This quantity will in fact appear again in the analysis of *plurality rule*, which we turn to next.

**Plurality rule.** The definition is as follows: For each candidate, we consider the number of first-place votes, and whoever has the most votes wins. For concreteness, ties are broken "alphabetically" — that is, in favor of earlier-numbered candidates; or earlier-lettered, when we use the notation $A, B, C, \ldots$ for candidates. (Most of our

results are not actually sensitive to how ties are broken).

**Proposition 3.2** *Let $\sigma_N^{plur}$ denote the susceptibility of plurality rule.*

*(a) For each $N$, $\sigma_N^{plur} \geq \sigma_N^*$.*

*(b) $\sigma_N^{plur}$ satisfies*

$$\frac{1}{2}\sqrt{\frac{1}{\pi}\cdot\frac{M}{N}} \lesssim \sigma_N^{plur} \lesssim \sqrt{\frac{1}{\pi}\cdot\frac{M}{N}}.$$

The lower bounds come from considering some potential critical distributions. One case where the manipulator has a relatively high probability of being pivotal is essentially when the manipulator's preferences are $ABC\ldots$ and the other voters split their first-place votes evenly between $B$ and $C$. Note that either $B$ or $C$ is sure to win, and the manipulator may want to vote for $B$ instead of $A$ in order to increase the chance of $B$ winning. This underlies part (a).

Another, related case is when the other voters split their votes almost evenly among all $M$ candidates, but with slightly higher (and equal) probabilities of voting for $B$ and $C$ than any of the others. In this case, again the outcome will almost certainly be either $B$ or $C$ (by Lemma 2.2(a)), incentivizing a vote for $B$ instead of $A$. Since the vote probabilities of $B$ and $C$ are equal and are approximately $1/M$ each, we can estimate the probability of being pivotal using Lemma 2.2(b); this probability is approximately $(1/2)\sqrt{M/\pi N}$. The lower bound in (b) follows.

It is not immediate, however, that the lower bound is sharp: By manipulating to $B$, the manipulator not only has a chance of changing the outcome from $C$ to $B$ but also a chance of changing from other undesirable outcomes $D, E, \ldots$ to $B$. Any upper bound on susceptibility must take account of all these possibilities.

The argument behind our upper bound runs as follows. Suppose the manipulator's true first choice is $A$ but he considers voting for $B$ as above. Consider the critical belief $\phi \in \Delta(C)$ that maximizes his probability of being pivotal. There must be *at least one* other candidate, say $C$, for which $\phi_C$ is close to $\phi_B$; otherwise the manipulator is unlikely to be pivotal. Now, beginning from any arbitrary opponent-profile, move along the $B - C$ axis — that is, hold constant the number of votes for all candidates

except $B$ and $C$, and vary the breakdown of the remaining votes into $B$ and $C$. We show that only one pivotal opponent-profile can be reached in this way. Consider the *conditional* probability of drawing this pivotal profile, given the number of votes for all candidates other than $B$ and $C$. Either the pivotal profile either has $B$ getting far more votes than $C$, in which case it is very unlikely; or it has both of them getting at least $1/M$ of the votes, in which case its probability is at most $\lesssim \sqrt{M/\pi N}$. So in either case, the conditional probability of the pivotal profile is $\lesssim \sqrt{M/\pi N}$. It follows that the unconditional probability of being pivotal is also $\lesssim \sqrt{M/\pi N}$, giving the upper bound.

The full proof of the proposition is in Appendix D.

Proposition 3.2 gives two different lower bounds on $\sigma_N^{plur}$, using two different beliefs. For small $M$, the bound in (a) is stronger than that in (b). Pivotality depends on the balance between larger population shares (1/2 for the belief used in (a), versus $1/M$ in (b)), which would tend to make the manipulator less likely to be pivotal under the belief used for (a), by the logic of Lemma 2.2(b) (the difference between these two shares has higher variance). On the other hand, in the case of (a), parity considerations add an extra factor of 2 to the probability of being pivotal, exactly as in the discussion following Lemma 2.2 above.

For the case of three candidates, we are able to extend this idea to show that the bound from (a) is exact — that is, the critical belief for a manipulator with preferences $ABC$ is that the opponents are split evenly between $B$ and $C$. However much or little probability of $A$ is introduced into the belief, the decrease in variance of the $B - C$ split is outweighed by the uncertainty over parity.

**Proposition 3.3** *If* $M = 3$, $\sigma_N^{plur} = \sigma_N^*$.

The proof is in Appendix D.

**$Q$-approval voting.** Next, we consider the voting system known as *$Q$-approval voting*, for any given $Q$ with $2 \leq Q \leq M - 1$. Each voter gives a point to each of his $Q$ favorite candidates. The candidate with the most points wins; ties are broken alphabetically. In the case $Q = M - 1$, this system is often known as *antiplurality*

voting.

Despite the superficial resemblance to plurality voting, this system is much easier to analyze, and also gives quite different results.

**Proposition 3.4** *For each $Q$, the susceptibility of $Q$-approval voting is* 1.

**Proof:** Let the manipulator's true preference be $BA\ldots$ and let $\phi$ be the distribution putting probability 1 on a preference of the form $AB\ldots$. So the manipulator's belief is that everyone else will report this preference, with probability 1. If the manipulator tells the truth, then $A$ receives $N+1$ points, the maximum possible, and hence (by alphabetical tie-breaking) $A$ wins, regardless of the other candidates' scores. If the manipulator instead reports any preference with $B$ ranked first and $A$ ranked last, then $A$ receives only $N$ points and $B$ receives $N+1$, so (again by alphabetical tie-breaking) $B$ must win. Thus, this manipulation improves the outcome from $A$ to $B$ with probability 1.

This example shows that the susceptibility of $Q$-approval voting is at least 1. Since susceptibility can never be more than 1, the result follows. □

The result is perhaps surprising, since standard approval voting (in which each voter approves any set of candidates, and whoever receives the most approvals wins) has often been specifically advocated as resistant to manipulation [11, 21]. We do not analyze this version of approval voting here, because it does not fit directly into our framework — in particular, it is unclear how a voter's default truthful vote should be defined. Appendix B discusses possible ways of extending our methods to treat approval voting.

**Borda count.** Another often-discussed voting system is the *Borda count*, which determines a winner as follows. Each voter assigns $M(M+1)/2$ points to the candidates: $M$ points to his first choice, $M-1$ to his second choice, ..., 1 point to his last choice. For each candidate, we compute a score by totaling across voters. The candidate with the highest score wins. Ties are again broken alphabetically.

We content ourselves to give a lower bound on susceptibility.

**Proposition 3.5** *The Borda count has susceptibility*

$$\sigma_N^{Borda} \gtrsim \left\lceil \frac{M-2}{2} \right\rceil \sqrt{\frac{2}{\pi N}}.$$

The argument is analogous to that of Proposition 3.2(a). Consider a manipulator with preferences $ABC\ldots$. Let the belief be as follows: opponents are evenly split between $ABC\ldots$ and $BAC\ldots$. Then the winner is surely either $A$ or $B$. By moving $B$ to the bottom of his reported preference ordering, instead of being truthful, the manipulator can improve the score of $A$ relative to $B$ by $M-2$ points. Hence, the manipulator is pivotal if, among the other voters, $A$ trails $B$ by more than 1 point but not more than $M-1$. Our lower bound follows by estimating the probability of this event.

The full detailed proof is in Appendix D.

To segue into the next section, we compare the results of Propositions 3.1, 3.2(b), and 3.5. Supermajority with status quo, plurality, and Borda count all have susceptibility declining as $N \to \infty$ at rate $1/\sqrt{N}$; but the constant factors (relative to $N$) are different. In particular, the constant factor for supermajority is constant in $M$; that for plurality is on the order of $\sqrt{M}$; and that for Borda count is linear in $M$. This allows unambiguous comparisons between these rules for sufficiently large $M$. For example, the comparison between Propositions 3.2(b) and 3.5 shows that, when $M \geq 5$, Borda count is more susceptible than plurality rule if the number of voters $N$ is large.

## 3.2 Low manipulability revisited

Next, we consider voting systems which have been specifically identified as resistant to manipulation in previous literature, using different measures, and ask whether they continue to fare well under our measure of susceptibility. To decide which voting systems to examine, we turn for guidance to the work of Aleskerov and Kurbanov [2], which appears to be the most extensive prior comparison of voting rules in terms of strategic manipulation. Aleskerov and Kurbanov used Monte Carlo simulations, with

small numbers of voters and candidates, to compare 25 voting systems according to several profile-counting-based measures of manipulability. We will consider the systems highlighted by their analysis, and give lower bounds on the susceptibility of each of these systems. As a benchmark for comparison, we use plurality rule, which is surely the most widespread voting rule in practice. Our lower bounds will imply that each of the systems picked out by [2] is actually more susceptible than plurality rule, under our measure. Table 3.1 gives a quick summary of our findings, and the details are explained below.

Like most of our results, the comparisons will be asymptotic (in the number of voters). For given $M$, we say that a voting system $f$ is *more susceptible* than $g$ if there is a positive constant $c$ such that the susceptibility of $f$ is at least $1 + c$ times the susceptibility of $g$, for all sufficiently large $N$. Thus, for example, we say that Borda count is more susceptible than plurality rule (for $M \geq 5$), even though both have susceptibility decaying at rate $1/\sqrt{N}$.

The comparison paper by Aleskerov and Kurbanov [2] does not conclusively favor some particular voting system. Instead, we consider all the systems that are identified by name in their concluding section. In addition to the Borda and $Q$-approval voting systems, which we have already considered, these include the Black, Copeland, Fishburn, minimax, and single transferable vote systems.

We will define these voting systems momentarily, but we first need a couple preliminary definitions. Given an $(N+1)$-profile $P$, we say that candidate $A_i$ *majority-defeats* candidate $A_j$ — notated $A_i \to A_j$ — if

- more than $(N+1)/2$ of the voters rank $A_i$ above $A_j$, or
- exactly $(N+1)/2$ of the voters rank $A_i$ above $A_j$, and $i < j$.

(The second case is used to ensure that among any two candidates, one majority-defeats the other. Again, our results are not sensitive to how such ties are broken.) A *Condorcet winner* is a candidate that majority-defeats every other candidate; if a Condorcet winner exists, she is unique.

The voting systems we consider are defined as follows:

- *Black's* system: If a Condorcet winner exists, that candidate is chosen; otherwise, Borda count is applied.

- *Copeland's* system: Define the *score* of each candidate $A_i$ to be the number of candidates $A_j$ such that $A_i \rightarrow A_j$. Choose the candidate with the highest score as the winner; break ties alphabetically.

- *Fishburn's* system (also known as the *uncovered set* system [56]): Say that a candidate $A_i$ *covers* another candidate $A_j$ if, for all $k$ such that $A_k \rightarrow A_i$, we also have $A_k \rightarrow A_j$. (In particular, this requires $A_i \rightarrow A_j$.) This is a partial ordering on the set of candidates, so there must exist at least one uncovered candidate. This candidate is the winner. If there is more than one uncovered candidate, we choose the alphabetically earliest.

- *Minimax* system (also known as *Simpson's* system): For each candidate $A_i$, let the *score* be the maximum, over all $j \neq i$, of the number of voters ranking $A_j$ above $A_i$. Choose the candidate with the lowest score as the winner, breaking ties alphabetically as usual.

- *Single transferable vote* system (also known as *successive elimination* or *Hare's* system): Each voter has one vote, initially assigned to his first-choice candidate. For each candidate, we determine the number of votes she receives. The candidate $A_{i_1}$ with the fewest votes is eliminated; ties are broken alphabetically (that is, in favor of keeping alphabetically earlier candidates). Each voter who ranked $A_{i_1}$ first has his vote reassigned to his second-choice candidate. Then, among the remaining candidates and new votes, we again eliminate the candidate $A_{i_2}$ with the fewest votes, reassign these votes, and so forth. The last candidate to escape elimination is the winner.

These voting systems are listed in the first column of Table 3.1. In the second column, we give an asymptotic lower bound on the susceptibility of each system. In each case, we prove the lower bound for all $M$ except possibly some small values. The table indicates exactly for which $M$ we prove the bound. (For the minimax system,

| System | Susceptibility Bound | $\sigma_N \gtrsim (1+c)\cdot\sigma_N^{plur}$? |
|---|---|---|
| Black | $\sigma_N^{Black} \gtrsim \lceil\frac{M-2}{2}\rceil\sqrt{\frac{2}{\pi N}}$ for $M \geq 5$ | $M \geq 5$ |
| Copeland | $\sigma_N^{Copeland} \gtrsim \lfloor\frac{M+1}{3}\rfloor\sqrt{\frac{2}{\pi N}}$ for $M \neq 5$ | $M \geq 6$ |
| Fishburn | $\sigma_N^{Fishburn} \gtrsim (M-3)\sqrt{\frac{2}{\pi N}}$ | $M \geq 5$ |
| Minimax | $\sigma_N^{minimax} \gtrsim \frac{c}{\sqrt[4]{N}}$ for $M \geq 4$ | $M \geq 4$ |
| STV | $\sigma_N^{STV} \gtrsim \sqrt{\frac{2^{M-1}}{\pi N}}$ | all $M$ |

Table 3.1: Comparison of voting systems identified in [2] against plurality rule. The second column gives lower bounds on susceptibility. Each system is more susceptible than plurality, for the values of $M$ indicated in the third column.

the statement is that there is some absolute constant $c$ such that $\sigma \gtrsim c/\sqrt[4]{N}$ for all $N$ and $M$.)

For most of the voting systems, our lower bound is decreasing in $N$ at rate $1/\sqrt{N}$, but with different constant factors. Each such constant factor grows at least linearly in $M$ — faster than the $\sqrt{M}$ factor for plurality rule (from Proposition 3.2(b)). Therefore, each voting system is more susceptible than plurality rule when $M$ is large enough. Specifically, by comparing the second column of the table with Proposition 3.2(b), we get the results shown in the third column: each voting system listed is more susceptible than plurality rule for the indicated values of $M$.

In particular, our lower bound for single transferable vote is exponential in $M$, so that it is substantially more susceptible than plurality rule for moderately large numbers of candidates; and our lower bound for minimax is on the order of $N^{-1/4}$ rather than $N^{-1/2}$, so it is *much* more susceptible than plurality rule, in large populations, as long as $M \geq 4$.

**Proposition 3.6** *The five voting systems listed in Table 3.1 satisfy the asymptotic lower bounds on susceptibility listed in the table. (In particular, all of them are more susceptible than plurality rule when $M \geq 6$.)*

The proof of Proposition 3.6 is in Appendix E. Here we give a sketch of the arguments used. In the process, we highlight the insights gained about the properties of these voting systems that make them particularly susceptible.

Broadly, the approach is the same as for the lower bounds in Propositions 3.2(b) and 3.5. For each system, we prove the lower bound by constructing a particular belief $\phi$ and proposed manipulation, and estimating the probability of being pivotal.

For minimax and single transferable vote, the crucial intuition is that a rule is highly susceptible if it is sensitive to the balance between two very small shares of the population.

In more detail: We construct the belief $\phi$ in such a way that pivotality occurs when the numbers of opponents reporting preferences $\succ$ and $\succ'$ are equal, for some particular $\succ, \succ' \in \mathcal{L}$. In this belief, $\succ$ and $\succ'$ occur with equal probability $\alpha$. Then, from Lemma 2.2(b), the probability of being pivotal is $\sim (1/2)\sqrt{1/\pi\alpha N}$. In particular, for small $\alpha$, the probability of being pivotal is high. For these two voting systems, we can construct beliefs with the relevant $\alpha$ quite small. In particular, in the case of minimax, we achieve the $N^{-1/4}$ convergence rate by varying the belief as $N$ increases, so that the population shares of the two relevant preference orders go to zero. Plurality rule, on the other hand, does not suffer from this sensitivity to small population shares, since the opportunity to be pivotal between some two potential winners only arises when each of them is the first choice of at least $1/M$ of the voters.

The Copeland and Fishburn systems are defined in terms of the majority defeat relation, which cannot hinge on small population shares, so we cannot use a similar construction to show that these systems have high susceptibility. Instead, the intuition we use here is that a rule is highly susceptible if the manipulator can simultaneously be pivotal in many independent ways.

Specifically, for each of these systems, we construct a belief with the following property: there are many pairs $\{A_i, A_j\}$ over which the population is close to evenly split, and if the manipulator is pivotal for any one of these pairs, he can manipulate advantageously. For each such pair, the probability of being pivotal is $\sim \sqrt{2/\pi N}$. The number of pairs is linear in $M$, and pivotality for any pair is independent of pivotality for any other pair, so that the overall probability of being pivotal is $\sim \sqrt{2/\pi N}$ times a coefficient linear in $M$.

One might at first think that plurality rule allows the same construction, since,

as pointed out in the discussion preceding Proposition 3.2, it is possible to be pivotal in many ways simultaneously: a manipulation from $A_i$ to $A_j$ can change the outcome from $A_k$ to $A_j$, for each $k \neq j$. But these pivotality conditions are not independent of each other, since the manipulator can only be pivotal from $A_k$ to $A_j$ when $A_k, A_j$ are the two candidates with the most first-place votes.

Finally, for Black's system, we exploit the same intuition as for the Borda count: a manipulation can have a large effect on the relative standing of two candidates, so that the slice of the vote simplex for which the manipulator is pivotal has "thickness" proportional to $M$. Indeed, the construction we give for Black's system is based on our construction for Borda count, with some extra foolery added to prevent the existence of a Condorcet winner.

Before closing this subsection, we should comment on the practical significance of a comparison like Proposition 3.6. Is it not enough to simply know that each voting system's susceptibility tends to zero for $N$ large?

In the context of our motivating model, with the $\epsilon$ cost of strategic behavior, a result comparing the susceptibility of two voting systems is most cogent if we believe that a plausible cost of behaving strategically would be on the same order of magnitude as the susceptibility of the two rules. In this case, there would be agents who would consider manipulating under one system but not the other.

Consider a six-candidate election, in an organization with 2,000 members (this could correspond to, say, a leadership election in a modest-sized professional organization). Treating the asymptotic bounds as exact, we have from Proposition 3.2 an upper bound of 0.031 for the susceptibility of plurality rule, whereas the lower bounds from Proposition 3.6 are 0.036 for Black and Copeland, 0.054 for Fishburn, and 0.071 for single transferable vote. These numbers are economically distinguishable from zero. More precisely, the differences in susceptibility between the voting systems are important if the voters' cost of behaving strategically is on the order of 3 to 7 percent of their concern about the outcome. This seems a reasonable estimate in many organizations, where most members' interest in the outcome of elections is modest.

## 3.3 A new voting system

We have now shown that a number of voting systems, previously identified as resistant to manipulation under profile-counting definitions, are in fact *more* susceptible to manipulation than the benchmark of plurality rule under our worst-case measure. A question which naturally presents itself is: is there any reasonable voting system that is *less* susceptible than plurality?

There are a couple of easy, but not entirely satisfactory, answers. In Section 4, we will indicate how to construct a unanimous voting system whose susceptibility is on the order of $1/N^{\kappa}$, for some $\kappa > 1/2$. Thus, such a rule is considerably less susceptible than any of the voting systems we have considered, for large $N$. However, that rule will be arguably artifical and violates almost any standard regularity condition.

Another possible answer is one we have already given, namely majority rule with status quo; our bounds imply that it is less susceptible than plurality rule if $M \geq 9$. However, this voting system treats the candidates in a very asymmetric manner.

We will give below a voting system that is less susceptible than plurality rule, for the special case $M = 3$. This voting system is well-behaved, in the sense of being unanimous and monotone, and arguably treats the candidates as fairly as possible. (Complete symmetry among candidates — often called *neutrality* in social choice theory — would be complicated by the need to break ties. Rather than formally define neutrality with exceptions for tie-breaking, we just argue intuitively that our rule breaks symmetry only in knife-edge cases.)

The construction is based on the following observation: Under plurality rule with $M = 3$, the strongest incentive to manipulate arises when voters split evenly between two candidates (see Proposition 3.3). In this case, however, deciding by majority rule between these two candidates (ignoring the third candidate), rather than using plurality, would eliminate the incentive to manipulate. This suggests constructing a voting rule such that

- when two candidates are "far ahead" of the third in terms of first-place votes, the winner is chosen by majority rule between the two leading candidates;

- when all three candidates are roughly evenly matched, plurality rule is used; and

- the transition between the two preceding cases is smooth enough to avoid creating other opportunities for manipulation.

We now construct a voting system $f$ along these lines, which we will call the *pair-or-plurality* voting system. For $N$ sufficiently large, let $K, L$ be positive integers with $2K < L < N/6$. (These values can depend on $N$, in ways to be specified later.)

Say that a candidate $A_i$ is *viable* if $A_i$ receives at least $K$ first-place votes. The winner is determined as follows:

(a) If there is only one viable candidate, she wins.

(b) If there are two viable candidates, the winner is determined by majority vote between them (with ties broken alphabetically).

(c) If all three candidates are viable, then we compute a *score* for each candidate. For each candidate $A_i$, consider the voters ranking her first. Let the number of voters reporting preferences $A_iA_jA_k, A_iA_kA_j$ be $x, y$ respectively. We will award $x + y$ corresponding points to the three candidates, as follows:

   - If $x + y \geq L$, then all $x + y$ points are awarded to $A_i$.
   - If $x + y < L$, then we award

$$\frac{L(x+y-K)}{L-K} \text{ points to } A_i,$$

$$\max\left\{0, \min\left\{x - \frac{(x+y-K)L}{2(L-K)}, \frac{K(L-x-y)}{L-K}\right\}\right\} \text{ points to } A_j,$$

$$\max\left\{0, \min\left\{y - \frac{(x+y-K)L}{2(L-K)}, \frac{K(L-x-y)}{L-K}\right\}\right\} \text{ points to } A_k.$$

After doing this for each candidate $A_i$, ultimately we have allocated $N + 1$ points, corresponding to the $N + 1$ voters. Then the candidate with the most points wins. Ties are broken alphabetically.

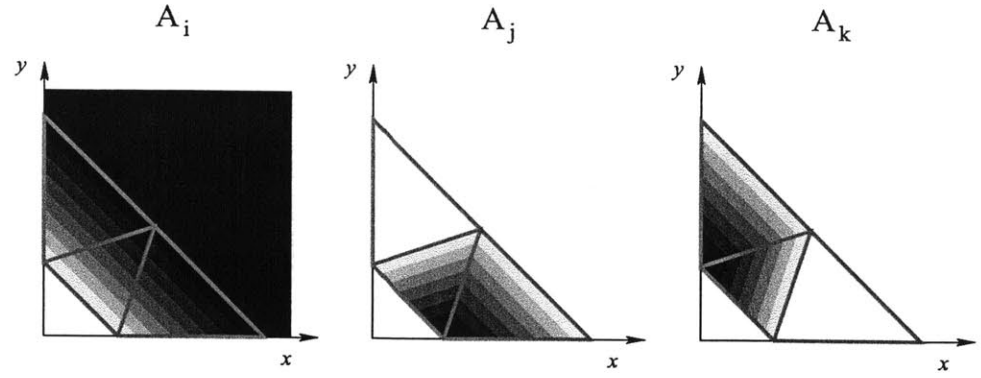


Figure 3.1: Scoring system in case (c) of the pair-or-plurality voting rule. The level plots show what fraction of the $x + y$ points are allocated to each candidate, as a function of $x$ and $y$. Darker regions represent more points for the candidate indicated. For reference, the gray lines connect the points $(x, y) = (K, 0), (0, K), (L, 0), (0, L)$, and $(L/2, L/2)$.

The scoring system in case (c) is illustrated in Figure 3.1, which shows the allocation of points as a function of $x$ and $y$. This system achieves a smooth transition between majority rule (in the case $x + y = K$, where the $x + y$ points are awarded to $A_j$ and $A_k$ based on pairwise preference) and plurality rule (when $x + y \geq L$, where all $x + y$ points go to $A_i$).

**Lemma 3.7** *For each $N$, the pair-or-plurality voting rule constructed above is monotone and Pareto efficient.*

We now give our main result for the pair-or-plurality voting rule. It applies when $K$ and $L$ are chosen to vary in the appropriate way as functions of $N$.

**Proposition 3.8** *If $K, L$ are chosen for each $N$ so that $L/K \to \infty$ and $K \to \infty$ as $N \to \infty$, then*

$$\sigma_N^{POP} \lesssim \frac{1}{2}\sqrt{\frac{3}{\pi N}}.$$

Comparing this upper bound to Proposition 3.2(a), we see that the pair-or-plurality rule is indeed less susceptible than plurality rule.

The proofs of both of the above results are in Appendix F.

Unfortunately, there is no obvious way to generalize the construction of the pair-or-plurality voting rule to a system that is less susceptible than plurality rule for arbitrary

$M$. For large $M$, the critical distribution for plurality no longer has opponents evenly split between two candidates, so our motivating idea does not apply. Finding a well-behaved voting system that is less susceptible than plurality rule for arbitrary $M$, or showing that no such voting system exists (under an appropriate definition of "well-behaved"), is a task for future research.

# 4 General lower bounds

The previous section gave comparisons of several specific voting systems. However, a mechanism designer may often approach her problem not with particular mechanisms in mind, but rather with a list of desired properties that a mechanism should satisfy, and then ask how well she can do in terms of strategic incentives while satisfying those other properties. In this section, we give several illustrative results to show how our measure of susceptibility can be used to address such questions. Each of our results is of the following form: for some combination of (efficiency, regularity, informational) properties, we provide an asymptotic lower bound on the susceptibility of any voting rule satisfying them. The properties we use are those defined in Subsection 2.1.

These lower bounds (together with some partial tightness results) offer insights into the quantitative tradeoffs between susceptibility to strategic manipulation and other desiderata. They can also be viewed, more pessimistically, as quantitative versions of the Gibbard-Satterthwaite theorem, analogous to the recent results of Isaksson, Kindler, and Mossel [26] and Mossel and Racz [40] which used a profile-counting measure. (A version of the Gibbard-Satterthwaite theorem for our IID setting was first proved by McLennan [38].)

For expositional smoothness, we begin by presenting all of the results, in Subsection 4.1. That subsection ends with a very brief sketch of the tools used in the proofs. Ensuing subsections give more careful outlines of the proofs. These outlines are of interest in themselves, as they illustrate more general techniques for working with our measure of susceptibility. The full proofs are for the most part left to Appendix G.

As before, our results are asymptotic in $N$, so we treat $M$ as fixed. Thus when

any result in this section refers to a "constant," it is understood that the constant may depend on $M$ but not $N$.

## 4.1 Statement of results

The discussion here will explain the motivation behind each result. A quick summary of the results is provided in Table 4.1 near the end of this subsection.

Since any constant voting rule obviously has susceptibility zero, some efficiency condition needs to be imposed to obtain any interesting results. A minimal such restriction is weak unanimity, which leads to the following general lower bound:

**Theorem 4.1** *There exists a constant $c > 0$ such that, for every value of $N$, every weakly unanimous voting rule $f$ has susceptibility $\sigma \geq cN^{-3/2}$.*

If we add tops-onliness, then we can improve the exponent from $-3/2$ to $-1$. (Note that a less negative exponent of $N$ means a higher value, thus a stronger lower bound.)

**Theorem 4.2** *There exists a constant $c > 0$ such that every unanimous, tops-only voting rule has susceptibility $\sigma \geq cN^{-1}$.*

(We simply say *unanimous* because weak and strong unanimity coincide for tops-only voting rules.)

It is unknown whether the bounds in Theorems 4.1 and 4.2 are tight. The voting systems considered in Section 3, which all had susceptibility of order $N^{-1/2}$ or larger, might suggest that a tight lower bound should have an exponent of $-1/2$. The following result shows that such a bound actually does not hold in general:

**Theorem 4.3** *There exist a number $\kappa > 1/2$ and a Pareto-efficient, tops-only voting system with susceptibility $\sigma \lesssim N^{-\kappa}$.*

The slower rate of decline in Section 3 exploited the interpretation of susceptibility as the probability of being pivotal. Theorem 4.3 instead depends on a construction for which the pivotal intuition does not apply.

Instead, we construct a low-susceptibility voting system based on the following ideas. Imagine temporarily that we allow voting rules to specify probabilistic outcomes. Thus instead of being a function $f : \mathcal{L}^{N+1} \to \mathcal{C}$, a voting rule is a function $f : \mathcal{L}^{N+1} \to \Delta(\mathcal{C})$. With expected utility over lotteries, our definition of susceptibility (2.1) remains applicable. But now the *random dictatorship* voting rule, which picks a voter uniformly at random and then chooses that voter's first choice as the winner, has susceptibility zero.

In this paper, we have forbidden explicitly random voting rules, so the random dictatorship is disallowed. However, there is still room for implicit randomization, via the manipulator's IID uncertainty about others' votes. This allows us to construct an $f$ that looks approximately like random dictatorship from the manipulator's point of view: For any $(N+1)$-profile $P$, we choose the values $f(Q)$ for profiles $Q$ close to $P$, so that the fraction of such profiles at which any candidate $A_i$ wins is close to the fraction of the population voting for $A_i$ at $P$. This is illustrated in Figure 4.1. The construction in Appendix H in effect achieves this for all $P$ simultaneously, to within an error of order strictly smaller than $N^{-1/2}$. (That construction requires some additional details not reflected in the figure.)

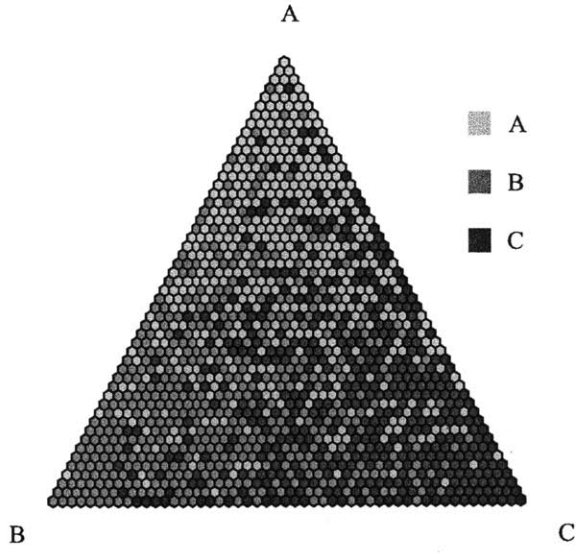


Figure 4.1: The approximate random dictatorship voting rule

This approximate random dictatorship is extremely sensitive to the exact vote

profile, so that the pivotal intuition does not apply. However, one might argue that it is not a realistic voting rule, and impose a regularity condition to rule out such a construction. For example, monotonicity does the trick, at least as long as we are also willing to strengthen unanimity to Pareto efficiency. This restores the $N^{-1/2}$ rate of decline in susceptibility that we saw in Section 3:

**Theorem 4.4** *There exists a constant $c$ such that every Pareto efficient and monotone voting rule $f$ has susceptibility $\sigma \geq cN^{-1/2}$.*

If we impose both monotonicity and tops-onliness, the problem becomes structured enough so that we can compute the minimum susceptibility exactly. Moreover, we can partially characterize the voting rules attaining the minimum. Say that a tops-only voting rule $f$ is a *majority rule* if it satisfies the following: for every profile $P$ at which more than half the voters rank the same candidate $A_i$ first, $f(P) = A_i$.

**Theorem 4.5** *Every unanimous, monotone, tops-only voting rule $f$ has susceptibility $\sigma \geq \sigma_N^*$. Moreover, if equality holds, and $N \geq 4$, then $f$ must be a majority rule.*

Equality is attained, for example, by majority rule with status quo (Proposition 3.1). Again, this contrasts with the results of [34, 37], using a profile-counting measure of manipulation; the least-manipulable voting rules they identify look qualitatively like unanimity rules, not majority rules.

Theorems 4.4 and 4.5 both give bounds on the order of $N^{-1/2}$. The example of Section 3.3 shows that Theorem 4.5 is not redundant: the bound there would not hold if we did not require tops-onliness.

Finally, we give two theorems showing that the relatively mild regularity condition of simplicity already makes some demands on incentives. By itself, it is enough to imply an $N^{-1}$ bound (where we had $N^{-3/2}$ otherwise); and combined with tops-onliness, it gives $N^{-1/2}$, the same order of magnitude as monotonicity.[6]

[6]The latter result, Theorem 4.7, does not even require an explicit efficiency condition: simplicity imposes enough efficiency to yield the bound. Note that even though simplicity only concerns two candidates, the usual method of giving perfect incentives by using majority vote between these two candidates is unavailable, because it violates tops-onliness.

**Theorem 4.6** *There is a constant $c > 0$ such that every weakly unanimous voting rule that is simple over some pair of candidates has susceptibility $\sigma \geq cN^{-1}$.*

**Theorem 4.7** *There is a constant $c > 0$ such that every voting rule that is simple over some pair of candidates and tops-only has susceptibility $\sigma \geq cN^{-1/2}$.*

In proving all of these lower bounds, we focus on profiles and beliefs $\phi$ that are concentrated on just two or three preference orderings. To understand why, recall that if we had not imposed any restrictions on beliefs in the definition (1.1) of susceptibility, then every voting rule would have susceptibility 1. Lower susceptibility is made possible by the smoothing of beliefs that the IID restriction achieves. A belief placing non-negligible probability on many preference orders is smoothed along many dimensions. Beliefs concentrated on a small number of orderings give coarser smoothing, and thus are more powerful in translating the discreteness of local changes in $f$ into lower bounds on susceptibility.

For the theorems involving monotonicity (4.4 and 4.5), the most important intuition behind the lower bounds is the interpretation of susceptibility as the probability of being pivotal. For the others, the main driving force is the coarseness of discrete approximation described in the previous paragraph.

| Efficiency | Regularity | Information | Bound | Theorem |
|---|---|---|---|---|
| Weakly unanimous | | | $\sigma \geq cN^{-3/2}$ | 4.1 |
| Weakly unanimous | Simple | | $\sigma \geq cN^{-1}$ | 4.6 |
| Pareto | Monotone | | $\sigma \geq cN^{-1/2}$ | 4.4 |
| Unanimous | | Tops-only | $\sigma \geq cN^{-1}$ | 4.2 |
| | Simple | Tops-only | $\sigma \geq cN^{-1/2}$ | 4.7 |
| Unanimous | Monotone | Tops-only | $\sigma \geq \sigma_N^*$ $(\sim cN^{-1/2})$ | 4.5 |

Table 4.1: Summary of lower-bound theorems

The remaining subsections sketch these proofs. Instead of following the order of exposition above, they are arranged in a more convenient way for presenting the tools. Subsection 4.2 covers Theorem 4.5. Since this is an exact bound, the proof is combinatorial. The remaining proofs are at least partly analytic, building on a

lemma introduced in Subsection 4.3 that bounds the variation in local averages of $f$ in terms of the susceptibility $\sigma$. Subsection 4.4 proves Theorem 4.4 for monotone voting rules, using the lemma to help formalize the pivotal intuition. Subsection 4.5 covers the results for tops-only voting rules, Theorems 4.2 and 4.7, while Subsection 4.6 proves the more general Theorems 4.1 and 4.6. These last two subsections exhibit a "meta-technique" for proving lower bounds on susceptibility: Begin with a proof by contradiction showing that susceptibility cannot be zero; then introduce error terms, and calculate how large the error terms need to be in order for the contradiction to disappear. In particular, Subsection 4.6 builds on Gibbard's [23] classic characterization of strategyproof probabilistic voting rules by including error terms in this way.

As for Theorem 4.3, we have already sketched the main idea of the construction; further details are left to Appendix H.

## 4.2 Monotone, tops-only voting rules

We begin with the proof of Theorem 4.5. Notice that for tops-only voting rules, monotonicity means that if a candidate $A_i$ wins at some profile $P$, and we change $P$ by replacing votes for candidates other than $A_i$ with votes for $A_i$, then $A_i$ remains the winner.

For intuition, consider the case of three candidates; an example of a monotone, tops-only voting rule is shown in Figure 4.2. Such a rule carves the simplex of possible vote profiles into a region where $A$ is chosen, a region where $B$ is chosen and a region where $C$ is chosen. Focus on the $B - C$ edge of the simplex. There is exactly one profile along this edge where the manipulator can be pivotal between $B$ and $C$ — either by changing his vote from $A$ to $B$, he changes the outcome from $C$ to $B$, or else (as in the figure) by changing his vote from $A$ to $C$, he changes the outcome from $B$ to $C$. Thus, if his true first choice is $A$, he can change the outcome from his third to second choice by manipulating. The critical distribution $\phi$ is then chosen to maximize the probability of this pivotal profile, and the bound follows via Lemmas 2.3 and 2.4.

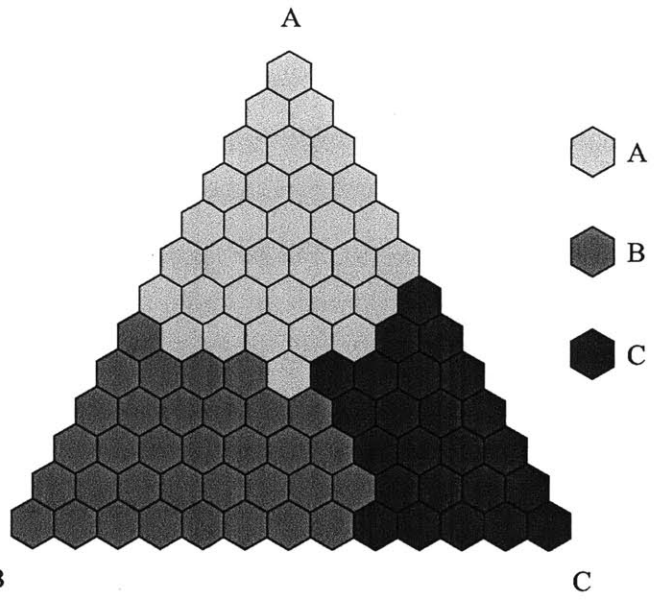


Figure 4.2: A unanimous, monotone, tops-only voting rule

The full proof, which is in Appendix G, modifies this argument to allow for arbitrarily many candidates. The proof also requires some extra work to deal with extreme shapes for the boundaries between regions, particularly when proving the equality case.

## 4.3 A crucial lemma

For the remaining results, we use analytic methods rather than purely combinatorial ones. Henceforth, we will need to refer to $N+1$ more often than $N$ directly, so put $\widetilde{N} = N + 1$.

The following definitions will be useful throughout the rest of this section. For any distribution $\phi \in \Delta(\mathcal{L})$, write $\overline{f}(\phi)$ for the distribution over candidates induced by $f$ when all $N+1$ votes are drawn IID from $\phi$. Also write $\overline{f}_{A_i}(\phi)$ for the probability of candidate $A_i$ in this distribution. Rather than studying $f$ directly, it will be more convenient to work with $\overline{f}$: the latter, being a continuous object, lends itself to analytic techniques.

From the point of view of the manipulator, reporting a preference $\succ'$, the distribution over outcomes is similar, but not identical, to $\overline{f}(\phi)$: the manipulator reports

$\succ'$ for sure, while the *other* $N$ preferences are drawn from $\phi$. The incentives to manipulate involve a comparison between two such distributions. As it turns out, this difference between two distributions is exactly equal to the directional derivative of $\overline{f}$, in the direction of changing preferences from $\succ$ to $\succ'$, up to a scaling factor. More precisely:

**Lemma 4.8** *Let $\succ, \succ'$ be any two orderings; let $\phi \in \Delta(\mathcal{L})$ and $\alpha \in [0,1]$. For $x \in [0,1]$, define*

$$\phi^x = \alpha((1-x) \ \succ + x \ \succ') + (1-\alpha)\ \phi.$$

*Then, the components of the derivative of the function $\overline{f}(\phi^x)$ are given by*

$$\frac{d}{dx}\left(\overline{f}_{A_i}(\phi^x)\right) = \alpha\widetilde{N} \ \cdot \ E_{IID(\phi^x)}[\mathbf{I}(f(\succ', P) = A_i) - \mathbf{I}(f(\succ, P) = A_i)].$$

The proof, by direct computation, is in Appendix D.

This leads to the following key lemma, which relates rates of change of $\overline{f}$ to the susceptibility of $f$.

**Lemma 4.9 (Local Average Lemma)** *Suppose the voting rule $f$ has susceptibility $\sigma$. There exists a constant $c$, independent of $N$ (or $f$ or $\sigma$), such that the following hold:*

*(a) Let $\succ, \succ'$ be any two orderings; let $\phi \in \Delta(\mathcal{L})$ and $\alpha \in [0,1]$. Then for any set $C^+$ consisting of the $L$ highest-ranked candidates under $\succ$, for some $L$, we have*

$$\sum_{A_k \in C^+} \overline{f}_{A_k}(\alpha \succ' + (1-\alpha)\phi) - \sum_{A_k \in C^+} \overline{f}_{A_k}(\alpha \succ + (1-\alpha)\phi) \leq \widetilde{N}\alpha\sigma. \qquad (4.1)$$

*(b) Let $\succ, \succ'$ be two orderings differing only by a switch of the adjacent candidates $A_i, A_j$; let $\phi \in \Delta(\mathcal{L})$ and $\alpha \in [0,1]$. Then for any set $C'$ of candidates not containing $A_i$ or $A_j$,*

$$\left| \sum_{A_k \in C'} \overline{f}_{A_k}(\alpha \succ' + (1-\alpha)\phi) - \sum_{A_k \in C'} \overline{f}_{A_k}(\alpha \succ + (1-\alpha)\phi) \right| \leq c\widetilde{N}\alpha\sigma. \qquad (4.2)$$

*(c) Suppose $f$ is tops-only. Let $\phi \in \Delta(\mathcal{C})$ and $\alpha \in [0,1]$. Then for any set $\mathcal{C}'$ of candidates not containing $A_i$ or $A_j$,*

$$\left| \sum_{A_k \in \mathcal{C}'} \overline{f}_{A_k}(\alpha A_i + (1-\alpha)\phi) - \sum_{A_k \in \mathcal{C}'} \overline{f}_{A_k}(\alpha A_j + (1-\alpha)\phi) \right| \leq c\widetilde{N}\alpha\sigma. \qquad (4.3)$$

**Proof:** We focus on proving (a), then check that the other parts follow immediately. Using the notation of Lemma 4.8, put $g(x) = \sum_{A_k \in \mathcal{C}^+} \overline{f}_{A_k}(\phi^x)$. We then have

$$\frac{dg}{dx} = \alpha\widetilde{N} \cdot E_{IID(\phi^x)}[\mathbf{I}(f(\succ', P) \in \mathcal{C}^+) - \mathbf{I}(f(\succ, P) \in \mathcal{C}^+)].$$

From (2.3), the right-hand side is at most $\sigma$. Therefore, $\frac{dg}{dx} \leq \alpha\widetilde{N}\sigma$ for all $x$, hence

$$\sum_{A \in \mathcal{C}'} \overline{f}_A(\alpha \succ' +(1-\alpha)\phi) - \sum_{A \in \mathcal{C}'} \overline{f}_A(\alpha \succ +(1-\alpha)\phi) = g(1) - g(0) \leq \alpha\widetilde{N}\sigma.$$

This proves (a).

For (b), notice that if $\mathcal{C}'$ consists of the $L$ highest-ranked candidates under $\succ$ (and hence also under $\succ'$) for some $L$, then (4.2) with $c = 1$ follows from part (a), applied once directly and once with $\succ$ and $\succ'$ reversed. If $\mathcal{C}'$ consists of the $L$ lowest-ranked candidates, then (4.2) with $c = 1$ likewise follows from part (a), taking $\mathcal{C}^+ = \mathcal{C} \setminus \mathcal{C}'$. Finally, any $\mathcal{C}'$ not containing $A_i$ or $A_j$ can be obtained by taking unions and differences of at most $M-2$ such highest- or lowest-ranked sets. Hence in general (4.2) holds with $c = M-2$, using the triangle inequality.

Part (c) is immediate from (b).

□

## 4.4 Monotone voting rules

We now take on Theorem 4.4, for monotone voting rules. Clearly it suffices to show the result when $N$ is sufficiently large.

Monotonicity again allows us to carve the simplex of vote profiles into regions where each candidate wins. The intuition of susceptibility as the probability of being

pivotal then applies: for the appropriate critical distribution, the probability of being on the boundary of two regions is of order $N^{-1/2}$, and we show that some such boundary is sloped so that a non-negligible fraction of the boundary profiles are in fact ones where manipulation is advantageous.

Lemma 4.10 formalizes this pivotal intuition, in the form that we need. The lemma focuses on a portion of the vote simplex spanned by three particular preferences $\succ, \succ', \succ''$. We suppose that there are two candidates $A_i, A_j$ who are ranked in the same way by $\succ'$ and $\succ''$; and that this simplex contains an $A_i$ region adjacent to an $A_j$ region, with the boundary between them sufficiently sloped relative to the $\succ' - \succ''$ edge of the simplex. If the manipulator expects the vote profile to lie near the boundary, he has an incentive to manipulate from $\succ'$ to $\succ''$ or vice versa, in order to help the more-preferred of the two candidates win. The size of this incentive is of order $N^{-1/2}$.

The formal statement of the lemma below is lengthy, but the idea is as above. The statement focuses on a parallelogram-shaped portion of the $\succ - \succ' - \succ''$ simplex, and assumes that throughout this parallelogram, $f$ chooses either $A_i$ or $A_j$, as illustrated in Figure 4.3. (The parallelogram shape makes the lemma easier to state, but is not crucial to the result.)

Condition (iii) of the lemma says the relevant regions are well-behaved enough to talk about the boundary between them. When applying the lemma, we use monotonicity to verify this condition. Conditions (iv) and (v) express that the boundary's slope is bounded below by $\kappa > 0$.

**Lemma 4.10** *Let $\kappa > 0$ be a constant. There exists a constant $c(\kappa) > 0$, depending only on $\kappa$, for which the following holds.*

*Suppose $f$ is a voting rule with $\widetilde{N}$ voters, having susceptibility $\sigma$. Let $\succ, \succ', \succ''$ be any three preference orderings. Let $0 \leq \underline{J} \leq \overline{J} \leq \widetilde{N}$ with $\overline{J} - \underline{J} > \kappa\widetilde{N}$. Let $0 \leq \overline{K} \leq \widetilde{N} - \underline{J}$. Define*

$$R = \{(J,K) \mid \underline{J} \leq J \leq \overline{J}; 0 \leq K \leq \overline{K}; J + K \leq \widetilde{N}\};$$

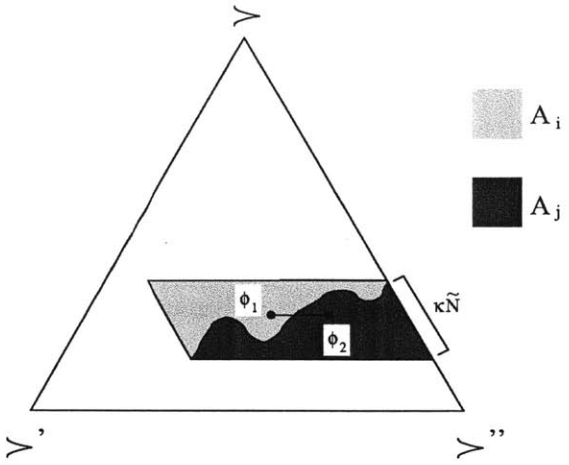


Figure 4.3: Illustration of Lemma 4.10

$$\text{and} \qquad P_{J,K} = \begin{pmatrix} J & \succ \\ K & \succ' \\ \widetilde{N} - J - K & \succ'' \end{pmatrix} \qquad \text{for } (J,K) \in R.$$

*Let $A_i, A_j$ be two different candidates. Suppose the following conditions hold:*

*(i)* $f(P_{J,K}) \in \{A_i, A_j\}$ *for all* $(J,K) \in R$;

*(ii)* $\succ'$ *and* $\succ''$ *rank* $A_i$, $A_j$ *in the same way relative to each other;*

*(iii) if* $(J,K) \in R$ *and* $f(P_{J,K}) = A_i$, *then* $f(P_{J+1,K}) = A_i$ *and* $f(P_{J+1,K-1}) = A_i$ *(whenever the relevant index pairs are in $R$);*

*(iv)* $f(P_{J,K}) = A_j$ *whenever* $K = 0$; *and*

*(v)* $f(P_{J,K}) = A_i$ *whenever* $K = \overline{K}$.

*Then*

$$\sigma \geq c(\kappa) N^{-1/2}. \tag{4.4}$$

The lemma is proven by identifying two distributions $\phi_1, \phi_2$ on either side of the boundary, with the distance between them on the order of $N^{-1/2}$, such that $\overline{f}(\phi_1) \approx A_i$ and $\overline{f}(\phi_2) \approx A_j$ (see the figure); and then applying the local average lemma. The proof is in Appendix G, as is the full proof of Theorem 4.4.

We proceed to describe the proof of Theorem 4.4 itself. The main strategy is illustrated in Figure 4.4, for the case of three candidates $A, B, C$. We focus on the behavior of $f$ on the $ABC-BCA-CAB$, $ABC-ACB-CAB$, $ACB-CAB-CBA$, and $ACB-CBA-BAC$ simplices, which are shown unfolded into a single plane in the figure. Monotonicity and Pareto efficiency give us $A$, $B$, and $C$ regions, with the shapes indicated. Note that $B$ cannot win anywhere in the middle two simplices, by Pareto efficiency. Consider the boundary between the $A$ and $C$ regions. If (as in the figure) the slope of this boundary is far from zero, then we can apply Lemma 4.10 to obtain the desired $cN^{-1/2}$ bound on susceptibility. (Actually, the application of the lemma is straightforward when the portion of the boundary in the middle two simplices of the figure is sloped. But when the sloped portion appears in the leftmost or rightmost simplex, a more detailed case analysis is needed, as sketched in Figure G.1 in Appendix G.)

It may be that the $A-C$ boundary is not sloped enough to apply the argument directly. However, Figure 4.4 shows only a part of the vote simplex. We can repeat the construction of this figure, replacing $A, B, C$ by $B, C, A$, respectively, or by $C, A, B$, respectively. Thus we obtain two more such figures. The proof of Theorem 4.4 shows that at least one of these figures contains a boundary whose slope is bounded away from zero, and then the argument goes through.

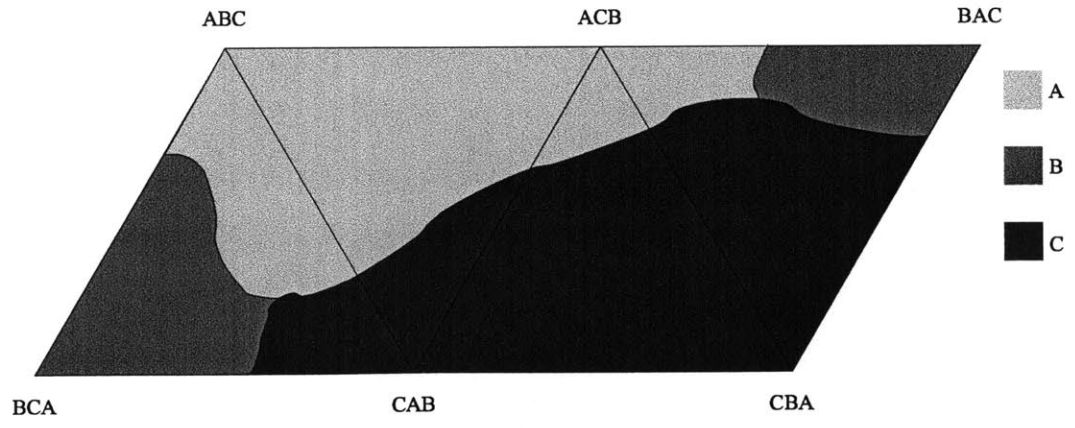


Figure 4.4: Proof of Theorem 4.4

## 4.5 Tops-only voting rules

Next, we show how to prove Theorems 4.2 and 4.7, on tops-only voting rules.

For Theorem 4.7, which gives a $cN^{-1/2}$ bound when the voting rule is simple, we take the approach of first sketching a proof that $\sigma > 0$, and then introducing error terms to find out explicitly how large $\sigma$ needs to be. Without loss of generality, suppose $f$ is simple over $B$ and $C$, and consider the values of $\overline{f}$ at several distributions in the $A - B - C$ simplex, as shown in Figure 4.5. We choose $\phi_1$ and $\phi_2$ so that $\overline{f}(\phi_1)$ puts high probability on $B$, $\overline{f}(\phi_2)$ puts high probability on $C$, and the distance between $\phi_1$ and $\phi_2$ is on the order of $N^{-1/2}$.

Suppose for contradiction $\sigma = 0$. Then $\overline{f}(\phi_1)$ and $\overline{f}(\phi_3)$ must put the same total weight on $A$ and $B$, by Lemma 4.9(c). Similarly, $\overline{f}(\phi_2), \overline{f}(\phi_3)$ put the same total weight on $A$ and $C$. We conclude that $\overline{f}(\phi_3)$ puts high probability on $A$. Next, again using Lemma 4.9(c), $\overline{f}(\phi_3), \overline{f}(\phi_4)$ put equal weight on $A$; and $\overline{f}(\phi_1), \overline{f}(\phi_4)$ put equal weight on $B$. Then $\overline{f}(\phi_4)$ puts high weight on both $A$ and $B$, which is a contradiction.

Now, repeat the argument without assuming $\sigma = 0$. Each time we apply Lemma 4.9, the conclusion remains the same as before, to within an approximation error of order $\sigma N^{-1/2}$. As long as the total approximation error accumulated in the course of the proof is smaller than some positive constant, we end with the same contradiction as before. Thus, the contradiction arises unless $\sigma > cN^{-1/2}$.

The formal proof of Theorem 4.7, following the above sketch, is short enough that we can include it here in the text.

**Proof of Theorem 4.7:** Assume that $f$ is simple over $B$ and $C$, and assume the threshold $K^*$ is $\leq \widetilde{N}/2$ (otherwise switch $B$ and $C$). Also let $c_0$ be the constant from Lemma 4.9.

We will assume that $f$ has susceptibility

$$\sigma < \frac{\sqrt{2}}{32c_0} \cdot \widetilde{N}^{-1/2} \tag{4.5}$$

and obtain a contradiction.

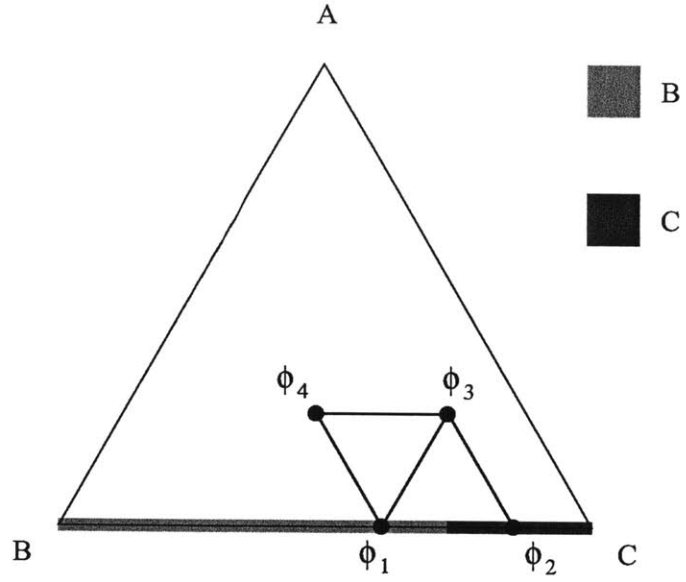


Figure 4.5: Proof of Theorem 4.7

Let

$$\phi_1 = (\alpha_1\ B, 1-\alpha_1\ C) \qquad \text{with} \qquad \alpha_1 = \frac{K^* + \sqrt{2\widetilde{N}}}{\widetilde{N}}.$$

Then $\overline{f}(\phi_1) = (\gamma_1\ B, 1-\gamma_1\ C)$, where $\gamma_1$ is the probability that at least $K^*$ voters vote $B$. The number $K$ of such voters is binomial with mean $\widetilde{N}\alpha_1 = K^* + \sqrt{2\widetilde{N}}$ and variance $\widetilde{N}\alpha_1(1-\alpha_1) \leq \widetilde{N}/4$, so by Chebyshev's inequality,

$$\Pr(K < K^*) \leq \Pr(|K - E[K]| \geq \sqrt{2\widetilde{N}}) \leq \frac{1}{8}.$$

Thus, $\gamma_1 \geq 7/8$.

Let

$$\phi_2 = (\alpha_2\ B, 1-\alpha_2\ C) \qquad \text{with} \qquad \alpha_2 = \max\left\{\frac{K^* - \sqrt{2\widetilde{N}}}{\widetilde{N}}, 0\right\}.$$

Then $\overline{f}(\phi_2) = (\gamma_2\ B, 1-\gamma_2\ C)$, where now $\gamma_2 \leq 1/8$ (this follows again by Chebyshev if $\alpha_2 > 0$, and if $\alpha_2 = 0$ then $\overline{f}(\phi_2) = C$).

We have $\phi_1 - \phi_2 = \Delta(B - C)$ where $\Delta = \alpha_1 - \alpha_2 \leq 2\sqrt{2/\widetilde{N}}$. By (4.5),

$$c_0 \widetilde{N} \Delta\sigma < \frac{1}{8}.$$

Let $\phi_3 = \phi_1 + \Delta(A - B) = \Phi_2 + \Delta(A - C)$ (this is again a valid probability distribution). Applying Lemma 4.9(c) to $\phi_1$ and $\phi_3$, with the set of candidates $\mathcal{C} \setminus \{A, B\}$, we find that $\overline{f}(\phi_3)$ places total weight at most $1/8 + c_0 \widetilde{N} \Delta\sigma < 1/4$ on candidates other than $A$ and $B$. Likewise, applying Lemma 4.9(c) to $\phi_2$ and $\phi_3$, with $\mathcal{C}' = \mathcal{C} \setminus \{A, C\}$, we conclude that $\overline{f}(\phi_3)$ places total weight $< 1/4$ on candidates other than $A$ and $C$. Consequently, $\overline{f}(\phi_3)$ places weight $> 1/2$ on $A$.

Now let $\phi_4 = \phi_1 + \Delta(A - C) = \phi_3 + \Delta(B - C)$. This is a valid distribution as long as $\phi_1$ places probability at least $\Delta$ on $C$. If $\widetilde{N}$ is large enough then

$$1 - \alpha_1 = \frac{\widetilde{N}/2 - \sqrt{2\widetilde{N}}}{\widetilde{N}} \geq \frac{2\sqrt{2\widetilde{N}}}{\widetilde{N}} \geq \Delta$$

so this requirement is satisfied.

Applying Lemma 4.9(c) to $\phi_1$ and $\phi_4$ with $\mathcal{C}' = \{B\}$ gives that $\overline{f}(\phi_4)$ places weight $> 3/4$ on $B$. Applying Lemma 4.9(c) again to $\phi_3$ and $\phi_4$ with $\mathcal{C}' = \{A\}$ gives that $\overline{f}(\phi_4)$ places weight $> 3/8$ on $A$. Since $3/4 + 3/8 > 1$, this is a contradiction. □

The proof of Theorem 4.2 builds on the above. We begin by considering various potential manipulations when the belief $\phi$ lies on the $B - C$ edge of the vote simplex. We show that if no such manipulation gives a gain greater than $cN^{-1}$ in expected utility, then $f$ is "approximately simple" over $B$ and $C$. From there we can repeat the proof of Theorem 4.7. The proof of Theorem 4.2 is in Appendix G.

## 4.6 General voting rules

Finally, we prove our most general result, Theorem 4.1, for any weakly unanimous voting rule. As an inexpensive by-product, we will also obtain Theorem 4.6, for simple and weakly unanimous voting rules.

The proof is closely modeled on Gibbard's [23] proof of the characterization of

strategyproof[7] probabilistic voting rules. Gibbard shows that any such voting rule is a convex combination of *unilateral* rules, in which only one agent's preference affects the outcome, and *duple* rules, where only two distinct outcomes are possible. Under our assumptions of anonymity and weak unanimity, the only such probabilistic voting rule is random dictatorship.

The connection between Gibbard's result and ours is made by the local average lemma, which says that if $f$ has low susceptibility, then the probabilistic voting rule $\widehat{f}(P) = \overline{f}(P/\widetilde{N})$ is approximately strategyproof. We retrace Gibbard's proof and keep track of error terms, showing that if $\widehat{f}$ is approximately strategyproof then it must be approximately a random dictatorship. Finally, we use the coarseness of approximation (since $f$ is deterministic) to show that $\widehat{f}$ cannot be too close to random dictatorship.

At a technical level, the proof of Gibbard's characterization of strategyproof probabilistic voting rules $g$ is based on equations of the form

$$g(\succ, P) - g(\succ', P) = g(\succ, P') - g(\succ', P') \tag{4.6}$$

for certain pairs of preferences $\succ, \succ'$ and opponent-profiles $P, P'$. If (4.6) were to hold for all $\succ, \succ', P, P'$, it would say that $g$ is linear (as a function of the number of voters having each preference). Combined with weak unanimity, this linearity would immediately imply that $g$ is random dictatorship. In fact, Gibbard's proof only shows (4.6) for certain $\succ, \succ', P, P'$, but these cover enough cases to give the needed linearity.

Gibbard's original proof was quite involved, but our assumptions of anonymity and weak unanimity make the argument less difficult. (See also [17] and [59] for streamlined versions of Gibbard's argument under the unanimity assumption only.)

The key tool used in our argument — a version of (4.6) with error terms — is given by the following lemma. The absolute value notation for vectors here refers to the $L_1$ norm.

**Lemma 4.11** *Let $\succ_1, \succ_2, \succ_3, \succ_4$ be preference orderings, and let $A_i, A_j, A_k, A_l$ be candidates (not necessarily distinct), with the following properties:*

[7] That is, those where truth-telling is a dominant strategy.

- $\succ_1, \succ_2$ *differ only by a switch of the adjacent candidates* $A_i, A_j$*;*

- $\succ_3, \succ_4$ *differ only by a switch of the adjacent candidates* $A_k, A_l$*;*

- $\{A_i, A_j\} \neq \{A_k, A_l\}$.

*Let* $\phi \in \Delta(\mathcal{L})$*, and let* $\alpha, \beta, \gamma \geq 0$ *with* $\alpha + \beta + \gamma = 1$*. Take* $c_0$ *to be the constant from Lemma 4.9. Then, if* $f$ *is a voting rule with susceptibility* $\sigma$*, we have the bound*

$$\left| \overline{f}\begin{pmatrix} \alpha & \succ_1 \\ \beta & \succ_3 \\ \gamma & \phi \end{pmatrix} - \overline{f}\begin{pmatrix} \alpha & \succ_2 \\ \beta & \succ_3 \\ \gamma & \phi \end{pmatrix} - \overline{f}\begin{pmatrix} \alpha & \succ_1 \\ \beta & \succ_4 \\ \gamma & \phi \end{pmatrix} + \overline{f}\begin{pmatrix} \alpha & \succ_2 \\ \beta & \succ_4 \\ \gamma & \phi \end{pmatrix} \right| \leq 16 c_0 \widetilde{N} \sigma. \quad (4.7)$$

The proof simply involves decomposing the four-way difference on the left-hand side of (4.7) into a sum of two differences in two ways, and applying Lemma 4.9 to each of these differences. The details are in Appendix G.

We now outline the proof of Theorem 4.1, via three lemmas, whose proofs are again in Appendix G. Focus on candidates $A, B, C$. We assume a fixed ordering for the remaining candidates, and write expressions such as $CAB\ldots$ to denote a preference beginning $CAB$, with the remaining candidates arranged in their fixed order.

We maintain throughout the assumption that $f$ is weakly unanimous, with susceptibility $\sigma$.

**Lemma 4.12** *There is a constant* $c_1 > 0$ *with the following property: if* $\sigma < c_1/N$*, then*

$$f(K\ CAB\ldots, \widetilde{N} - K\ CBA\ldots) = C \qquad \text{for all } K. \quad (4.8)$$

This is easy to show using beliefs along near the $CAB\ldots - CBA\ldots$ edge. If (4.8) were violated, we could find some such belief where the manipulator can increase the probability of $C$ by $c_1/N$ by manipulating from $CAB\ldots$ to $CBA\ldots$ or vice versa.

**Lemma 4.13** *Assume (4.8) holds. Let* $x, y, z, x', z'$ *be nonnegative numbers with* $x +$

$y + z = x' + y + z' = 1$. *Then*

$$\left| \left( \overline{f} \begin{pmatrix} x & ABC\ldots \\ y+z & BAC\ldots \end{pmatrix} - \overline{f} \begin{pmatrix} x+y & ABC\ldots \\ z & BAC\ldots \end{pmatrix} \right) - \left( \overline{f} \begin{pmatrix} x' & ABC\ldots \\ y+z' & BAC\ldots \end{pmatrix} - \overline{f} \begin{pmatrix} x'+y & ABC\ldots \\ z' & BAC\ldots \end{pmatrix} \right) \right| \leq 192 c_0 \widetilde{N} \sigma, \tag{4.9}$$

*where $c_0$ is the constant from Lemma 4.9.*

This key step is proven by repeated applications of Lemma 4.11. The bound (4.10) says that if we start at some distribution concentrated on the preference orderings $ABC\ldots$ and $BAC\ldots$, and move some fixed amount $y$ of mass from $ABC\ldots$ to $BAC\ldots$, then the change in $\overline{f}$ cannot depend too much on where we started. More simply put, $\overline{f}$ is approximately linear along the $ABC\ldots - BAC\ldots$ edge of the preference simplex.

**Lemma 4.14** *There exists some absolute constant $c_2$, independent of $N$, with the following property: for any weakly unanimous $f$, there exist some nonnegative values $x, y, z, x', z'$ with*

$$\left| \left( \overline{f} \begin{pmatrix} x & ABC\ldots \\ y+z & BAC\ldots \end{pmatrix} - \overline{f} \begin{pmatrix} x+y & ABC\ldots \\ z & BAC\ldots \end{pmatrix} \right) - \left( \overline{f} \begin{pmatrix} x' & ABC\ldots \\ y+z' & BAC\ldots \end{pmatrix} - \overline{f} \begin{pmatrix} x'+y & ABC\ldots \\ z' & BAC\ldots \end{pmatrix} \right) \right| \geq c_2 / \sqrt{\widetilde{N}}. \tag{4.10}$$

This simply quantifies how much the discreteness forces $\overline{f}$ to be far from linearity along the $ABC\ldots - BAC\ldots$ edge.

Theorem 4.1 now follows directly.

**Proof of Theorem 4.1:** Let $c_0, c_1, c_2$ be as in the three preceding lemmas. Either $\sigma \geq c_1/N$, and we are done; or else Lemma 4.12 applies, in which case the ensuing two lemmas imply that (4.10) and (4.11) both hold, from which $\sigma \geq c_2/192 c_0 \widetilde{N}^{3/2}$. □

If we impose the additional requirement of simplicity, the bound $c_2/\sqrt{\widetilde{N}}$ on the right side of (4.11) can be sharpened to a constant $c_3$, because $\overline{f}$ is not close to linear along the $ABC\ldots - BAC\ldots$ edge — its values are always close to $A$ or close to $B$, except right near the threshold. By repeating the proof of Theorem 4.1, we then find a lower bound for susceptibility of order $N^{-1}$ rather than $N^{-3/2}$, thus proving Theorem 4.6. The details are in Appendix G.

# 5 Conclusion

## 5.1 Summary

This paper has advanced a new way to quantify the susceptibility of decision-making mechanisms to strategic misbehavior, and argued its usefulness. We have focused here on voting rules as a canonical choice of application, but our approach is applicable quite broadly to other classes of mechanisms. Our measure of susceptibility is defined as the maximum expected utility an agent could gain by acting strategically rather than truthfully. To make this measure operational for voting rules, we needed a normalization of utility to the range $[0, 1]$, and an IID restriction on beliefs. Our measure has a simple interpretation in terms of behavior, in which agents trade off the benefits to manipulation against some (computational or psychological) costs.

To demonstrate the usefulness of this measure of susceptibility, we gave two classes of results. The first consisted of concrete estimates of the susceptibility of various voting systems. In particular (Table 3.1), we found that other systems previously identified as resistant to manipulation, including the Black, Copeland, Fishburn, minimax, and single transferable vote systems, actually are more susceptible than plurality rule, by our worst-case measure of incentives. We also identified qualitative properties of these voting systems that make them susceptible.

The second class of results consisted of lower bounds for the susceptibility of voting rules satisfying various efficiency, regularity, and informational properties (Table 4.1). These bounds illustrate how our measure can be used to study tradeoffs between

susceptibility and other properties. The proofs are built on a few, widely generalizable key ideas — such as susceptibility as the probability of being pivotal, the coarse smoothing provided by the IID assumption, and the broader technique of introducing quantitative error terms into impossibility proofs — thus showing how our measure of susceptibility can be worked with in practice.

## 5.2 Onwards

This is an appropriate place to discuss directions for future research.

At the most immediate level, there are many ways to extend the analysis here in technical directions. For example, one could seek lower bounds on susceptibility under other regularity conditions, or consider probabilistic voting rules. One could also consider different classes of probabilistic beliefs, in place of the IID model we have used here. For example, we have stuck to a model in which the number of other voters, $N$, is known with certainty, because this makes conditions such as monotonicity easy to formulate; but one might find the Poisson model [42, 43], which describes uncertainty about the population size as well as the distribution of preferences, to be more realistic. Our approach could also be extended to consider manipulation by coalitions.

A more important direction would be to apply our approach to measuring susceptibility to other classes of mechanism design problems. The third chapter of this dissertation, which studies the quantitative tradeoff between incentives to manipulate and efficiency in double auction environments, provides an example.

On a conceptual level, the approach to measuring susceptibility presented here would be greatly improved by incorporating some description of the decision process behind manipulation. The positive interpretation of our approach is based on a comparison of costs and benefits to the manipulator, but the modeling of costs here is simplistic — behaving strategically just always costs $\epsilon$. More realistically, it might be harder to manipulate in some mechanisms than others. A computational model that captures such distinctions would help in better understanding manipulative behavior.

Finally, a few words on how our approach fits into a broader agenda. There

are two main paradigms in mechanism design theory. One is the dominant-strategy paradigm [6, 58, 57]. This paradigm in effect evaluates mechanisms by their worst-case performance. Positive results, when they exist, are extremely robust to uncertainty about agents' beliefs, their assumptions about each other's strategic behavior, or the details of their preferences over lotteries; but existence of dominant strategies is a stringent requirement, and for many problems no dominant-strategy mechanism exists.

The second paradigm is Bayesian: the theorist presumes a common prior distribution over agents' types, assumes that agents maximize expected utility, and shows how to construct a mechanism that maximizes the expectation of some objective, such as welfare or revenue. The Bayesian paradigm allows for more positive results than dominant strategies (e.g. [15]), but often depends on stringent common knowledge assumptions that limit its practical usefulness [60].

The space in between the dominant-strategy and Bayesian approaches — explored by the recent literature on robust mechanism design [9, 14, 61] — may offer new avenues to obtain robust positive results. The approach of the present paper fits into this intermediate space: in the motivating model sketched in Subsection 1.3, we assume that the *voters* are Bayesian expected utility maximizers, but the *planner* takes a worst-case approach, with no probabilistic assumptions about the voters' preferences or beliefs (nor any requirement that voters' beliefs about each other correspond to the truth). More generally, integrating elements of the Bayesian and worst-case approaches will be valuable in bringing mechanism design theory closer to practice.

# A A consequentialist model

This appendix presents a game-theoretic model of voting rule choice by a social planner who cares about how well the outcome of the vote reflects the voters' preferences (but not about whether manipulation occurs per se). The model fleshes out the argument sketched verbally in Section 1.3 to describe how our measure of susceptibility would be involved in the choice of a voting rule. It is a formalization of the informal arguments that have long been used to justify dominant-strategy mechanisms, with a small cost of strategic behavior added in.

We imagine a planner choosing a voting rule for a society with $N$ voters and $M$ candidates. After the planner chooses the rule, the voters' types — meaning their preferences, beliefs, and their individual costs of manipulation — are realized. The voters cast their votes, and the election result is determined.

In the main model, the planner evaluates voting rules by their worst-case performance and is totally agnostic about what strategic voters will do, except that she believes voters will not strategize if they cannot benefit by more than $\epsilon$ from doing so. This extreme agnosticism is meant to represent the idea that the planner finds estimating strategic incentives to be much easier than predicting in detail how strategic voters will actually behave. (This models the trend in recent market design literature, such as [4, 12, 27, 28], which argues that incentives to manipulate in particular mechanisms go to zero, without going into exactly what the optimal manipulations would be.) However, our general point — that a quantitative measure of incentives to manipulate is relevant to choice of mechanism — does not depend on extreme agnosticism, as discussed further in Subsection A.5.

## A.1 Planner's preferences

We assume the planner cares ultimately about the relationship between the voters' preferences and the candidate who is elected. Thus, the planner has a utility function $U : \mathcal{C} \times ([0,1]^M)^{N+1} \to \mathbb{R}$, specifying her utility for each candidate contingent on all voters' preferences. To follow the ordinal framework of the main paper, we

assume that the planner's preferences depend only on the voters' ordinal rankings of candidates. So let $\succ^*$: $[0,1]^M \to \mathcal{L}$ be a given function, such that for each possible utility function $u \in [0,1]^M$, $u$ weakly represents $\succ^*(u)$. (The function $\succ^*$ describes how to convert cardinal preferences $u$ to ordinal rankings; the choice of $\succ^*(u)$ is nontrivial only when tie-breaking is necessary.) We assume there exists a function $V : \mathcal{C} \times \mathcal{L}^{N+1} \to \mathbb{R}$ such that

$$U(A_i; u_1, \ldots, u_{N+1}) = V(A_i; \succ^*(u_1), \ldots, \succ^*(u_{N+1}))$$

for all $A_i$ and all $u_1, \ldots, u_{N+1}$. Let $\underline{V}$ denote the minimum value attained by $V$, over all preference profiles and all outcomes $A_i$.

The planner is to choose from some nonempty set $\mathcal{F}$ of possible voting rules. We assume that every $f \in \mathcal{F}$ is surjective. We further assume that every $f$ satisfies

$$V(f(P); P) > \underline{V}$$

for every profile $P$. That is, the planner only considers voting rules with the following property: as long as all voters vote honestly, catastrophically bad outcomes are avoided.

## A.2 Mathematical states of nature

The planner expects that voters will behave strategically, if doing so is worth the cost $\epsilon$. In this case, she expects they will correctly solve their strategic optimization problem. However, the planner's task of predicting voters' behavior is much more complex than each individual voter's problem, since there may be many voting rules that the planner could consider, and many preferences and beliefs that each voter could potentially have. So we imagine that the planner does not know the solution to each voter's problem. We represent the planner's ignorance by ambiguity about how a voter's choice of vote maps to a distribution over outcomes (for a fixed distribution over others' votes).

More specifically, we model the planner's ignorance via *mathematical states of nature*. A mathematical state is a continuous function

$$\omega : \mathcal{F} \times \mathcal{L} \times \Delta(\mathcal{L}) \to \Delta(\mathcal{C}).$$

(Continuity is relevant only to the third argument, since $\mathcal{F}$ and $\mathcal{L}$ have discrete topologies.) Let $\Omega$ be the set of all possible mathematical states.

A mathematical state $\omega$ has the following interpretation: in this state, if the voting rule in use is $f$, a voter expects others' votes to follow distribution $\phi$, and he reports preference $\succ$, then he expects the outcome of the election will be distributed according to $\omega(f, \succ, \phi)$. There is one "true" mathematical state $\omega_0$, described by the actual outcomes of each voting rule: for all $f, \succ, \phi$, the distribution $\omega_0(f, \succ, \phi)$ is equal to the actual distribution over $f(\succ, P)$ that results if $P \sim IID(\phi)$. But the planner does not know the true state.

In any state $\omega$, the susceptibility of a voting rule $f$ is given by the analogue of (2.1):

$$\sigma_\omega(f) = \sup_{(\succ,\succ',u,\phi)\in\mathcal{Z}} \Big( u(\omega(f, \succ', \phi)) - u(\omega(f, \succ, \phi)) \Big).$$

(Here, and subsequently, we extend $u$ to lotteries over $\mathcal{C}$ by linearity.) This definition coincides with (2.1) in state $\omega_0$.

We assume that, although the planner does not know the true state, she has estimates on the susceptibility of each voting rule, which serve to narrow down the possible states. Specifically, for each $f \in \mathcal{F}$, she knows that the susceptibility of $f$ is less than some exogenous upper bound $\overline{\sigma}(f)$. We may have $\overline{\sigma}(f) > 1$, which corresponds to no knowledge about the susceptibility of $f$. (We do not model the process by which the planner learns of these upper bounds. We could also assume the planner knows lower bounds on susceptibilities; this would not change our results.) With these upper bounds, the set of states the planner considers possible is

$$\Omega^* = \{\omega \in \Omega \mid \sigma_\omega(f) < \overline{\sigma}(f) \text{ for all } f \in \mathcal{F}\}.$$

We assume that the planner's bounds are consistent with the truth: $\omega_0 \in \Omega^*$.

We will not need to specify a prior belief for the planner over $\Omega^*$, because we will assume she has maxmin preferences, as detailed below.

## A.3 Voters' preferences

Each voter has a utility function on candidates, $u : \mathcal{C} \to [0, 1]$, and a cost of behaving strategically, $\epsilon \in [\underline{\epsilon}, \overline{\epsilon}]$. Thus, the space of *basic types* of the voters is

$$\mathcal{T}_0 = [0, 1]^M \times [\underline{\epsilon}, \overline{\epsilon}].$$

The bounds $\underline{\epsilon}, \overline{\epsilon}$ are commonly known parameters, with $0 < \underline{\epsilon} < \overline{\epsilon}$ and $\underline{\epsilon} < 1$.

We assume there is some *rich type space* $\mathcal{T}$ of possible types for each voter, a compact Polish space, together with two continuous maps: a *basic type map* $\rho : \mathcal{T} \to \mathcal{T}_0$ and a *belief map* $\beta : \mathcal{T} \to \Delta(\mathcal{T})$. When a voter has rich type $t$, $\rho(t)$ is his basic type, and he believes other voters' rich types are drawn IID from the distribution $\beta(t)$. Let $\overline{\rho} : \Delta(\mathcal{T}) \to \Delta(\mathcal{T}_0)$ be the induced map: if $t$ is distributed according to $\psi$ on $\mathcal{T}$, then $\overline{\rho}(\psi)$ is the distribution of $\rho(t)$.

We assume the type space is rich enough so that the map

$$\rho \times (\overline{\rho} \circ \beta) : \mathcal{T} \to \mathcal{T}_0 \times \Delta(\mathcal{T}_0)$$

is surjective. That is, any combination of own basic type and (first-order) belief about others' basic types is possible.

Voters know the true mathematical state $\omega$.[8] Thus, each voter's type in the game-theoretic sense consists of his type in $\mathcal{T}$ as well as the state $\omega \in \Omega$. A (mixed) strategy for a voter specifies a distribution over $\mathcal{L}$, as a function of $t \in \mathcal{T}$ and $\omega \in \Omega$.

Voters have expected utility with respect to lotteries over candidates. The lottery that results from any particular vote is determined by the mathematical state. Thus,

[8]This assumption is not intended to mean literally that voters are computationally stronger than the planner; it is simply a technical shortcut to express that each voter can solve his own optimization problem.

in state $\omega$, for a voter with utility function $u$, if he votes $\succ$ and expects others to vote according to $\phi$, then his material payoff is $u(\omega(f, \succ, \phi))$.

## A.4 The game

The full timing of the game is as follows:

- The planner publicly announces a voting rule $f \in \mathcal{F}$.
- The voters' types in $\mathcal{T}$ are realized, as is the state $\omega \in \Omega^*$.

  (The fact that the true state is always $\omega_0$ will not be relevant, since we are studying the behavior of the planner, who does not know the true state.)
- Each voter chooses a preference ordering in $\mathcal{L}$ to report.
- The winning candidate is determined by applying $f$ to the reported preferences.

Now, we need to specify payoffs. Consider a voter in state $\omega$, with utility function $u$, and strategizing cost $\epsilon$. His utility if he truthfully reports preference $\succ^*(u)$, and other voters' votes are IID draws from $\phi$, is

$$u(\omega(f, \succ^*(u), \phi)).$$

If the voter reports any other preference $\succ'$, then his utility is

$$u(\omega(f, \succ', \phi)) - \epsilon.$$

As for the planner, her ex post preferences (given voters' utility functions and the outcome of the vote) are given by the function $U$. Her ex ante preferences are maxmin with respect to the voters' type profile and the mathematical state of nature: she wishes to maximize

$$\inf_{\substack{(t_1,\ldots,t_{N+1})\in\mathcal{T}^{N+1} \\ \omega\in\Omega^*}} E[U(f(\widehat{\succ}_1,\ldots,\widehat{\succ}_{N+1}); u_1,\ldots,u_{N+1})] \tag{A.1}$$

where the inf is over type profiles and mathematical states; each $u_i$ is the utility component of voter $i$'s basic type $\rho(t_i)$; and the expectation is over the reported preferences $\widehat{\succ}_i$ determined by the (possibly mixed) strategies of the voters in state $\omega$.

Finally, our solution concept is perfect Bayesian equilibrium, symmetric among the voters. That is, in each state, the voters play a symmetric Bayesian equilibrium (where the incomplete information is about each other's types); and given the strategies of the voters, the planner chooses a voting rule to maximize her utility (A.1).

With the game laid out in detail, we can finally state the proposition tying susceptibility to the planner's choice of a rule.

**Proposition A.1** *If there exists a voting rule $f \in \mathcal{F}$ whose known susceptibility bound $\overline{\sigma}(f)$ is at most $\underline{\epsilon}$, then in any equilibrium, the planner will choose such a rule. Specifically, she will choose $f$ to maximize $\min_{P \in \mathcal{L}^{N+1}} V(f(P), P)$, subject to $\overline{\sigma}(f) \leq \underline{\epsilon}$.*

*If no such $f$ exists, then in any equilibrium, the planner is indifferent among all voting rules; they all give her utility $\underline{V}$.*

The full proof is in Appendix D, but the argument is quite straightforward. If the planner can choose a voting rule with susceptibility less than $\underline{\epsilon}$, then she will be certain that all voters will vote truthfully, giving the outcome that the voting rule prescribes. On the other hand, if the planner cannot choose such a voting rule, then she cannot rule out the possibility that the voters will manipulate in the worst possible way, because the mathematical state and the voters' beliefs may be such that this manipulation is optimal for each voter.

## A.5 Variants

The preceding positive model gives a simple connection from our measure of susceptibility to a planner's choice of voting rule. We briefly sketch here several ways to extend the model, that would retain or strengthen this connection.

(a) We have considered here a model of a single election, leading to the conclusion that the planner would choose a voting rule whose susceptibility is known to be less than $\epsilon$, if one exists. With a large number of elections, the model could justify choosing a voting rule $f$ whose known susceptibility bound $\overline{\sigma}_f$ is as small as possible.

To be more specific, suppose that the planner anticipates the voting rule being used for many elections, some more important than others. Importance is represented by an upper bound $\overline{u}$ on voters' utilities from the outcome. Thus for each election there is a type space $\mathcal{T}_{\overline{u}}$, in which voters' utility functions have range $[0, \overline{u}]$ rather than $[0, 1]$; whereas the bounds $\underline{\epsilon}, \overline{\epsilon}$ on manipulation costs are constant across elections. The planner has a belief $\xi$ about the distribution of $\overline{u}$ across elections, with full support $[0, \infty]$. The planner's total utility is the long-run average of her utilities from each election. For a large number of elections, we can express this as an expectation. Thus the planner's utility becomes

$$\int_0^\infty \left( \inf_{\substack{(t_1,\ldots,t_{N+1}) \in \mathcal{T}_{\overline{u}}^{N+1} \\ \omega \in \Omega^*}} E[U(f(\widehat{\succ}_1, \ldots, \widehat{\succ}_{N+1}); u_1, \ldots, u_{N+1})] \right) d\xi(\overline{u}).$$

Then the planner's choice of voting rule depends on the tradeoff between susceptibility and the desirability of the outcomes that result under honest behavior. If the planner is very risk-averse in terms of outcomes — i.e. $\underline{V}$ is very low compared to other values of $V$ — then in equilibrium she will simply choose a voting rule $f \in \mathcal{F}$ whose susceptibility bound is as low as possible.

(b) We could also suppose that the planner has some inherent preference for non-consequentialist properties of the voting rule — say, regularity properties. This could be represented by preferences of the form

$$\inf_{\substack{(t_1,\ldots,t_{N+1}) \\ \omega}} E[U(f(\widehat{\succ}_1, \ldots, \widehat{\succ}_{N+1}); u_1, \ldots, u_{N+1})] + H(f)$$

where $H : \mathcal{F} \to \mathbb{R}$ is some function expressing the planner's preference over

these other properties. In such a model, the choice of voting rule would depend on the tradeoff between susceptibility and other properties.

(c) The preceding model makes extreme assumptions in terms of the players' knowledge. On one hand, the voters know the mathematical state perfectly: they are able to optimize their material payoffs exactly (if they choose to do so). On the other hand, the planner knows nothing about how voters will behave, except that they will not manipulate when the gain is definitely less than $\epsilon$.

However, in a model where agents might not manipulate optimally, or where the planner had some idea how agents manipulate, our general approach to quantifying incentives would remain relevant. Susceptibility would just have to be redefined, not as the maximum incentive for any manipulation, but as the maximum incentive specifically for manipulations that could potentially lead to undesirable outcomes (suitably defined).

The third chapter of this dissertation, on double auctions, explores the consequences of one such model in more detail. There, we assume no uncertainty about mathematical states. On the other hand, rather than optimizing exactly, the agents may potentially attempt any manipulation that gives them at least $\epsilon$ expected utility gain over truthfulness. The planner would like to minimize the maximum amount of inefficiency that can result from such manipulations. The analysis of this problem uses quite similar methods to the analysis of the tradeoff between susceptibility (as originally defined) and inefficiency.

# B Approval voting

In *approval voting*, each voter names a set of candidates, interpreted as the candidates who receive his approval. Whichever candidate receives the largest number of approvals wins. (As usual, we assume ties are broken alphabetically.)

Approval voting has often been specifically advocated as resistant to strategic manipulation [11, 21], so it is natural to ask how it fares under our approach to

measuring susceptibility. We have not addressed approval voting in the main paper because it does not fit into our framework. It requires voters to submit a set of approved candidates, rather than a ranking. More importantly, we have presumed that there is an unambiguous way to vote truthfully, for any given utility function $u$. In the case of approval voting, it is unclear how a voter should decide how many candidates to approve. This clashes with our motivating assumption — that truthful voting is costless — since the need for strategic calculation is now unavoidable. (Niemi [44] also argued that approval voting actually encourages strategic behavior for this reason.)

Still, it is possible to adapt our framework to formally cover approval voting, or vice versa. Here we present two possible ways of doing so. The discussion will be less detailed than in the main text.

## B.1 Multiple truthful strategies

We could simply allow that multiple strategies by voters are deemed truthful. In the case of approval voting, we might specify that it is truthful to approve a set $S \subseteq \mathcal{C}$ if $S$ consists of the $L$ most-preferred candidates, for some $L$. That is, $S$ is *sincere* for a utility function $u$ if, whenever $A_j \in S$ and $u(A_i) > u(A_j)$, then $A_i \in S$ as well. (This is the definition used in previous literature on approval voting [11, 21].) We could then define the susceptibility of approval voting to be the maximum gain from voting strategically, relative to voting sincerely.

To be precise, let $\mathcal{S}$ denote the set of all subsets of $\mathcal{C}$. The natural modification of the definition (2.1) for approval voting would then be

$$\sigma = \sup_{u,\phi} \left( \sup_{S'} (E_{IID(\phi)}[u(f(S', P))]) - \sup_{S} (E_{IID(\phi)}[u(f(S, P))]) \right), \tag{B.1}$$

where

- the outer supremum is over preferences $u \in [0,1]^M$ and beliefs $\phi \in \Delta(\mathcal{S})$;
- the first inner supremum is over arbitrary $S' \subseteq \mathcal{C}$;

- the second inner supremum is over $S$ that are sincere for $u$.

Notice that all suprema are taken over compact sets, so in fact we could write max instead of sup. (Alternatively, we could continue restricting $u$ to have no indifferences, as in the main text.)

With this approach, we can show that when $M \geq 4$, approval voting has susceptibility $\gtrsim 1/4$. In particular, its susceptibility does not go to zero as $N \to \infty$.

Let the manipulator's true preference be $BADC\ldots$, with the utility function

$$u(B) = 1, \qquad u(A) = 1/2 + \epsilon, \qquad u(D) = 1/2 - \epsilon, \qquad u(\text{any other candidate}) \leq \epsilon$$

for arbitrarily small $\epsilon$. Suppose the manipulator's belief $\phi$ is that each other voter approves $\{A, B\}$ with probability 1/2 and $\{C, D\}$ with probability 1/2.

With probability $\sim 1/2$, a majority of other voters vote $\{A, B\}$. In this case, the manipulator is pivotal between $A$ and $B$: if he votes for $B$ but not $A$, then $B$ wins; otherwise, $A$ wins. With probability $\sim 1/2$, a majority of other voters vote $\{C, D\}$, and the manipulator is pivotal between $C$ and $D$. (The other voters may be exactly evenly split, but the probability of this event goes to 0 as $N \to \infty$, so we disregard it.)

Hence, if the manipulator votes $\{B\}$, his expected utility is $\sim \frac{1}{2}u(B) + \frac{1}{2}u(C) \approx 1/2$. If he votes $\{B, A, D\}$, then his expected utility is $\sim \frac{1}{2}u(A) + \frac{1}{2}u(D) \approx 1/2$. And with any other sincere vote, his expected utility is $\sim \frac{1}{2}u(A) + \frac{1}{2}u(C) \approx 1/4$.

However, with the manipulation $S' = \{B, D\}$, his expected utility is $\sim \frac{1}{2}u(B) + \frac{1}{2}u(D) \approx 3/4$. Thus the gain from strategic voting expressed in (B.1) is approximately 1/4 as $N \to \infty$. Only by being insincere can the manipulator ensure that he gets the preferred outcome in both likely situations.

Why do our results here conflict with the view of previous literature, that approval voting resists manipulation? Unlike in Section 3, where the main issue was how to quantify manipulation, the basic difference here is one of modeling assumptions. The arguments in [11, 21] in favor of the strategic properties of approval voting assume that voters partition the candidates into three or fewer indifference classes. Indeed,

in the case $M = 3$, voting sincerely is always optimal ($\sigma$ as defined in (B.1) is zero). However, our argument shows that this finding breaks down severely as soon as $M \geq 4$. Indeed, Brams and Fishburn [11] were aware of this; they give an example that is almost identical to ours.

## B.2 Approval with status quo

An alternative way to model approval voting, without leaving the framework of the main paper, would be to specify an unambiguous choice of truthful vote for each preference order. For example, we could choose a particular candidate (here we will use $A$) as status quo, and declare that voters should approve all candidates who are preferred to the status quo.

Thus, the voting system *approval voting with status quo* is defined as follows: Each voter submits a preference order. Each candidate receives a score, defined as the number of voters who prefer her over $A$. The candidate with the highest score wins; ties are broken alphabetically. If every voter ranks $A$ first, then $A$ wins.

Then we can apply our usual definition (2.1) of susceptibility. In this case, we find that approval voting with status quo has susceptibility 1, similarly to $Q$-approval voting. Indeed, suppose that the manipulator has preference $CBA\ldots$ but expects that every other voter will vote $BCA\ldots$ with probability 1. Then, with probability 1, sincere voting will lead to the outcome $B$ (by alphabetical tie-breaking); whereas the manipulation $CA\ldots$ will lead to the better outcome $C$.

Thus, with this modeling approach, we again find approval voting to be highly susceptible to manipulation.

# C Computational tools

The present section gathers a collection of technical tools used in subsequent calculations. It includes proofs of the preliminary results stated in Subsection 2.2 of the main paper.

The following notation, not introduced in the main paper, will be useful here and

subsequently. For a vector $x = (x_1, \ldots, x_r)$, we will write $x_{-k}$ for the vector of all components except $x_k$, and $x_{-jk}$ for the vector of all components except $x_j$ and $x_k$; and $x_{-ijk}$ similarly. We will write $x_{j+k}$ for the sum of components $x_j$ and $x_k$. Notice that if $x_{-jk}$ and $\sum_l x_l$ are given then $x_{j+k}$ is uniquely determined.

One other useful bit of notation: if $f$ is a function of $N$ and $c$ a constant, we write $f(N) \overset{e}{\sim} c$ to say that $f(N)$ converges exponentially fast to $c$, i.e. $|f(N) - c| \lesssim e^{-\lambda N}$ for some $\lambda > 0$ (as in the statement of Lemma 2.2(a)).

**Lemma C.1 (Stirling's approximation)** *For any positive integer $K$,*

$$K! = \sqrt{2\pi K}(K/e)^K \iota \qquad \textit{with } 1 < \iota < e^{1/12K}.$$

We cite this without proof; see e.g. [1, eq. 6.1.38].

**Proof of Lemma 2.1:** Expand the probability explicitly, and apply Lemma C.1 to the factorials. Since the $\iota$ factors all tend to 1 as $N \to \infty$, we get

$$\begin{aligned}
\mathbf{P}\left( \begin{array}{c|c} x_N & \\ N - x_N & N;\ \begin{array}{c} \beta_N \\ 1-\beta_N \end{array} \end{array} \right) & \\
&= \frac{\sqrt{2\pi N}\left(\frac{N}{e}\right)^N \cdot \beta_N^{x_N}(1-\beta_N)^{N-x_N}}{\sqrt{2\pi x_N}\left(\frac{x_N}{e}\right)^{x_N} \cdot \sqrt{2\pi(N-x_N)}\left(\frac{N-x_N}{e}\right)^{N-x_N}} \\
&\sim \left(\frac{\beta_N N}{x_N}\right)^{x_N} \left(\frac{(1-\beta_N)N}{N-x_N}\right)^{N-x_N} \frac{1}{\sqrt{2\pi N \beta_N(1-\beta_N)}}.
\end{aligned}$$

We know $\beta_N \to \beta$ (since $|(\beta - \beta_N)N| < 2c$), so the result will follow if we can show

$$\left(\frac{\beta_N N}{x_N}\right)^{x_N} \left(\frac{(1-\beta_N)N}{N-x_N}\right)^{N-x_N} \to 1. \tag{C.1}$$

Now, the logarithm of the left-hand side of (C.1) is $Nh(x_N/N, \beta_N)$, where

$$h(\gamma, \delta) = \gamma(\ln\delta - \ln\gamma) + (1-\gamma)(\ln(1-\delta) - \ln(1-\gamma)).$$

The derivative of $h$ with respect to its first argument is

$$\frac{\partial h}{\partial \gamma} = \ln \frac{\delta}{\gamma} - \ln \frac{1-\delta}{1-\gamma}.$$

In particular, $\partial h/\partial \gamma$ is continuous on $(0,1) \times (0,1)$ and is zero when $\gamma = \delta$. We also have $h(\beta_N, \beta_N) = 0$, and so

$$\begin{aligned}
\left| Nh\left(\frac{x_N}{N}, \beta_N\right) \right| &\leq N \left| \frac{x_N}{N} - \beta_N \right| \cdot \max_{\gamma \in \left[\frac{x_N}{N}, \beta_N\right]} \left| \frac{\partial h}{\partial \gamma}(\gamma, \beta_N) \right| \\
&< c \cdot \max_{\gamma \in \left[\frac{x_N}{N}, \beta_N\right]} \left| \frac{\partial h}{\partial \gamma}(\gamma, \beta_N) \right| \\
&\to c \cdot \frac{\partial h}{\partial \gamma}(\beta, \beta) \\
&= 0.
\end{aligned}$$

(The notation assumes $x_N/N \leq \beta_N$, but of course an identical argument applies when $\beta_N < x_N/N$.) Then (C.1) follows. □

**Proof of Lemma 2.3:** Taking logs and ignoring the constant, we see the problem is to maximize $\sum_i x_i \ln \alpha_i$ subject to $\sum_i \alpha_i = 1$. This is a concave maximization problem; the solution is given by the first-order condition $x_i/\alpha_i = \lambda$ for all $i$, where $\lambda$ is the Lagrange multiplier on the constraint. Hence the $\alpha_i$ must be proportional to the $x_i$ at the maximum. □

**Lemma C.2** *For $1 \leq q < r$, we have*

$$\mathbf{P}\left( \begin{array}{c|c} \begin{matrix} x_1 \\ \vdots \\ x_r \end{matrix} & \begin{matrix} K; & \alpha_1 \\ & \vdots \\ & \alpha_r \end{matrix} \end{array} \right) = \mathbf{P}\left( \begin{array}{c|c} \begin{matrix} x_1 \\ \vdots \\ x_q \\ x \end{matrix} & \begin{matrix} K; & \alpha_1 \\ & \vdots \\ & \alpha_q \\ & \alpha \end{matrix} \end{array} \right) \cdot \mathbf{P}\left( \begin{array}{c|c} \begin{matrix} x_{q+1} \\ \vdots \\ x_r \end{matrix} & \begin{matrix} x; & \alpha_{q+1}/\alpha \\ & \vdots \\ & \alpha_r/\alpha \end{matrix} \end{array} \right)$$

*where $x = x_{q+1} + \cdots + x_r$ and $\alpha = \alpha_{q+1} + \cdots + \alpha_r$ (assuming $\alpha > 0$).*

This is the familiar decomposition property of the multinomial distribution: given that $K - x$ voters are of the first $q$ types and the remaining $x$ voters are of the

remaining $r-q$ types, the distribution of types among the last $x$ voters is independent of the distribution among the first $K-x$ voters (and in particular is again multinomial $\mathbf{M}(x; \alpha_{q+1}/\alpha, \ldots, \alpha_r/\alpha)$.

**Proof:** Immediate from the definitions. □

**Lemma C.3**

$$\sum_{x \text{ even}} \mathbf{P}(x, N-x \mid N; \alpha, 1-\alpha) = (1 + (1-2\alpha)^N)/2.$$

**Proof:** Write the right-hand side as $(((1-\alpha)+\alpha)^N + ((1-\alpha)-\alpha)^N)/2$; expanding by the binomial theorem, the terms with odd powers of $\alpha$ cancel and we get $\sum_{x \text{ even}} \binom{N}{x}(1-\alpha)^{N-x}\alpha^x$ which is the left-hand side. □

**Proof of Lemma 2.2:**

(a) Choose $\epsilon$ sufficiently small such that if $(\alpha_1, \ldots, \alpha_r) \in J$ and $|\beta_j - \alpha_j| < \epsilon$ for each index $j$, then $\beta_1, \ldots, \beta_r$ must still satisfy the inequalities $\mathcal{I}$. (We can do this since $J$ is compact and the inequalities $\mathcal{I}$ carve out an open set.) We can find $\kappa < 1$ such that

$$\frac{\alpha^\beta(1-\alpha)^{1-\beta}}{\beta^\beta(1-\beta)^{1-\beta}} < \kappa \qquad \text{for all } \alpha, \beta \in [0,1] \text{ with } \qquad |\beta - \alpha| \geq \epsilon, \qquad \text{(C.2)}$$

where we interpret $0^0$ as 1. Indeed, the denominator of the left side of (C.2) is bounded away from 0, whereas as $\alpha \to 0$ the numerator is $\leq \alpha^\epsilon$ and so converges uniformly to 0 for $\beta \in [\epsilon, 1]$; likewise as $\alpha \to 1$ the numerator is $\leq (1-\alpha)^{1-\epsilon}$ and so converges uniformly to 0 for $\beta \in [0, 1-\epsilon]$. This shows that for some $\eta > 0$, we can choose $\kappa < 1$ to ensure that (C.2) holds when $\alpha \leq \eta$ or $\alpha \geq 1-\eta$. Otherwise, use the fact that the logarithm of the left side of (C.2) is $\beta(\ln\alpha - \ln\beta) + (1-\beta)(\ln(1-\alpha) - \ln(1-\beta))$. This expression is continuous on the rectangle $[\alpha, \beta] \in [\eta, 1-\eta] \times [0,1]$, and takes its maximum value of zero only at $\alpha = \beta$ (by Lemma 2.3), and therefore is bounded strictly below 0 for $|\alpha - \beta| \geq \epsilon$. Statement (C.2) follows.

Now take any $(\alpha_1, \ldots, \alpha_r) \in J$. Consider any given index $j$, and any value $x_j$ with $|x_j/N - \alpha_j| > \epsilon$. Let $\beta_j = x_j/N$. The probability that the realized $j$-th component is $x_j$ is

$$\begin{aligned}
\mathbf{P}\left( \begin{array}{c} x_j \\ N - x_j \end{array} \middle| N; \begin{array}{c} \alpha_j \\ 1 - \alpha_j \end{array} \right) &= \mathbf{P}\left( \begin{array}{c} x_j \\ N - x_j \end{array} \middle| N; \begin{array}{c} \beta_j \\ 1 - \beta_j \end{array} \right) \times \\
&\qquad \left( \frac{\alpha_j^{\beta_j}(1-\alpha_j)^{1-\beta_j}}{\beta_j^{\beta_j}(1-\beta_j)^{1-\beta_j}} \right)^N \\
&\leq \kappa^N.
\end{aligned}$$

There are $r$ possible choices of index $j$ and at most $N+1$ values $x_j$ to consider for any given $j$, so the total probability that some event $|x_j/N - \alpha_j| > \epsilon$ occurs is at most $r(N+1)\kappa^N$. This bound still decays exponentially in $N$, and is independent of the choice of $(\alpha_1, \ldots, \alpha_r) \in J$.

(b) Fix arbitrarily small $\epsilon > 0$. We will show that

$$\frac{1}{2}\sqrt{\frac{2}{\pi(2\alpha_i + \epsilon)N}} \lesssim \mathbf{P}(S_N^{\mathcal{I}} \cap T_{ij,y} \mid N; \alpha_1, \ldots, \alpha_r) \lesssim \frac{1}{2}\sqrt{\frac{2}{\pi(2\alpha_i - \epsilon)N}} \qquad \text{(C.3)}$$

and the conclusion will follow by taking $\epsilon \to 0$.

Let $S'_N = \{(x_1, \ldots, x_r) \mid (2\alpha_i - \epsilon)N < x_i + x_j < (2\alpha_i + \epsilon)N\}$. By (a), the probability of drawing a profile in $S_N^{\mathcal{I}}$ and the probability of drawing a profile in $S'_N$ both go to 1 exponentially as $N \to \infty$.

Let $S_N^{par}$ be the set of profiles such that $x_i - x_j - y$ is even, or equivalently $x_i + x_j - y$ is even. Certainly $T_{ij,y} \subseteq S_N^{par}$. From Lemma C.2, $(x_i + x_j, \sum_{k \neq i,j} x_k)$ is multinomial with parameters $N; \alpha_i + \alpha_j, 1 - (\alpha_i + \alpha_j)$. So the probability of drawing a profile in $S_N^{par}$ is $(1 \pm (1 - 2(\alpha_i + \alpha_j))^N)/2$, by Lemma C.3. This converges exponentially to $1/2$ as $N \to \infty$.

Write $p_N$ for the probability that $P \in T_{ij,y}$, *conditional* on $P \in S'_N \cap S_N^{par}$. Because the probabilities of drawing profiles in $S_N^{\mathcal{I}}, S'_N, S_N^{par}$ converge exponentially

to $1, 1, 1/2$ respectively, it suffices to show that $p_N$ satisfies

$$\sqrt{\frac{2}{\pi(2\alpha_i + \epsilon)N}} < p_N < \sqrt{\frac{2}{\pi(2\alpha_i - \epsilon)N}} \tag{C.4}$$

and then (C.3) will follow.

For any given $N$, fix any value of the subvector $x_{-ij}$, such that

$$(1 - 2\alpha_i - \epsilon)N < \sum_{k \neq i,j} x_k < (1 - 2\alpha_i + \epsilon)N$$

and

$$x_{i+j} = N - \sum_{k \neq i,j} x_k \text{ is the same parity as } y.$$

Also write $x_{i+j}^{max}$ and $x_{i+j}^{min}$ for the maximum and minimum possible values of $x_{i+j}$ subject to these conditions. Note that whether or not $P \in S'_N \cap S_N^{par}$ depends only on $x_{-ij}$.

Conditional on the values $x_{-ij}$, the remaining coordinates $(x_i, x_j)$ are distributed $\mathbf{M}(x_{i+j}; 1/2, 1/2)$ by Lemma C.2. Moreover $x \in T_{ij,y}$ if and only if $x_i - x_j = y$, or equivalently $x_i = (x_{i+j} + y)/2$ (which is an integer). Hence, conditional on $x_{-ij}$, the probability that $x \in T_{ij,y}$ is

$$h_y(x_{i+j}) = \mathbf{P}\left( \begin{array}{c|cc} (x_{i+j} + y)/2 & x_{i+j}; & 1/2 \\ (x_{i+j} - y)/2 & & 1/2 \end{array} \right).$$

Applying Lemma 2.1 together with $x_{i+j}^{min} \sim (2\alpha_i - \epsilon)N, x_{i+j}^{max} \sim (2\alpha_i + \epsilon)N$ gives

$$\sqrt{\frac{2}{\pi(2\alpha_i + \epsilon)N}} \lesssim \min_{x_{i+j}} h_y(x_{i+j}) \leq \max_{x_{i+j}} h_y(x_{i+j}) \lesssim \sqrt{\frac{2}{\pi(2\alpha_i - \epsilon)N}},$$

where the maxima are taken over $x_{i+j} \in [x_{i+j}^{min}, x_{i+j}^{max}]$. For each realization of $x_{-ij}$, the conditional probability of $x \in T_{ij,y}$ lies between $\min h_y(x_{i+j})$ and $\max h_y(x_{i+j})$, so the overall probability of $x \in T_{ij,y}$ also lies in between these

bounds. At this point (C.4) follows.

As already shown, this in turn implies (C.3), and the proof of part (b) is complete.

□

**Proof of Lemma 2.4:** For any $K$, the maximum is attained by $\alpha = K/N$ by Lemma 2.3. Hence it suffices to study the behavior with respect to $K$ of the expression $\binom{N}{K}(K/N)^K((N-K)/N)^{N-K}$, or equivalently of $b(K) = K^K(N-K)^{N-K}/K!(N-K)!$. In particular, by symmetry it suffices to show that $b(K)$ is strictly increasing for $K \geq N/2$.

Put $c(K) = K^K/K!$. Notice that $c(K+1)/c(K) = (1+1/K)^K$ which is increasing in $K$ (this can be verified directly by taking the logarithm and differentiating). Hence for $K \geq N/2$ we have

$$\frac{b(K+1)}{b(K)} = \frac{c(K+1)c(N-K-1)}{c(K)c(N-K)} = \frac{c(K+1)}{c(K)} \Big/ \frac{c(N-K)}{c(N-K-1)} > 1$$

because $K > N-K-1$. □

Next we give a simple bound on the probability of large deviations under multinomial distributions.

**Lemma C.4** *For all $N, K, \alpha$,*

$$\mathbf{P}(K, N-K \mid N; \alpha, 1-\alpha) \leq e^{-N \cdot \frac{(\alpha - K/N)^2}{2}}.$$

(One can obtain a slightly stronger bound from Hoeffding's Inequality [24], but the proof here is self-contained.)

**Proof:** Consider the function $h(\alpha) = \ln \mathbf{P}(K, N-K \mid N; \alpha, 1-\alpha)$, whose maximum is at $\alpha = K/N$ by Lemma 2.3, and its value there is certainly at most 0. Moreover $d^2h/d\alpha^2 = -(K/\alpha^2 + (N-K)/(1-\alpha)^2)$. Now by Cauchy-Schwarz,

$$\left(\frac{K}{\alpha^2} + \frac{N-K}{(1-\alpha)^2}\right)(\alpha^2 + (1-\alpha)^2) \geq (\sqrt{K} + \sqrt{N-K})^2 \geq N.$$

Then $d^2h/d\alpha^2 \leq -N$ so $h(\alpha) \leq -N(\alpha - K/N)^2/2$. □

The next two results concern the quantity $\sigma_N^*$, defined in Subsection 3.1.

**Lemma C.5**

$$e^{-\frac{1}{3(N-1)}}\sqrt{\frac{2}{\pi N}} < \sigma_N^* < e^{\frac{1}{12N}}\sqrt{\frac{2N}{\pi(N^2-1)}}.$$

**Proof:** Put $\alpha = 1/2$ if $N$ is even and $(N-1)/2N$ if $N$ is odd. By Lemma C.1, write

$$\sigma_N^* = \binom{N}{\alpha N}\alpha^{\alpha N}(1-\alpha)^{(1-\alpha)N} =$$

$$\frac{\iota_N(N/e)^N\sqrt{2\pi N}}{[\iota_{\alpha N}(\alpha N/e)^{\alpha N}\sqrt{2\pi\alpha N}]\cdot[\iota_{(1-\alpha)N}((1-\alpha)N/e)^{(1-\alpha)N}\sqrt{2\pi(1-\alpha)N}]}\alpha^{\alpha N}(1-\alpha)^{(1-\alpha)N}$$

where the three $\iota_x$ terms satisfy $1 < \iota_x < e^{1/12x}$. Cancelling common factors reduces to

$$\frac{\iota_N}{\iota_{\alpha N}\iota_{(1-\alpha)N}}\cdot\sqrt{\frac{1}{2\pi N\alpha(1-\alpha)}}.$$

Both $\alpha N$ and $(1-\alpha)N$ are at least $(N-1)/2$, hence $e^{-1/3(N-1)} < \iota_N/\iota_{\alpha N}\iota_{(1-\alpha)N} < e^{1/12N}$; and $\alpha(1-\alpha) \in \{(N-1)^2/4N^2, 1/4\}$, hence the square-root term is either $\sqrt{2/\pi N}$ or $\sqrt{2N/\pi(N^2-1)}$. □

**Corollary C.6** *$\sigma_N^*$ is decreasing in $N$.*

**Proof:** For $N < 15$, $\sigma_N^* < \sigma_{N-1}^*$ can be verified by direct computation. For $N \geq 15$, Lemma C.5 implies that it is sufficient to check that

$$e^{\frac{1}{12N}}\sqrt{\frac{N}{N^2-1}} < e^{-\frac{1}{3(N-2)}}\sqrt{\frac{1}{N-1}} \tag{C.5}$$

or equivalently

$$e^{\frac{1}{12N}+\frac{1}{3(N-2)}}\sqrt{\frac{N}{N+1}} < 1. \tag{C.6}$$

Since $((N+1)/N)^{N+1} > e$, we have $\sqrt{N/(N+1)} < e^{-1/2(N+1)}$, so (C.6) follows from the inequality $1/12N + 1/3(N-2) < 1/2(N+1)$ which holds for $N \geq 15$. □

We provide a few more useful bounds.

**Lemma C.7** *If $x, y \geq K > 0$, then for all $\alpha$ we have*

$$\mathbf{P}(x, y \mid x+y; \alpha, 1-\alpha) \leq \frac{e^{1/12}}{\sqrt{\pi K}}.$$

**Proof:** By Lemma 2.3, the probability is maximized by taking $\alpha = x/(x+y)$. In this case, we can write the probability explicitly using Lemma C.1 and simplify as in Lemma C.5 to obtain

$$\mathbf{P}\left( \begin{array}{c|cc} x & x+y; & x/(x+y) \\ y & & y/(x+y) \end{array} \right) \leq \frac{e^{1/12}\sqrt{2\pi(x+y)}}{\sqrt{2\pi x}\sqrt{2\pi y}}.$$

Either $(x+y)/x \leq 2$ or $(x+y)/y \leq 2$, so we can cancel the numerator radical with one of the denominator radicals and a $\sqrt{2}$ factor, and the result follows. □

**Lemma C.8** *There exists an absolute constant $c > 0$ with the following property. For every positive integer $N$ and every nonempty subset $S \subseteq \{0, \ldots, N\}$, there exists $\alpha \geq \max(S)/N$ such that*

$$\sum_{K \in S} \left[ \mathbf{P}\left( \begin{array}{c|cc} K & N; & \alpha \\ N-K & & 1-\alpha \end{array} \right) - \mathbf{P}\left( \begin{array}{c|cc} K-1 & N; & \alpha \\ N-K+1 & & 1-\alpha \end{array} \right) \right] \geq \frac{c}{N}.$$

**Proof:** It suffices to prove the lemma when $S = \{K\}$. Indeed, since $\mathbf{P}(K, N-K \mid N; \alpha, 1-\alpha)$ is increasing in $K$ when $K \leq \alpha(N+1)$, every term on the left-hand side of the inequality in the lemma is nonnegative as long as $\alpha \geq \max(S)/N$, so it suffices to show that the term corresponding to $K = \max(S)$ is at least $c/N$.

So let $S = \{K\}$. If $K = N$ then take $\alpha = 1$. If $K = 0$ then take $\alpha = 0$. Otherwise, let $L = K + \lfloor \sqrt{K(N-K)/N} \rfloor$; we will show that $\alpha = L/N$ does the job. (Note that

$L \leq N$, i.e. $\alpha \leq 1$.) We have

$$\begin{aligned}
&\mathbf{P}\left(\begin{array}{c|cc} K & N; & \alpha \\ N-K & & 1-\alpha \end{array}\right) - \mathbf{P}\left(\begin{array}{c|cc} K-1 & N; & \alpha \\ N-K+1 & & 1-\alpha \end{array}\right) \\
&\qquad = \mathbf{P}\left(\begin{array}{c|cc} K & N; & \alpha \\ N-K & & 1-\alpha \end{array}\right) \cdot \left[1 - \frac{K}{N-K+1} \cdot \frac{1-\alpha}{\alpha}\right] \\
&\qquad = \mathbf{P}\left(\begin{array}{c|cc} L & N; & \alpha \\ N-L & & 1-\alpha \end{array}\right) \cdot \\
&\qquad\qquad \left[\prod_{k=K}^{L-1} \frac{k+1}{N-k} \frac{1-\alpha}{\alpha}\right] \cdot \left[1 - \frac{K}{N-K+1} \cdot \frac{1-\alpha}{\alpha}\right].
\end{aligned}$$

Now, the middle bracketed expression is a product consisting of $L-K$ factors, each of which is greater than

$$\frac{K}{N-K} \cdot \frac{1-\alpha}{\alpha} = \frac{K(N-L)}{L(N-K)} \geq 1 - \frac{1}{L-K}$$

(to verify the last inequality, cross-multiply and rearrange terms to find that it is equivalent to $(L-K)^2 N \leq L(N-K)$, which is true). Hence this product is

$$> \left(1 - \frac{1}{L-K}\right)^{L-K} \geq \frac{1}{4}$$

as long as $L - K \geq 2$. Otherwise, $L - K = 0$ and the middle product is empty, or else $L - K = 1$ and the middle product equals $(N-K-1)/(N-K) \geq 1/2$ (notice that if $K = N-1$ then $L = K$). Hence in every case the middle bracketed expression is $\geq 1/4$.

It therefore suffices to show that there is some constant $c'$ such that the bound

$$\mathbf{P}\left(\begin{array}{c|cc} L & N; & \alpha \\ N-L & & 1-\alpha \end{array}\right) \cdot \left[1 - \frac{K}{N-K+1} \cdot \frac{1-\alpha}{\alpha}\right] \geq \frac{c'}{N} \qquad \text{(C.7)}$$

always holds. We split into three cases.

- Suppose $K \leq N/2$ and $L > K$. The $\mathbf{P}(\cdots)$ factor is bounded below by $\sigma_N^* \gtrsim$

$\sqrt{2/\pi N}$, by Lemma 2.4. Also, $L - K > 0$ implies $(1/2)\sqrt{K(N-K)/N} \leq L - K \leq K$, so

$$\begin{aligned} 1 - \frac{K}{N-K+1} \cdot \frac{1-\alpha}{\alpha} &\geq 1 - \frac{K(N-L)}{L(N-K)} \\ &= \frac{(L-K)N}{L(N-K)} \\ &\geq \frac{L-K}{L} \\ &\geq \frac{L-K}{2K} \\ &\geq \frac{1}{4}\sqrt{\frac{N-K}{NK}} \\ &= \frac{1}{4}\sqrt{\frac{1}{K} - \frac{1}{N}} \\ &\geq \frac{1}{4}\sqrt{\frac{1}{N}} \end{aligned}$$

where there last step uses the assumption $K \leq N/2$. So each of the two factors on the left side of (C.7) is bounded below by a constant times $\sqrt{1/N}$.

- Suppose $K > N/2$ and $L > K$. In this case, we apply Stirling's approximation (C.1) as usual to observe that $\mathbf{P}(L, N-L \mid N; \alpha, 1-\alpha)$ is bounded below by a constant times $\sqrt{N/L(N-L)}$. Combining with the chain of inequalities from the previous case, we see that the left side of (C.7) is bounded below by a constant times

$$\sqrt{\frac{N}{L(N-L)}} \cdot \frac{1}{4}\sqrt{\frac{N-K}{NK}} \geq \frac{1}{4}\sqrt{\frac{N}{K(N-K)}} \cdot \sqrt{\frac{N-K}{NK}} = \frac{1}{4K} \geq \frac{1}{4N}.$$

- Finally suppose $L = K$. This can only happen for $K = 1$ or $N-1$, or for small $N$ (which we can ignore since the result is asymptotic), and so we verify (C.7) directly in these cases. We have $\mathbf{P}(L, N-L \mid N; \alpha, 1-\alpha) = ((N-1)/N)^{N-1} \geq 1/e$, a constant. If $K = 1$ then the second factor in (C.7) is $1/N$; if $K = N-1$ then this factor is $1/2$.

This verifies that (C.7) holds in every case.

□

**Lemma C.9** *Fix any positive constant $c$. If $N$ is taken large enough and $\alpha \leq c/\sqrt{N}$ then*

$$\sum_{K=\lceil 3c\sqrt{N}\rceil}^{N} \mathbf{P}\left( \begin{array}{c|cc} K & N; & \alpha \\ N-K & & 1-\alpha \end{array} \right) \leq \frac{1}{N}.$$

(Actually the left side goes to zero exponentially fast in $\sqrt{N}$, but this very crude bound is all we will need.)

**Proof:** Put $p(K) = \mathbf{P}(K, N-K \mid N; \alpha, 1-\alpha)$. We have

$$\frac{p(K+1)}{p(K)} = \frac{N-K}{K+1} \cdot \frac{\alpha}{1-\alpha} \leq \frac{N-K}{K} \cdot \frac{\alpha}{1-\alpha} \leq \frac{1}{2}$$

whenever $K \geq 2N\alpha$. Since $p(K) \leq 1$ for $K = \lceil 2N\alpha \rceil$, we have by induction $p(K) \leq 1/2^{K-\lceil 2N\alpha \rceil}$ for $K \geq 2N\alpha$, and therefore by the expression in the lemma statement is at most

$$\sum_{K=\lceil 3N\alpha \rceil}^{\infty} \frac{1}{2^{K-\lceil 2N\alpha \rceil}} = \frac{1}{2^{\lceil 3N\alpha \rceil - \lceil 2N\alpha \rceil - 1}} \leq \frac{1}{2^{c\sqrt{N}-2}} \lesssim \frac{1}{N}.$$

□

The remaining lemmas in this section are bounds on certain alternating sums of multinomial probabilities. These bounds are useful for the construction in Appendix H.

If $S$ is a set of positive integers, let $\sigma(S)$ and $\pi(S)$ denote, respectively, the sum and the product of elements of $S$ (with $\sigma(\emptyset) = 0, \pi(\emptyset) = 1$).

**Lemma C.10** *Fix $\epsilon > 0$ and $\underline{\alpha} \in (0, 1/2)$, and fix a positive integer $d$. There exists a threshold $N_0$ with the following property: For all $N > N_0$, all $\alpha \in [\underline{\alpha}, 1-\underline{\alpha}]$, all integers $K$, and all sets $S$ of positive integers with $|S| = d$,*

$$\left| \sum_{T \subseteq S} (-1)^{|T|} \mathbf{P}\left( \begin{array}{c|cc} K - \sigma(T) & N; & \alpha \\ N - K + \sigma(T) & & 1-\alpha \end{array} \right) \right| \leq \pi(S) N^{-d\left(\frac{1}{2}-\epsilon\right)}.$$

**Proof:** The expression inside the absolute value is (up to a sign) the coefficient of $z^K$ in the polynomial

$$Q_{\alpha,S}(z) = \left[\prod_{s\in S}(z^s - 1)\right] \cdot\cdot\, (\alpha z + (1-\alpha))^N.$$

This coefficient is also expressible as

$$\frac{1}{L}\sum_{l=1}^{L} \zeta^{-Kl} Q_{\alpha,S}(\zeta^l)$$

where $L$ is any integer greater than the degree of $Q_{\alpha,S}$ and $\zeta$ is a primitive $L$th root of unity. Therefore, it suffices to show that for some $N_0$ the following holds: whenever $N > N_0$, for all choices of $S$ and $\alpha$ and every complex number $z$ with $|z| = 1$,

$$|Q_{\alpha,S}(z)| \leq \pi(S) N^{-d\left(\frac{1}{2}-\epsilon\right)}. \tag{C.8}$$

We consider two cases for $z$. Let $\theta = \arg z$.

- Suppose $|\theta| < N^{-1/2-\epsilon}$. Then $|z-1| < N^{-(1/2-\epsilon)}$, from which

$$|z^s - 1| = \left|\sum_{t=0}^{s-1} z^t (z-1)\right| \leq s|z-1| < sN^{-(1/2-\epsilon)}$$

  and then multiplying across all $s \in S$, together with $|\alpha z + (1-\alpha)| \leq 1$, gives the result.

- Otherwise, $|\theta| \geq N^{-1/2-\epsilon}$. As long as $N$ is not too small,

$$\begin{aligned} |\alpha z + (1-\alpha)|^2 &= (1 - \alpha + \alpha\cos\theta)^2 + (\alpha \sin\theta)^2 \\ &= (1-\alpha)^2 + \alpha^2 + 2(1-\alpha)\alpha\cos\theta \\ &< (1-\alpha)^2 + \alpha^2 + 2(1-\alpha)\alpha\sqrt{1 - \frac{1}{4N^{1-2\epsilon}}} \end{aligned}$$

(this follows from $\cos^2 N^{-1/2-\epsilon} = 1 - \sin^2 N^{-1/2-\epsilon} > 1 - 1/4N^{1-2\epsilon}$)

$$\begin{aligned} &< (1-\alpha)^2 + \alpha^2 + 2(1-\alpha)\alpha\left(1 - \frac{1}{8N^{1-2\epsilon}}\right) \\ &= 1 - \frac{(1-\alpha)\alpha}{4N^{1-2\epsilon}} \\ &\leq 1 - \frac{c'}{N^{1-2\epsilon}} \end{aligned}$$

where $c' = (1-\underline{\alpha})\underline{\alpha}/4$. Hence

$$\begin{aligned} |\alpha z + (1-\alpha)|^N &< \left(1 - c'N^{-(1-2\epsilon)}\right)^{N/2} \\ &= \left[\left(1 - c'N^{-(1-2\epsilon)}\right)^{N^{1-2\epsilon}/2}\right]^{N^{2\epsilon}} \\ &< \left[\exp(-c')^{1/2}\right]^{N^{2\epsilon}} \\ &\leq N^{-d\left(\frac{1}{2}-\epsilon\right)}/2^d \end{aligned}$$

as long as $N$ is larger than some threshold that depends only on $\underline{\alpha}, \epsilon, d$. Since also $|z^s - 1| \leq 2$ for each $s \in S$, the bound (C.8) follows.

□

**Lemma C.11** *Fix a positive integer $d$. For any positive integer $h$, there exists a partition of the set $Z = \{0, 1, \ldots, 2^{hd} - 1\}$ into $2^h$ subsets $Z_0, Z_1, \ldots, Z_{2^h-1}$, of size $2^{h(d-1)}$ each, so that the following property is satisfied:*

*For any $\epsilon > 0$ and $\underline{\alpha} \in (0, 1/2)$, there exists a threshold $N_0$ such that for all $N > N_0$ all $\alpha \in [\underline{\alpha}, 1-\underline{\alpha}]$, and all integers $K$,*

$$\left|\sum_{x\in Z_i} \mathbf{P}\left(\begin{array}{c} K-x \\ N-K+x \end{array}\middle| N; \begin{array}{c} \alpha \\ 1-\alpha \end{array}\right) - \sum_{x\in Z_j} \mathbf{P}\left(\begin{array}{c} K-x \\ N-K+x \end{array}\middle| N; \begin{array}{c} \alpha \\ 1-\alpha \end{array}\right)\right| \leq 2^{h(d^2+d-1)} h N^{-d\left(\frac{1}{2}-\epsilon\right)}$$

*for any two sets $Z_i, Z_j$ of the partition.*

**Proof:** We first describe the partition. Consider each of the numbers $0, 1, \ldots,$

$2^{hd} - 1$ written out as a binary string with $hd$ digits. We assign each such number $x$ to a subset $Z_i$ as follows:

- Divide the $hd$ digits of $x$ into $h$ segments of $d$ digits each;
- next, replace each segment with a 0 or a 1, depending whether the number of 1's in that segment is even or odd;
- finally, read the resulting $h$-digit string as a binary number $i \in \{0, 1, \ldots, 2^h - 1\}$, and assign $x$ to $Z_i$.

It should be clear that each $Z_i$ consists of exactly $2^{h(d-1)}$ values $x$.

Now let $N_0$ be the threshold given by Lemma C.10, with the same $\epsilon, \underline{\alpha}, d$ as in the current lemma. Assume $N > N_0$, and let $\alpha \in [\underline{\alpha}, 1 - \underline{\alpha}]$ be arbitrary.

Define

$$\Sigma(\alpha, Z_i, N, K) = \sum_{x \in Z_i} \mathbf{P}\left( \begin{array}{c|cc} K - x & & \alpha \\ N - K + x & N; & 1 - \alpha \end{array} \right).$$

It suffices to show that if the binary representations of $i$ and $j$ differ by just one digit, then for all $K$,

$$|\Sigma(\alpha, Z_i, N, K) - \Sigma(\alpha, Z_j, N, K)| \leq 2^{h(d^2+d-1)} N^{-d\left(\frac{1}{2}-\epsilon\right)}. \tag{C.9}$$

Indeed, since one can get from any $i$ to any $j$ by at most $h$ single-digit changes, applying (C.9) repeatedly will then imply

$$|\Sigma(\alpha, Z_i, N, K) - \Sigma(\alpha, Z_j, N, K)| \leq 2^{h(d^2+d-1)} h N^{-d\left(\frac{1}{2}-\epsilon\right)} \tag{C.10}$$

which is exactly the assertion of the current lemma.

Without loss of generality, $i$ has a 0 in the $(r+1)$th position from the right, while $j$ has a 1 in that position; all other digits in the binary representations of $i$ and $j$ are the same. Then define three sets $Z'_\emptyset, Z'_i, Z'_j$:

- $Z'_\emptyset$ consists of all values of $x \in Z_i$ such that the $(dr+1)$th, $(dr+2)$th, $\ldots$, $(dr+d)$th digits from the right are all 0;

- $Z'_i$ consists of all numbers that can be represented as a sum of an even number of elements of the set $\{2^{dr}, 2^{dr+1}, \ldots, 2^{dr+d-1}\}$;
- $Z'_j$ consists of all numbers that can be represented as a sum of an odd number of elements of $\{2^{dr}, 2^{dr+1}, \ldots, 2^{dr+d-1}\}$.

Then, $Z_i$ consists of numbers that can be represented as a sum of an element of $Z'_\emptyset$ and one of $Z'_i$, and for each such number, the representation is unique. Likewise $Z_j$ consists of numbers that can be represented (uniquely) as a sum of an element of $Z'_\emptyset$ and one of $Z'_i$.

Applying the conclusion of Lemma C.10 with $S = \{2^{dr}, 2^{dr+1}, \ldots, 2^{dr+d-1}\}$, and using the easy bound $\pi(S) \le 2^{d(dr+d-1)}$, gives the following: for any $K$,

$$\left| \sum_{x \in Z'_i} \mathbf{P}\left( \begin{array}{c|cc} K-x & N; & \alpha \\ N-K+x & & 1-\alpha \end{array} \right) - \sum_{x \in Z'_j} \mathbf{P}\left( \begin{array}{c|cc} K-x & N; & \alpha \\ N-K+x & & 1-\alpha \end{array} \right) \right| \le 2^{d(dr+d-1)} N^{-d\left(\frac{1}{2}-\epsilon\right)}. \quad \text{(C.11)}$$

Now replace $K$ by $K-y$ for each possible $y \in Z'_\emptyset$, and sum over all $y$. We have

$$\sum_{y \in Z'_\emptyset} \sum_{x \in Z'_i} \mathbf{P}\left( \begin{array}{c|cc} K-y-x & N; & \alpha \\ N-K+y+x & & 1-\alpha \end{array} \right) = \sum_{x \in Z_i} \mathbf{P}\left( \begin{array}{c|cc} K-x & N; & \alpha \\ N-K+x & & 1-\alpha \end{array} \right)$$

and likewise for $Z'_j$ and $Z_j$. Thus, summing (C.11) over the $2^{(d-1)(h-1)}$ choices of $y \in Z'_\emptyset$ and then applying the triangle inequality gives

$$\begin{aligned} |\Sigma(\alpha, Z_i, N, K) - \Sigma(\alpha, Z_j, N, K)| &\le 2^{(d-1)(h-1)} \cdot 2^{d(dr+d-1)} N^{-d\left(\frac{1}{2}-\epsilon\right)} \\ &\le 2^{(d-1)(h-1)+d(hd-1)} N^{-d\left(\frac{1}{2}-\epsilon\right)}. \end{aligned}$$

Since $(d-1)(h-1) + d(hd-1) \le h(d^2+d-1)$, (C.9) follows. □

# D Assorted shorter proofs

**Proof of Proposition 3.2:**

(a) The argument is actually slightly more complex than that given in the main text, because the alphabetical tie-breaking leads to different cases depending on the parity of $N$.

If $N$ is even, let the manipulator's preferences be $ACB\ldots$, and let the opponent-profile $P$ be distributed according to $\phi = \left(\frac{1}{2}\ B, \frac{1}{2}\ C\right)$ (only voters' top choices matter). Then the manipulator cannot change the outcome except in the case $P = \left(\frac{N}{2}\ B, \frac{N}{2}\ C\right)$, in which case strategically voting for $C$ instead of $A$ beneficially changes the outcome from $B$ to $C$. If $N$ is odd, let the preferences be $ABC\ldots$, and let $\phi = \left(\frac{N-1}{2N}\ B, \frac{N+1}{2N}\ C\right)$. Then the manipulator is pivotal precisely when the opponent-profile is $P = \left(\frac{N-1}{2}\ B, \frac{N+1}{2}\ C\right)$, in which case voting for $B$ changes the outcome from $C$ to $B$. In both cases, the probability of being pivotal (2.3) is $\sigma_N^*$.

(b) First we prove the lower bound. Consider any small $\epsilon > 0$. Let the manipulator's preference be $ABC\ldots$, and consider a distribution $\phi \in \Delta(\mathcal{C})$ of the other voters' first-place votes such that $B$ and $C$ are each chosen with probability $\frac{1}{M} + \epsilon$, and every other candidate is chosen with probability $\frac{1}{M} - \frac{2\epsilon}{M-2}$.

Consider susceptibility as formulated in (2.3), where the proposed manipulation $\succ'$ is one that ranks $B$ first, and the set $\mathcal{C}^+$ of desirable candidates is $\{A, B\}$. Write the relevant expectation as

$$\sigma \geq \sum_P \left[\mathbf{I}(f(\succ', P) \in \mathcal{C}^+) - \mathbf{I}(f(\succ, P) \in \mathcal{C}^+)\right]\mathbf{P}(P \mid N; \phi). \tag{D.1}$$

(We write $\geq$ rather than $=$, since we are considering a specific distribution $\phi$ rather than the max.) Say that an opponent-profile $P$ is *relevant* if $B$ and $C$ both receive a vote share between $1/M + \epsilon/2$ and $1/M + 3\epsilon/2$, and every other candidate receives less than $1/M$ of the vote. By Lemma 2.2(a), the probability that the realized profile is relevant is $\overset{e}{\sim} 1$, so we need only consider the contribution of the relevant profiles to (D.1). For any such profile (assuming $N$ is large enough), no matter what the manipulator does, the outcome will

be either $B$ or $C$. The relevant profiles that contribute to (D.1) are exactly the ones where the manipulator is pivotal in changing the outcome from $C$ to $B$ — that is, the ones for which $B$ receives exactly one less vote than $C$. It follows from Lemma 2.2(b) that the total probability of these profiles is $\sim (1/2)\sqrt{1/\pi\left(\frac{1}{M}+\epsilon\right)N}$. (Here the lemma applies with $B, C$ corresponding to the indices $i, j$, and $y = -1$. Note that the definition of a relevant profile is a set of linear inequalities on the vote shares.)

Thus we have

$$\sigma \gtrsim \frac{1}{2}\sqrt{\frac{1}{\pi\left(\frac{1}{M}+\epsilon\right)N}}.$$

Taking $\epsilon \to 0$ gives the lower bound in Proposition 3.2(b).

Now we prove the upper bound. For each value of $N$, consider the true preference, manipulation, and belief $\phi$ that attain the maximum in (2.3). (These may vary depending on $N$, but we will not bother to make this dependence explicit in the notation.) Suppose that, for a given $N$, the manipulator's true first choice is $A_i$ and the reported first choice is $A_j$. This manipulation can be beneficial only if it changes the outcome from $A_k$, for some $k \neq i, j$, to $A_j$. For each $k$, let $S_{kj}$ be the set of all $N$-profiles $P$ such that $f(A_i, P) = A_k$ and $f(A_j, P) = A_j$; and let $S_{\to j} = \cup_{k \neq i,j} S_{kj}$. We wish to show that $\mathbf{P}(S_{\to j} \mid N; \phi) \lesssim \sqrt{M/\pi N}$.

Now, consider again any fixed $\epsilon > 0$. For each $k \neq i, j$, we have

$$\max_{\phi:\ \phi_j \geq (1+\epsilon)\phi_k} \mathbf{P}(S_{kj} \mid N; \phi) \overset{e}{\sim} 0. \tag{D.2}$$

Indeed, each opponent-profile $P = (x_1, \ldots, x_M) \in S_{kj}$ has $x_j + 1, x_k \geq N/M$, and also $x_j = x_k$ or $x_j = x_k - 1$. Consider such a profile $P$. Let $p(x_{-jk})$ be the conditional probability of realizing $P$, given that the components $x_{-jk}$ are realized. By Lemmas C.2 and C.4,

$$p(x_{-jk}) = \mathbf{P}\left( \begin{matrix} x_j \\ x_k \end{matrix} \;\middle|\; x_j + x_k;\ \begin{matrix} \phi_j/(\phi_j+\phi_k) \\ \phi_k/(\phi_j+\phi_k) \end{matrix} \right) \leq e^{-(x_j+x_k)\cdot\left(\frac{x_j}{x_j+x_k} - \frac{\phi_j}{\phi_j+\phi_k}\right)^2/2}.$$

The squared expression in the exponent is bounded away from zero, while the $x_j + x_k$ factor is $\geq N/M$, so the upper bound goes to zero exponentially in $N$. So, given any value of $x_{-jk}$, the conditional probability of realizing values of $x_j$ and $x_k$ for which the resulting profile is in $S_{kj}$ is bounded above by an expression that decays exponentially in $N$. Hence the *unconditional* probability of $S_{kj}$ satisfies this same exponential bound, and (D.2) holds.

On the other hand, the worst-case belief $\phi$ cannot have $\mathbf{P}(S_{\rightarrow j} \mid N; \phi) \stackrel{e}{\sim} 0$, since we already proved this probability satisfies a lower bound on the order of $\sqrt{1/N}$. Thus, as long as $N$ is large enough, there must be some $k^*$ such that $\phi_j < (1+\epsilon)\phi_{k^*}$. (This $k^*$ may not be unique, and may vary depending on $N$.)

Next, we claim that for any value of $x_{-jk^*}$ there is at most one way of choosing $x_j, x_{k^*}$ (given the additional constraint $\sum_l x_l = N$) so that the resulting $N$-profile lies in $S_{\rightarrow j}$. Indeed, suppose for a contradiction that $(x_j, x_{k^*}, x_{-jk^*}) \in S_{\rightarrow j}$, and also $(x_j + s, x_{k^*} - s, x_{-jk^*}) \in S_{\rightarrow j}$ for some positive integer $s$. Then, in particular,

$$f(x_i, x_j + 1, x_{k^*}, x_{-ijk^*}) = A_j; \tag{D.3}$$

$$f(x_i + 1, x_j + s, x_{k^*} - s, x_{-ijk^*}) = A_l \neq A_i, A_j. \tag{D.4}$$

If $s \geq 2$, then the profile in (D.4) gives a (weakly) greater advantage for $j$ relative to $l$ than the profile in (D.3) does, so if plurality rule chooses $A_j$ in (D.3) it should choose $A_j$ in (D.4) also, a contradiction. And if $s = 1$, then the profile in (D.4) differs from that in (D.3) by a vote shift from $A_{k^*}$ to $A_i$, which cannot change the winner from $A_j$ to $A_l$ — a contradiction again. Thus the claim holds.

Consider any $x_{-jk^*}$ such that there exist $x_j, x_{k^*}$ for which the resulting profile lies in $S_{\rightarrow j}$. We will again bound the probability of realizing this profile, conditional on $x_{-jk*}$. For this pivotal profile, we must have $x_{j+k^*} \geq x_j \geq (N+1)/M - 1 \geq N(1/M - \epsilon)$ (as long as $N$ is large). The conditional probability of realizing

$(x_j, x_{k^*})$ given $x_{-jk^*}$ is

$$p(x_{-jk^*}) = \mathbf{P}\left( \begin{array}{c|cc} x_j & x_{j+k^*}; & \phi_j/(\phi_j+\phi_{k^*}) \\ x_{k^*} & & \phi_{k^*}/(\phi_j+\phi_{k^*}) \end{array} \right).$$

(i) If $x_j > (1+2\epsilon)x_{k^*}$ then this probability $p(x_{-jk^*})$ is bounded above by an expression that decays exponentially in $x_{j+k^*}$ (by Lemma 2.2(a) and $\phi_j < (1+\epsilon)\phi_{k^*}$). In particular, across all choices of $x_{-jk^*}$ such that the corresponding profile in $S_{\to j}$ satisfies $x_j > (1+2\epsilon)x_{k^*}$, the probability $p(x_{-jk^*})$ is bounded above uniformly by a quantity that decays exponentially in $N$.

(ii) If $x_j \leq (1+2\epsilon)x_{k^*}$, then (since we also have $x_j + 1 \geq x_{k^*}$) we get $x_{j+k^*} \geq N(2/M - 3\epsilon)$. Hence

$$p(x_{-jk^*}) \leq \max_{\substack{x+y \geq N(2/M-3\epsilon) \\ x \leq (1+2\epsilon)y \\ y \leq (1+2\epsilon)x}} \mathbf{P}(x, y \mid x+y; \alpha_x, \alpha_y).$$

For given $x+y$, the choices of $x, y, \alpha_x, \alpha_y$ that attain the max are given by Lemmas 2.3 and 2.4, and we obtain

$$p(x_{-jk^*}) \leq \max_{K \geq N(2/M-3\epsilon)} \mathbf{P}\left(x, y \;\middle|\; K; \frac{x}{K}, \frac{y}{K}\right)$$

$$\text{with } x = \left\lceil \frac{K}{2+2\epsilon} \right\rceil, y = K - x.$$

Denote the expression inside this maximum by $\widetilde{p}(K)$.

We have thus shown that the conditional probability of realizing $(x_j, x_{k^*})$ forming a profile in $S_{\to j}$, given $x_{-jk^*}$, satisfies

$$p(x_{-jk^*}) \leq \max\{ce^{-\lambda N}, \max_{K \geq N(2/M-3\epsilon)} \widetilde{p}(K)\}.$$

(Here $c, \lambda$ are some positive values.) This inequality applies to the conditional

probability of obtaining a profile $x \in S_{\to j}$, given $x_{-jk^*}$. So it also applies to the unconditional probability of drawing a profile in $S_{\to j}$:

$$\mathbf{P}(S_{\to j} \mid N; \phi) \leq \max\{ce^{-\lambda N}, \max_{K \geq N(2/M - 3\epsilon)} \widetilde{p}(K)\}.$$

Now, Lemma 2.1 gives

$$\widetilde{p}(K) \sim \sqrt{\frac{1}{2\pi K \left(\frac{1}{2+2\epsilon}\right)\left(\frac{1+2\epsilon}{2+2\epsilon}\right)}}.$$

Hence

$$\mathbf{P}(S_{\to j} \mid N; \phi) \lesssim \max\left\{ce^{-\lambda N}, \sqrt{\frac{1}{2\pi N \left(\frac{2}{M} - 3\epsilon\right)\left(\frac{1}{2+2\epsilon}\right)\left(\frac{1+2\epsilon}{2+2\epsilon}\right)}}\right\}.$$

Clearly, for $N$ large enough the square-root term dominates.

Finally, taking $\epsilon \to 0$ gives us the simpler asymptotic upper bound $\sqrt{M/\pi N}$, which is what we wanted to show.

□

**Proof of Proposition 3.3:**

Given Proposition 3.2(a), we need only show $\sigma_N^{plur} \leq \sigma_N^*$. Consider any true preference for the manipulator and proposed manipulation. For this proof only, label the candidates so that the manipulator's preference is $ABC$, not necessarily corresponding to the tie-breaking order. A manipulation from $A$ to $C$ can never be beneficial; manipulation to $B$ can be beneficial only when it changes the winner from $C$ to $B$. So we need to show that the probability of being pivotal from $C$ to $B$ is at most $\sigma_N^*$. Let

$$S_0 = \{(x_A, x_B, x_C) \mid x_B = x_C - 1 \geq x_A\},$$

$$S_1 = \{(x_A, x_B, x_C) \mid x_B = x_C \geq x_A + 1\}.$$

The relevant set of pivotal profiles is contained either in $S_0$ or $S_1$ (depending on which of $B, C$ wins a tiebreaker), so we just need to show that for any $\phi$, both $S_0$ and $S_1$

are events of total probability at most $\sigma_N^*$.

Consider the $\phi$ that maximizes $\mathbf{P}(S_0 \mid N; \phi)$. Write $\phi = (\phi_A, \phi_B, \phi_C)$. We then have $\phi_C \geq \phi_A$. Proof: Suppose not. Then

$$\begin{aligned}
&\frac{d}{d\epsilon}\left[\mathbf{P}(S_0 \mid N; \phi_A - \epsilon, \phi_B, \phi_C + \epsilon)\right] \\
&\quad = \frac{d}{d\epsilon}\left[\sum_{(x_A,x_B,x_C)\in S_0} \frac{N!}{x_A!x_B!x_C!}(\phi_A - \epsilon)^{x_A}\phi_B^{x_B}(\phi_C + \epsilon)^{x_C}\right] \\
&\quad = \sum_{(x_A,x_B,x_C)\in S_0} \frac{N!}{x_A!x_B!x_C!}(\phi_A - \epsilon)^{x_A}\phi_B^{x_B}(\phi_C + \epsilon)^{x_C} \cdot \left(\frac{x_C}{\phi_C + \epsilon} - \frac{x_A}{\phi_A - \epsilon}\right).
\end{aligned}$$

For $\epsilon$ close to 0, the last factor in parentheses is always positive (since $x_C \geq x_A$ throughout $S_0$). So changing the belief from $(\phi_A, \phi_B, \phi_C)$ to $(\phi_A - \epsilon, \phi_B, \phi_C + \epsilon)$ increases the probability of drawing a profile in $S_0$, contrary to the assumption that the belief was chosen to maximize this probability.

Exactly the same reasoning applies for $S_1$. Thus it suffices to show that each of $S_0, S_1$ has probability at most $\sigma_N^*$, assuming that the belief $\phi = (\phi_A, \phi_B, \phi_C)$ satisfies $\phi_A \leq \phi_C$. In particular, we may assume $\phi_A \leq 1/2$.

We need to show four things:

(i) when $N$ is odd, the probability of drawing a profile in $S_0$ is at most $\sigma_N^*$;

(ii) when $N$ is odd, the probability of $S_1$ is at most $\sigma_N^*$;

(iii) when $N$ is even, the probability of $S_0$ is at most $\sigma_N^*$;

(iv) when $N$ is even, the probability of $S_1$ is at most $\sigma_N^*$.

First consider (i), so $N$ is odd. Then, for $(x_A, x_B, x_C) \in S_0$, we have $x_A$ even and at most $x_A^{max} = 2\lfloor N/6 \rfloor$, so

$$\mathbf{P}(S_0 \mid N; \phi) = \sum_{\substack{x_A \text{ even} \\ 0 \leq x_A \leq x_A^{max}}} \mathbf{P}\left(\begin{array}{c} x_A \\ N - x_A \end{array} \middle| N; \begin{array}{c} \phi_A \\ 1 - \phi_A \end{array}\right) \mathbf{P}\left(\begin{array}{c} x_B \\ x_C \end{array} \middle| N - x_A; \begin{array}{c} \phi'_B \\ \phi'_C \end{array}\right)$$

by Lemma C.2 (where $\phi'_B = \frac{\phi_B}{\phi_B + \phi_C}$, $\phi'_C = \frac{\phi_C}{\phi_B + \phi_C}$). Since the relevant $x_B, x_C$ are equal

or differ by 1, Lemma 2.3 gives $\mathbf{P}(x_B, x_C \mid N - x_A; \phi'_B, \phi'_C) \leq \sigma^*_{N-x_A}$, which in turn is at most $\sigma^*_{N-x_A^{max}}$ by Corollary C.6. Hence, the above sum is at most

$$\begin{aligned}
&\mathbf{P}\left( \begin{array}{c|c} 0 & \\ N & N; \end{array} \begin{array}{c} \phi_A \\ 1-\phi_A \end{array} \right) \sigma^*_N + \\
&\qquad \sum_{\substack{x_A \text{ even} \\ 2 \leq x_A \leq x_A^{max}}} \mathbf{P}\left( \begin{array}{c|c} x_A & \\ N-x_A & N; \end{array} \begin{array}{c} \phi_A \\ 1-\phi_A \end{array} \right) \sigma^*_{N-x_A^{max}} \\
\leq \quad &\mathbf{P}\left( \begin{array}{c|c} 0 & \\ N & N; \end{array} \begin{array}{c} \phi_A \\ 1-\phi_A \end{array} \right) (\sigma^*_N - \sigma^*_{N-x_A^{max}}) + \\
&\qquad \left[ \sum_{x_A \text{ even}} \mathbf{P}\left( \begin{array}{c|c} x_A & \\ N-x_A & N; \end{array} \begin{array}{c} \phi_A \\ 1-\phi_A \end{array} \right) \right] \sigma^*_{N-x_A^{max}}.
\end{aligned}$$

In this last line, the first probability is $(1-\phi_A)^N$, and the bracketed sum is the probability that a binomial distribution with parameters $N; \phi_A$ produces an even number of successes, which is $(1+(1-2\phi_A)^N)/2$ (Lemma C.3). Thus, the probability of drawing a profile in $S_0$ is at most

$$h(\phi_A) = (1-\phi_A)^N(\sigma^*_N - \sigma^*_{N-x_A^{max}}) + \frac{1+(1-2\phi_A)^N}{2}\sigma^*_{N-x_A^{max}}.$$

Let us find the maximum of $h$ on $[0, 1/2]$ (since by assumption $\phi_A$ lies in this interval). Differentiating gives

$$\frac{dh}{d\phi_A} = -N\left[(1-\phi_A)^{N-1}(\sigma^*_N - \sigma^*_{N-x_A^{max}}) + (1-2\phi_A)^{N-1}\sigma^*_{N-x_A^{max}}\right].$$

This is negative if

$$\left(\frac{1-2\phi_A}{1-\phi_A}\right)^{N-1} > \frac{\sigma^*_{N-x_A^{max}} - \sigma^*_N}{\sigma^*_{N-x_A^{max}}},$$

which holds precisely when $\phi_A$ is sufficiently small. Therefore $h$ is initially decreasing and then increasing, so the maximum occurs at one of the endpoints of the interval,

$$h(0) = \sigma^*_N \qquad \text{or} \qquad h\left(\frac{1}{2}\right) = \frac{1}{2^N}\sigma^*_N + \left(\frac{1}{2} - \frac{1}{2^N}\right)\sigma^*_{N-x_A^{max}}.$$

The first of these is larger as long as $\sigma^*_N \geq \sigma^*_{N-x_A^{max}}/2$. Using the fact that $N - x_A^{max} \geq 2N/3$ and the bounds in Lemma C.5, we can verify that this always holds. Thus, we have shown that the probability of drawing a profile in $S_0$ is

$$\mathbf{P}(S_0 \mid N; \phi) \leq h(\phi_A) \leq h(0) = \sigma^*_N.$$

That takes care of (i).

Next we turn to (ii), where we consider the probability of drawing a profile in $S_1$. In this case, each such profile has $x_A$ odd and at most $x_A^{max} = 2\lfloor N/6 \rfloor + 1$. Hence, by similar calculations, the relevant probability is

$$\sum_{\substack{x_A \text{ odd} \\ 1 \leq x_A \leq x_A^{max}}} \mathbf{P}\left( \begin{array}{c|cc} x_A & N; & \phi_A \\ N - x_A & & 1 - \phi_A \end{array} \right) \mathbf{P}\left( \begin{array}{c|cc} x_B & N - x_A; & \phi'_B \\ x_C & & \phi'_C \end{array} \right)$$

$$\leq \left[ \sum_{x_A \text{ odd}} \mathbf{P}\left( \begin{array}{c|cc} x_A & N; & \phi_A \\ N - x_A & & 1 - \phi_A \end{array} \right) \right] \sigma^*_{N - x_A^{max}}.$$

The bracketed expression is the probability that a binomial distribution with parameters $N; \phi_A$ produces an odd number of successes, which is $(1 - (1 - 2\phi_A)^N)/2 \leq 1/2$. (Remember that $\phi_A \leq 1/2$.) Therefore the probability of drawing a profile in $S_1$ is at most $\sigma^*_{N - x_A^{max}}/2$. This is less than $\sigma^*_N$, again by straightforward use of the bounds from Lemma C.5.

In case (iii), the relevant set of profiles again has $x_A$ odd and at most $x_A^{max} = 2\lfloor N/6 \rfloor + 1$, so the reasoning used for (ii) applies again word for word.

Finally, in (iv), the relevant set of profiles has $x_A$ even and at most $x_A^{max} = 2\lfloor N/6 \rfloor$. In this case the reasoning used for (i) applies again.

This covers all four cases (i)-(iv), so the probability that the manipulator is pivotal is never more than $\sigma^*_N$.

□

**Proof of Proposition 3.5:** Again, the tie-breaking assumption leads us to split into cases depending on parity. First suppose $M$ is even. Let the manipulator's

preferences be $A_1A_2\ldots A_M$. Suppose the belief $\phi$ is

$$\begin{array}{l} \frac{1}{2}\ A_1A_2A_3\ldots A_M \\ \frac{1}{2}\ A_2A_1A_3\ldots A_M \end{array}.$$

That is, all the other voters prefer $A_1$ and $A_2$, then the remaining candidates in numerical order, but are evenly split between ranking $A_1$ first or $A_2$ first. The manipulator considers manipulating by moving $A_2$ to the bottom, thus reporting $A_1A_3\ldots A_MA_2$.

Regardless of whether the manipulator tells the truth or lies, $A_1$ will have a higher score than $A_3,\ldots,A_M$, so the winner must be $A_1$ or $A_2$. Suppose $x$ of the other voters rank $A_1$ first, and the remaining $N-x$ rank $A_2$ first. The difference in scores between $A_1$ and $A_2$ is $(x+1)-(N-x)$ if the manipulator tells the truth and $(x+M-1)-(N-x)$ if he lies. Therefore, manipulation improves the outcome from $A_2$ to $A_1$ if

$$2x-N+1<0\leq 2x-N+M-1$$

or equivalently

$$\frac{N-(M-1)}{2}\leq x<\frac{N-1}{2}.$$

Otherwise, manipulation has no effect on the outcome.

Given that $x$ has to be an integer, the possible values of $x$ in this range are $\lfloor N/2-K\rfloor$ for $K=1,2,\ldots,(M-2)/2$. For each such $K$, Lemma 2.1 tells us that the probability that $x=\lfloor N/2-K\rfloor$ is $\sim\sqrt{2/\pi N}$. Therefore, the total probability of being pivotal is $\sim\frac{M-2}{2}\sqrt{2/\pi N}$, and the result follows via (2.3).

Now suppose $M$ is odd. The argument is essentially the same, except that we have to consider different cases depending on the parity of $N$. If $N$ is even, then we consider exactly the same preferences, the same manipulation, and the same belief as before. Again, the manipulator is pivotal if $(N-(M-1))/2\leq x<(N-1)/2$. The integer values of $x$ in this range are $N/2-K$ for $K=1,2,\ldots,(M-1)/2$.

If $N$ is odd, then we reverse the roles of $A_1$ and $A_2$ throughout. Thus, the manipulator's belief is the same as before, but his true preference is $A_2A_1A_3\ldots A_M$, and the proposed manipulation is $A_2A_3\ldots A_MA_1$. Let $x$ now denote the number

of other voters who rank $A_2$ first. Then the score of $A_2$ minus the score of $A_1$ is $(x+1)-(N-x)$ if the manipulator tells the truth and $(x+M-1)-(N-x)$ if he lies; in view of alphabetical tie-breaking, the manipulator is pivotal if

$$2x - N + 1 \leq 0 < 2x - N + M - 1.$$

The integer values of $x$ satisfying these inequalities are $x = (N+1)/2 - K$ for $K = 1, 2, \ldots, (M-1)/2$.

So for both $N$ even and $N$ odd, the manipulator is pivotal when $x = \lceil N/2 \rceil - K$ for some $K = 1, 2, \ldots, (M-1)/2$. The total probability of this event is $\sim \frac{M-1}{2}\sqrt{2/\pi N}$.

□

We next prove Lemma 4.8, the ancillary result en route to the local average lemma. We use the notation $\overline{f}(\phi)$, $\overline{f}_{A_i}(\phi)$ developed in Subsection 4.3 of the main paper.

**Proof of Lemma 4.8:** Put $g(x) = \overline{f}_{A_i}(\phi^x)$.

The proof is based on the following observation. Consider the definition (2.4) of $\mathbf{P}$, and take the partial derivative with respect to a parameter $\alpha_i$ (ignoring the fact that our interpretation of (2.4) required $\alpha_1 + \cdots + \alpha_r = 1$). We obtain

$$\frac{\partial}{\partial \alpha_i}\mathbf{P}\left(\begin{array}{c|cc} x_1 & & \alpha_1 \\ \vdots & & \vdots \\ x_i & K; & \alpha_i \\ \vdots & & \vdots \\ x_r & & \alpha_r \end{array}\right) = K \cdot \mathbf{P}\left(\begin{array}{c|cc} x_1 & & \alpha_1 \\ \vdots & & \vdots \\ x_i - 1 & K-1; & \alpha_i \\ \vdots & & \vdots \\ x_r & & \alpha_r \end{array}\right). \qquad \text{(D.5)}$$

On the right-hand side, $x_i$ has been replaced by $x_i - 1$ and all other $x_j$ are unchanged.

Now consider the function of $x$,

$$\overline{f}_{A_i}(\phi^x) = \sum_{f(P)=A_i} \mathbf{P}\left(\begin{array}{c|cc} P & & \alpha(1-x) \quad \succ \\ & N+1; & \alpha x \quad \succ' \\ & & 1-\alpha \quad \phi \end{array}\right).$$

The sum is over all $(N+1)$-profiles $P$ such that $f(P) = A_i$. Differentiating this sum

term-by-term with respect to $x$, and applying (D.5), we obtain

$$\frac{d}{dx}\left(\overline{f}_{A_i}(\phi^x)\right) = \sum_{f(P)=A_i} \left(\alpha\widetilde{N} \cdot \mathbf{P}(P- \succ' \mid N;\phi^x) - \alpha\widetilde{N} \cdot \mathbf{P}(P- \succ \mid N;\phi^x)\right). \quad \text{(D.6)}$$

The interpretation of the $\mathbf{P}(P- \succ' \mid \cdots)$ term is that if $P$ contains at least one $\succ'$ vote, then $P- \succ'$ is the $N$-profile consisting of $P$ with a $\succ'$ removed, and otherwise we simply interpret the whole term to be zero; similarly for the $\mathbf{P}(P- \succ \mid \cdots)$ term.

Now (D.6) can be rewritten

$$\frac{d}{dx}\left(\overline{f}_{A_i}(\phi^x)\right) = \alpha\widetilde{N}\left[\sum_{f(\succ',P)=A_i} \mathbf{P}(P \mid N;\phi^x) - \sum_{f(\succ,P)=A_i} \mathbf{P}(P \mid N;\phi^x)\right].$$

Here the first sum is over $N$-profiles $P$ with $f(\succ', P) = A_i$, and the second is over $P$ with $f(\succ, P) = A_i$. This in turn is equivalent to the difference given in the lemma statement. □

We also include here the proof of the result in Appendix A. It is basically a routine unwinding of definitions.

**Proof of Proposition A.1:** It suffices to show that for any symmetric equilibrium strategies of the voters, the following holds:

(a) if the planner chooses a voting rule $f$ with $\overline{\sigma}(f) \leq \underline{\epsilon}$ then her utility is given by $\min_{P \in \mathcal{L}^{N+1}} V(f(P), P)$;

(b) if she chooses $f$ with $\overline{\sigma}(f) > \underline{\epsilon}$, then her utility is $\underline{V}$.

Statement (a) holds because the voters will never manipulate. Specifically, suppose the state is $\omega \in \Omega^*$. Then, $\sigma_\omega(f) < \overline{\sigma}(f) \leq \underline{\epsilon}$. Consider a voter with utility function $u$, manipulation cost $\epsilon$, and belief $\psi$ about the types of the other voters. Composing the strategy $\tau$ of the other voters with $\psi$ gives a probability distribution $\phi \in \Delta(\mathcal{L})$, so that other voters' actual reports are expected to be independent draws from $\phi$. Consider any manipulation $\succ' \in \mathcal{L}$. From the definition of $\sigma_\omega(f)$ we have

$$u(\omega(f, \succ', \phi)) - u(\omega(f, \succ^*(u), \phi)) \leq \sigma_\omega(f) < \underline{\epsilon} \leq \epsilon.$$

Equivalently,

$$u(\omega(f, \succ', \phi)) - \epsilon < u(\omega(f, \succ^* (u), \phi)).$$

So the voter will choose to simply report the true preference $\succ^* (u)$. Thus, in all possible states $\omega \in \Omega^*$, each equilibrium strategy $\tau$ of the voters will specify that they always tell the truth. Then, whenever the voters' true preferences realize the (ordinal) profile $P$, the planner's utility is $V(f(P), P)$, regardless of the state. From the maxmin specification of the planner's utility, claim (a) follows.

For (b), consider any $f$ with $\overline{\sigma}(f) > \underline{\epsilon}$. We know that there exist some preferences $\succ_1, \ldots, \succ_{N+1}$ and reports $\widehat{\succ}_1, \ldots, \widehat{\succ}_{N+1}$ such that

$$V(f(\widehat{\succ}_1, \ldots, \widehat{\succ}_{N+1}); \succ_1, \ldots, \succ_{N+1}) = \underline{V}.$$

(This follows from the definition of $\underline{V}$ as the minimum value of $V$, and the fact that $f$ is surjective.) So our strategy will be to construct some state $\omega \in \Omega^*$, and some types $t_i \in \mathcal{T}$ for the voters, such that each voter $i$ has true preference $\succ_i$ but reports $\widehat{\succ}_i$ in any equilibrium.

First we construct the state $\omega$, as follows. Fix a number $\widetilde{\sigma}$ with $\underline{\epsilon} < \widetilde{\sigma} < \min\{1, \overline{\sigma}(f)\}$. We first define $\xi : \mathcal{L} \times \Delta(\mathcal{L}) \to \Delta(\mathcal{C})$ to be any continuous function such that for all preferences $\succ, \succ', \succ'' \in \mathcal{L}$,

$$\xi\left(\succ, \frac{2}{3} \succ' + \frac{1}{3} \succ''\right) = \begin{cases} \text{the top candidate in } \succ', \text{ with certainty} & \text{if } \succ = \succ''; \\ \text{the bottom candidate in } \succ', \text{ with certainty} & \text{otherwise.} \end{cases}$$

This can be done, since we have only specified the values of $\xi$ at finitely many points.

Now, for the given voting rule $f$, we define $\omega(f, \succ, \phi) \in \Delta(\mathcal{C})$ for all preferences $\succ \in \mathcal{L}$ and all beliefs $\phi \in \Delta(\mathcal{L})$, by

$$\omega(f, \succ, \phi) = \widetilde{\sigma} \xi(\succ, \phi) + (1 - \widetilde{\sigma})\, A_1.$$

That is, if the voting rule is $f$, then $\omega$ chooses the output of $\xi$ with probability $\widetilde{\sigma}$, and otherwise just chooses the fixed candidate $A_1$ as winner.

For every other voting rule $f' \neq f$, any $\succ \in \mathcal{L}$ and any $\phi \in \Delta(\mathcal{L})$, put

$$\omega(f', \succ, \phi) = \omega_0(f', \succ, \phi).$$

This completes the definition of $\omega$. It is straightforward to check that $\omega$ is indeed a continuous function: we need $\omega(f, \succ, \phi)$ to be continuous in $\phi$, but this follows from continuity of $\xi$; and for each $f' \neq f$ we need $\omega(f', \succ, \phi)$ to be continuous in $\phi$, but this follows from continuity for $\omega_0$.

We check that $\omega \in \Omega^*$. Notice that under voting rule $f$ in state $\omega$, each voter cannot affect more than $\widetilde{\sigma}$ probability mass of the outcome by changing his vote. It immediately follows that

$$\sigma_\omega(f) \leq \widetilde{\sigma} < \overline{\sigma}(f).$$

And for any other voting rule $f'$, we have

$$\sigma_\omega(f') = \sigma_{\omega_0}(f') < \overline{\sigma}(f')$$

by the assumption $\omega_0 \in \Omega^*$. Thus, the susceptibility bounds are satisfied, and $\omega \in \Omega^*$.

Next, for each voter $i$, we construct a type $t_i$ as follows:

- the utility function $u_i$ represents the preference $\succ_i$, and values the top candidate at 1 and the bottom candidate at 0;
- the manipulation cost is $\underline{\epsilon}$;
- the first-order belief about others' preferences is that every other voter
  - with probability 2/3, has a utility function that represents $\succ_i$ and has range smaller than $\underline{\epsilon}$; and
  - with remaining probability 1/3, has a utility function that represents $\widehat{\succ}_i$ and has range smaller than $\underline{\epsilon}$.

(The first-order belief about others' manipulation costs may be arbitrary.)

By the richness assumption, there exists a type $t_i \in \mathcal{T}$ having this basic type and first-order belief.

Now we consider $t_i$'s equilibrium behavior in state $\omega$. First, in any equilibrium, any voter whose utility function has range smaller than $\underline{\epsilon}$ always votes truthfully (since his material gain from lying is less than $\underline{\epsilon}$). Therefore, voter $i$'s induced belief $\phi$ about others' behavior is that each other voter will report $\succ_i$ with probability 2/3 and report $\widehat{\succ}_i$ with probability 1/3. Then:

- $\omega(f, \widehat{\succ}_i, \phi)$ is the distribution that chooses the candidate ranked first by $\succ_i$ with probability $\widetilde{\sigma}$, and chooses $A_1$ with remaining probability $1 - \widetilde{\sigma}$. Thereforefore, if voter $i$ reports $\widehat{\succ}_i$, his expected material utility is $\widetilde{\sigma} + (1 - \widetilde{\sigma})u_i(A_1)$.

- For any $\succ' \neq \widehat{\succ}_i$, $\omega(f, \succ', \phi)$ chooses the candidate ranked last by $\succ_i$ with probability $\widetilde{\sigma}$, and $A_1$ with remaining probability $1 - \widetilde{\sigma}$. Therefore, $i$'s expected material utility from reporting any such $\succ'$ is $(1 - \widetilde{\sigma})u_i(A_1)$.

Since $\widetilde{\sigma} > \underline{\epsilon}$, voter $i$'s unique best reply is to report $\widehat{\succ}_i$.

Thus, in state $\omega \in \Omega^*$, the types $t_1, \ldots, t_{N+1}$ of the voters have true preferences $\succ_1, \ldots, \succ_{N+1}$ but necessarily report $\widehat{\succ}_1, \ldots, \widehat{\succ}_{N+1}$. This leaves the planner with utility

$$V(f(\widehat{\succ}_1, \ldots, \widehat{\succ}_{N+1}); \succ_1, \ldots, \succ_{N+1}) = \underline{V},$$

her worst possible. Statement (b) follows. □

# E Proofs for comparison of voting systems

Here we prove Proposition 3.6, giving lower bounds on the susceptibility of five voting systems from [2].

**Proof of Proposition 3.6:** We give the proofs for the voting systems one by one in order.

**Black's system.** This is just an embellishment of the construction given for the Borda count, performed so as to ensure the nonexistence of a Condorcet winner (with

probability close to 1). We first present the construction for $M = 5$. For readability we refer to the candidates using letters $A, B, C, D, E$. Take small $\epsilon > 0$. Consider the following belief of the manipulator: the other voters report

$$\left.\begin{array}{lll} CDABE & DEABC & ECABD \\ CDBAE & DEBAC & ECBAD \end{array}\right\} \quad \text{each with probability} \quad 1/12 + \epsilon;$$

$$\left.\begin{array}{lll} CABDE & ABDEC & ABECD \\ CBADE & BADEC & BAECD \end{array}\right\} \quad \text{each with probability} \quad 1/12 - \epsilon.$$

Each other voter then:

- prefers $C$ over $D$ with probability $2/3$;
- prefers $D$ over $E$ with probability $2/3$;
- prefers $E$ over $C$ with probability $2/3$;
- prefers $C$ over $A$ and $B$ with probability $1/2 + 2\epsilon$.

By Lemma 2.2(a), with probability converging exponentially to 1, each of these pairwise preferences will be held by a share at least $1/2 + \epsilon$ of opposing voters, so no matter what the manipulator does we will end up with $C \to A, B, D$; $D \to E$; and $E \to C$. In particular, no candidate can then be a Condorcet winner.

Also, each other voter awards, on average,

- $40/12 - 10\epsilon$ points each to $A$ and $B$;
- $36/12 + 4\epsilon$ points to $C$;
- $32/12 + 8\epsilon$ points each to $D$ and $E$.

Using Lemma 2.2(a) again, we see that with probability exponentially converging to 1, candidates $A$ and $B$ will end up with higher scores than $C, D$ or $E$, no matter what the manipulator does.

So, neglecting events of exponentially small probability, we can focus on the realizations where there is no Condorcet winner and only $A$ or $B$ can win whatever

the manipulator does. Then, if $N$ is even, let the manipulator's true preference be $ABCDE$ and consider the manipulation $ACDEB$; if $N$ is odd, let the true preference be $BACDE$ and the manipulation be $BCDEA$. Exactly as in the proof of Proposition 3.5, the manipulation improves the outcome from the manipulator's second-ranked to his first-ranked candidate with probability $\sim 2/\sqrt{2/\pi N}$, and has no effect otherwise.

This covers the case $M = 5$. For $M > 5$, construct a belief by supposing each other voter ranks the first five candidates $A_1, \ldots, A_5$ $(= A, \ldots, E)$ at the top according to the distribution above, and then has all remaining candidates in numerical order after them. Then none of the extra candidates can ever be a Condorcet winner, nor a Borda winner, since they receive lower scores than (say) $A_1$. So again, with probability converging exponentially to 1, the winner will be either $A_1$ or $A_2$ no matter what the manipulator does. Let the manipulator's preferences and proposed manipulation be as for Proposition 3.5; then manipulation succeeds with probability $\sim \lceil((M-2)/2)\rceil\sqrt{2/\pi N}$ by the same argument as before.

**Copeland's system.** We will give a construction supposing that $M = 3K - 1$, where $K \geq 3$. If $M \geq 9$ is instead of the form $3K$ or $3K+1$, then we can modify the construction by the usual method of appending the extra one or two candidates at the end of everyone's preferences, and the same argument will apply. At the end of the proof we will also show how to modify the construction for the remaining cases $M = 3, 4, 6, 7$.

It will be convenient to depart from our usual notation for candidates and instead let the candidates be called $A, B, C_1, \ldots, C_K, D_1, \ldots, D_{2K-3}$, where ties are broken in that order. We will also let the $D$-candidates be numbered cyclically, so that $D_{i+(2K-3)} = D_i$.

Let the manipulator's true preference be $C_1 \ldots C_K D_1 \ldots D_{2K-3} AB$. To describe the belief $\phi$, we will not list out all the preferences that other voters may have, as there are too many to list individually; instead, we will describe a process by which a random preference is constructed. In this description, we will refer to choosing a *random cyclic permutation* of the $D_i$, which means an ordering of the form $D_j D_{j+1} \ldots D_{j+2K-4}$, where each possible value of $j \in \{1, 2, \ldots, 2K-3\}$ is chosen with probability $1/(2K-$

3).

- With probability 1/3, do the following: Begin with $BA$, then, for each $i = 1, \ldots, K$, append $C_i$ either at the beginning or at the end, independently each with probability 1/2. Finally, attach a random cyclic permutation of the $D_i$ at the beginning of the preference order.

- With probability 2/3, do the following: Begin with $BA$, immediately followed by a random cyclic permutation of the $D_i$; then successively append each $C_i$ either at the beginning or at the end, each with probability 1/2.

Whenever one candidate is preferred to another candidate with probability strictly greater than 1/2 under this distribution, the usual application of Lemma 2.2(a) ensures that the former candidate majority-defeats the latter with probability $\overset{e}{\sim} 1$. Thus, we can see that with probability $\overset{e}{\sim} 1$, all of the following majority-defeat relations hold:

- $B \rightarrow A$;

- $D_i \rightarrow D_{i+1}, D_{i+2}, \ldots, D_{i+K-2}$, for each $i$;

- $B, A \rightarrow D_i$ for each $i$;

- $D_i \rightarrow C_j$, for all $i$ and $j$.

We henceforth assume that these relations hold. Moreover, for each $C_j$, each of the other voters either prefers both $A$ and $B$ over $C_j$ or prefers $C_j$ over both $A$ and $B$; each case occurs with probability 1/2.

Each candidate $D_i$ majority-defeats exactly half of the other $D$-candidates and all of the $C$-candidates, for a Copeland score of $2K - 2$. Each of the $C$-candidates is majority-defeated by all of the $D$-candidates and so has a score of no more than $K + 1 \leq 2K - 2$. On the other hand, $B$ defeats all of the $D$-candidates and $A$, and so has a score of at least $2K - 2$. So by alphabetical tie-breaking, no matter what the manipulator does, either $A$ or $B$ must win.

Call a candidate $C_j$ *defeated* if there are at least $\lfloor N/2 \rfloor + 1$ other voters ranking $A, B$ above $C_j$. Let $d$ be the number of defeated candidates. If the manipulator tells the truth, then $A$ majority-defeats all the $D_i$ and the defeated $C_j$, for a score of $2K - 3 + d$; $B$ majority-defeats all the $D_i$, the defeated $C_j$, and $A$, for a score of $2K - 2 + d$. So $B$ wins.

Now suppose the manipulator reports the ranking $AC_1 \ldots C_K D_1 \ldots D_{2K-3} B$. Say that the manipulator is *pivotal for* $C_j$ if there are exactly $\lfloor N/2 \rfloor$ other voters ranking $A, B$ above $C_j$. If the manipulator is pivotal for $c$ candidates, then $B$ still has a score of $2K - 3 + d$, but $A$ now majority-defeats all the candidates for which the manipulator is pivotal and so has score $2K - 3 + d + c$. Thus, $A$ wins if $c \geq 1$.

The probability of being pivotal for any given $C_j$ is $\sim \sqrt{2/\pi N}$, and pivotality for $C_j$ is independent of pivotality for $C_k$ for $j \neq j$. Hence, the probability of being pivotal for at least one candidate $C_j$ is $\sim K\sqrt{2/\pi N}$. The lower bound for susceptibility follows.

We still need to give the construction for the cases $M = 3, 4, 6, 7$. For $M = 6$, let the candidates be $A, B, C_1, C_2, D, E$, and let the true preference be $C_1 C_2 DEAB$. The belief $\phi$ is given as follows:

- With probability 1/3, do the following: Begin with $BAE$; successively append $C_1$ and then $C_2$ either at the beginning or the end each with probability 1/2; finally, append $D$ at the beginning.

- With probability 1/3, do the following: Begin with $BADE$; then successively append $C_1$ and then $C_2$ either at the beginning or the end each with probability 1/2.

- With probability 1/3, do the following: Begin with $BAD$; successively append $C_1$ and then $C_2$ either at the beginning or at the end each with probability 1/2; finally, add $E$ at the beginning.

Now with probability $\overset{e}{\sim} 1$ we have the following majority-defeat relations:

- $A \to D, E$;

- $B \to A, D, E$;
- $D \to C_1, C_2, E$;
- $E \to C_1, C_2$.

Then $C_1, C_2$ both have score at most 3 since they are majority-defeated by $D$ and $E$. Under truth-telling, $A, B, D, E$ have respective scores $2+d, 3+d, 3, 2$, so that $B$ wins. Under the proposed manipulation, $A, B, D, E$ have scores $2+d+c, 3+d, 3, 2$, so that $A$ wins if the manipulator is pivotal for either $C_1$ or $C_2$. The same argument as before shows that this occurs with probability $\sim 2\sqrt{2/\pi N}$.

If $M = 3$, let the manipulator's true preference be $CBA$, and the belief $\phi$ be

$$1/4\ ACB, \qquad 1/2\ BAC, \qquad 1/4\ CBA.$$

With probability $\overset{e}{\sim} 1$, the resulting profile will have $B \to A \to C$. If exactly $\lfloor N/2 \rfloor$ of the other voters have $B \succ C$, then the manipulator is pivotal for this pair: Telling the truth leads to $C \to B$, in which case $A$ is the winner; manipulation leads to $B \to C$, so that $B$ wins, a more preferred outcome. If the manipulator is not pivotal, then the manipulation has no effect. So the manipulation is successful when the manipulator is pivotal, which happens with probability $\sim \sqrt{2/\pi N}$.

Finally, for $M = 4$ or 7, we take the construction for 3 or 6, respectively, and add an extra candidate at the end of everyone's preference ranking.

**Fishburn's system.** Assume $M \geq 4$, since the statement is trivial for $M = 3$. We return to the usual numerical labeling of the candidates. Let the manipulator's true preferences be $A_1 A_2 \ldots A_M$. As with the Copeland system, in order to describe the belief $\phi$, we give a process for generating a random preference, and let $\phi$ denote the resulting distribution over $\mathcal{L}$.

- With probability 2/3, we construct a preference as follows: Begin with $A_2 A_1 A_3$, and then for each $i = 4, \ldots, M$ in succession, randomly append $A_i$ either at the beginning of the existing ordering or at the end, independently with probability 1/2.

- With probability 1/3, we instead do the following: Begin with $A_2A_1$, then for each $i = 4, \ldots, M$ in succession, append $A_i$ either at the beginning or the end, independently with probability 1/2; finally, append $A_3$ at the beginning.

A preference $\succ$ drawn according to this distribution has the following properties:

- With probability 1, $A_2 \succ A_1$.
- With probability 2/3, $A_1, A_2 \succ A_3$.
- For each $i \geq 4$, with probability 2/3, $A_3 \succ A_i$.
- For each $i \geq 4$, with probability 1/2, $A_1, A_2 \succ A_i$; and with probability 1/2, $A_2, A_1 \succ A_i$.

Let the proposed manipulation consist of moving $A_2$ to the bottom of the ranking, thus reporting $A_1A_3 \ldots A_MA_2$. Let the utility $u$ be 1 for $A_1$ and 0 for every other candidate, so the manipulator is concerned only with the probability of $A_1$ winning.

As usual, with probability $\overset{e}{\sim} 1$, we have $A_2 \to A_1$, $A_1 \to A_3$, $A_2 \to A_3$, and $A_3 \to A_i$ for all $i \geq 4$. We may assume these relations hold.

If the manipulator tells the truth, then for each $i \geq 4$, either $A_1, A_2$ both majority-defeat $A_i$, or both are majority-defeated by $A_i$. In this case, $A_1$ is covered by $A_2$, and so cannot win.

Now suppose the manipulator lies. For each $i \geq 4$, if there are exactly $\lfloor N/2 \rfloor$ other voters who report $A_2 \succ A_i$, then the lie leads to $A_1 \to A_i \to A_2$ in the resulting profile. Say that the manipulator is *pivotal for* $A_i$ if this occurs. In this case, $A_2$ no longer covers $A_1$. Notice that $A_3$ also cannot cover $A_1$, nor can any $A_j$ for $j \geq 4$, since $A_1 \to A_3 \to A_j$. Hence, $A_1$ is uncovered and so wins.

So the manipulation is successful whenever the manipulator is pivotal for any $A_i$, $i \geq 4$. For each such $A_i$, the probability of being pivotal is $\sim \sqrt{2/\pi N}$. Moreover, for distinct $i, j \geq 4$, our construction of the opponents' preferences assigned $A_i$ to be ranked above or below $A_2$ independently of $A_j$; as a result, pivotality for $A_i$ is independent of pivotality for $A_j$. Consequently, the probability of being pivotal for both $A_i$ and $A_j$ simultaneously is $\sim 2/\pi N$. So as $N$ becomes large, the probability

of being pivotal for more than one $A_i$ becomes negligible compared to the probability of being pivotal for any given $A_i$, and so the probability of being pivotal for at least one $A_i$ is asymptotically $(M-3)$ times the probability of being pivotal for any given $A_i$; that is, $\sim (M-3)\sqrt{2/\pi N}$.

**Minimax system.** The argument here is similar to that used for Black's system above, and for single transferable vote below, but we vary the beliefs $\phi$ as $N$ varies. Doing so allows us to obtain a lower bound on susceptibility that converges more slowly than $N^{-1/2}$, but at the cost of requiring some additional computation.

For any given candidate $A_i$, we will use the term *defeater of* $A_i$ to refer to any candidate $A_j$ achieving the maximum, over $j \neq i$, of the number of voters preferring $A_j$ to $A_i$.

We first prove the bound for $M = 4$, with the candidates labeled $A, B, C, D$. Let the manipulator's preference be $ABCD$, and take the set of desirable candidates in (2.3) to be $\mathcal{C}^+ = \{A\}$. Consider the following belief $\phi$:

$$\begin{array}{ll} \frac{1}{\sqrt{N}} & ACBD \\ \frac{1}{2} - \frac{1}{\sqrt{N}} & ADBC \\ \frac{1}{2} - \frac{1}{\sqrt{N}} & CBAD \\ \frac{1}{\sqrt{N}} & DBAC \end{array}.$$

Let the proposed manipulation be $ACBD$.

In order to keep track of the consequences of manipulation, let the number of other voters reporting each of the four preferences be $w, x, y, z$, respectively. Then, the score of each candidate under truth-telling and under manipulation are as follows:

| | Truth | Manipulation |
|---|---|---|
| $A$ | $y+z$ | $y+z$ |
| $B$ | $\max\{w+x+1, w+y, x+z\}$ | $\max\{w+x+1, w+y+1, x+z\}$ |
| $C$ | $w+x+z+1$ | $w+x+z+1$ |
| $D$ | $\max\{w+x+y+1, w+y+z+1\}$ | $\max\{w+x+y+1, w+y+z+1\}$ |

(These values are obtained after performing obvious restrictions on the set of defeaters for each candidate. For example, every voter who ranks either $C$ or $D$

above $A$ also ranks $B$ above $A$, so $B$ is always a defeater for $A$, making the score of $A$ necessarily $y+z$. Likewise, $A$ is always a defeater for $C$, and either $A$ or $B$ is a defeater for $D$.)

We see that the only possible effect of manipulation is to increase the score of $B$ from $w+y$ to $w+y+1$. Hence, manipulation can change the outcome of the vote in only two situations:

- Manipulation can change the outcome from $B$ to $A$ if it causes $A$ and $B$ to have equal scores, and $C$'s score is at least as high. This requires $w+y \geq w+x+1, x+z$; $w = z-1$; and $y \leq w+x+1$.

  If $w = z-1$ then $w+x+1 = x+z$, so we actually need only $x+1 \leq y \leq w+x+1$ and $w = z-1$.

- Manipulation can also change the outcome from $B$ to $C$. However, since both $B$ and $C$ are undesirable outcomes, this case contributes nothing to the expectation in (2.3).

Thus, we are left to estimate the probability that $x+1 \leq y \leq w+x+1$ and $w = z-1$.

Write $s$ for the sum $w+z$, and $t$ for $N-s = x+y$. We use Lemma C.2 to decompose the probability of a profile $(w,x,y,z)$ into the probability of given values of $s,t$, times the probabilities of $w$ conditional on $s$ and of $y$ conditional on $t$. Thus, the probability we want becomes

$$\sum_{\substack{s \text{ odd} \\ t=N-s}} \left[ \mathbf{P}\left( \begin{matrix} s \\ t \end{matrix} \;\middle|\; N; \begin{matrix} 2/\sqrt{N} \\ 1-2/\sqrt{N} \end{matrix} \right) \times \mathbf{P}\left( \begin{matrix} (s-1)/2 \\ (s+1)/2 \end{matrix} \;\middle|\; s; \begin{matrix} 1/2 \\ 1/2 \end{matrix} \right) \right.$$

$$\left. \times \sum_{\frac{t+1}{2} \leq t \leq \frac{2t+s+1}{4}} \mathbf{P}\left( \begin{matrix} y \\ t-y \end{matrix} \;\middle|\; t; \begin{matrix} 1/2 \\ 1/2 \end{matrix} \right) \right]. \tag{E.1}$$

A lower bound for this outer sum is given by considering only the terms where $\sqrt{N} < s < 3\sqrt{N}$. In this case, by Lemma 2.1, $\min_s \mathbf{P}((s-1)/2, (s+1)/2 \mid s; 1/2, 1/2) \sim$

$\sqrt{2/3\pi\sqrt{N}}$. Also, the probability in the inner sum of (E.1) is decreasing as a function of $y$ for $y > t/2$, so each such term is at least

$$\begin{aligned}
&\mathbf{P}\left(\begin{array}{c|c} \lceil (2t+s+1)/4 \rceil & t;\ 1/2 \\ t - \lceil (2t+s+1)/4 \rceil & \ \ 1/2 \end{array}\right) \\
&\quad = \mathbf{P}\left(\begin{array}{c|c} \lceil t/2 \rceil & t;\ 1/2 \\ \lfloor t/2 \rfloor & \ \ 1/2 \end{array}\right) \times \prod_{k=1}^{\lceil (2t+s+1)/4 \rceil - \lceil t/2 \rceil} \frac{\lfloor t/2 \rfloor + 1 - k}{\lceil t/2 \rceil + k} \\
&\quad \geq \min_{\substack{\sqrt{N}<s<3\sqrt{N} \\ t=N-s}} \left[ \mathbf{P}\left(\begin{array}{c|c} \lceil t/2 \rceil & t;\ 1/2 \\ \lfloor t/2 \rfloor & \ \ 1/2 \end{array}\right) \times \left( \frac{t/2 - \lceil \frac{s+1}{4} \rceil}{t/2 + 1 + \lceil \frac{s+1}{4} \rceil} \right)^{\lceil \frac{s+1}{4} \rceil} \right] \\
&\quad \gtrsim \min_{N-3\sqrt{N}<t<N-\sqrt{N}} \left[ \mathbf{P}\left(\begin{array}{c|c} \lceil t/2 \rceil & t;\ 1/2 \\ \lfloor t/2 \rfloor & \ \ 1/2 \end{array}\right) \times \left( 1 - \frac{2\sqrt{N}}{N/3} \right)^{\lceil \sqrt{N} \rceil} \right] \\
&\quad \sim \sqrt{\frac{2}{\pi N}} \times e^{-6}.
\end{aligned}$$

Hence the expression in (E.1) is

$$\begin{aligned}
&\gtrsim \sum_{\substack{s \text{ odd} \\ \sqrt{N}<s<3\sqrt{N} \\ t=N-s}} \left[ \mathbf{P}\left(\begin{array}{c|c} s & N;\ 2/\sqrt{N} \\ t & \ \ 1-2/\sqrt{N} \end{array}\right) \times \sqrt{\frac{2}{3\pi\sqrt{N}}} \times \left\lfloor \frac{s+1}{4} \right\rfloor e^{-6} \sqrt{\frac{2}{\pi N}} \right] \\
&\qquad \gtrsim \left[ \sum_{\substack{s \text{ odd} \\ \sqrt{N}<s<3\sqrt{N}}} \mathbf{P}\left(\begin{array}{c|c} s & N;\ 2/\sqrt{N} \\ N-s & \ \ 1-2/\sqrt{N} \end{array}\right) \right] \times \frac{e^{-6}}{2\pi\sqrt{3}\sqrt[4]{N}}.
\end{aligned}$$

To evaluate the bracketed probability sum, notice that if $(s, N-s)$ follows a binomial distribution $\mathbf{M}(N; 2/\sqrt{N}, 1-2/\sqrt{N})$ then $s$ has mean $2\sqrt{N}$ and variance $2\sqrt{N}-4 < 2\sqrt{N}$; by Chebyshev's inequality, the probability that it differs from its mean by more than $\sqrt{N}$ is less than $2/\sqrt{N}$. Hence this probability sum is $\sim 1$. Consequently, the probability that the manipulator is pivotal, given by (E.1), is

$$\gtrsim \frac{e^{-6}}{2\pi\sqrt{3}} \cdot \frac{1}{\sqrt[4]{N}}.$$

Thus, the susceptibility is asymptotically bounded below by this quantity, as claimed.

Finally, if $M > 4$, simply modify the belief $\phi$ above, the manipulator's true preference, and the proposed manipulation by appending each of the extra candidates in order at the end of each ranking. None of the extra candidates can ever win (they all receive a score of $N$), so the rest of the preceding analysis applies unchanged.

**Single transferable vote system.** Fix small $\epsilon > 0$. Let the manipulator's true preferences be $A_1 A_2 \ldots A_M$, and let the belief $\phi$ be as follows:

$$\begin{array}{ll}
\frac{1}{2^{M-1}} + \epsilon & A_1 A_2 A_3 \ldots \\
\frac{1}{2^{M-1}} + \epsilon & A_2 A_3 \ldots \\
\frac{1}{2^{M-2}} - \frac{2}{M-2}\epsilon & A_3 A_2 \ldots \\
\frac{1}{2^{M-3}} - \frac{2}{M-2}\epsilon & A_4 A_2 A_3 \ldots \\
\frac{1}{2^{M-4}} - \frac{2}{M-2}\epsilon & A_5 A_2 A_3 \ldots \\
 & \vdots \\
\frac{1}{2} - \frac{2}{M-2}\epsilon & A_M A_2 A_3 \ldots
\end{array}$$

The end of each preference ranking (denoted by ...) may be filled in arbitrarily. (In case the pattern is unclear, the first 3 preference types are $A_1 A_2 A_3 \ldots$, $A_2 A_3 \ldots$, and $A_3 A_2 \ldots$, and then the remaining preferences — if $M > 3$ — are all of the form $A_i A_2 A_3 \ldots$.)

By the usual application of Lemma 2.2(a), for $N$ large, we can focus on the realizations such that for each preference ordering $\succ$, the share of the population reporting $\succ$ is within $\epsilon/M^2$ of the weight put on $\succ$ by distribution $\phi$.

In this case, the single transferable vote procedure follows one of two possible executions. Either $A_1$ or $A_2$ is eliminated in the first round.

- Suppose $A_1$ is eliminated first. Then the candidates $A_1, A_3, A_4, A_5, \ldots, A_M$ are eliminated in succession. Indeed, we can show by induction that at the beginning of the $k$th round of elimination ($k > 1$) that candidates $A_2$ and $A_{k+1}, \ldots, A_M$ remain. If this holds for some $k$, then $A_2$ receives the votes of the first $k$ preference types of voters, thus getting a vote share of at least

$1/2^{M-k} + 2\epsilon - (k-2)2\epsilon/(M-2) - k\epsilon/M^2 > 1/2^{M-k}$. $A_{k+1}$ has a vote share at most $1/2^{M-k} - 2\epsilon/(M-2) + \epsilon/M^2 < 1/2^{M-k}$, and each of the other remaining candidates has vote share at least $1/2^{M-k-1} - 2\epsilon/(M-2) - \epsilon/M^2$. Thus, $A_{k+1}$ is eliminated next, and the voters who ranked $A_{k+1}$ first have their votes transferred to $A_2$, giving the induction step.

Thus, in this case, $A_2$ ends up winning.

- Suppose $A_2$ is eliminated first. Then the voters who ranked $A_2$ first have their votes transferrred to $A_3$. In the second round, $A_1$ is eliminated, and the voters who ranked $A_1$ first have their votes transferred to $A_3$. An induction identical to the previous case now shows that $A_4, A_5, \ldots, A_M$ are eliminated in successive rounds. Thus $A_3$ ends up winning.

Consider a proposed manipulation of the form $A_2A_3\ldots$. The manipulator is pivotal either when $A_1$ and $A_2$ receive the same number of first-place votes among the other voters, or when $A_2$ receives one more first-place vote than $A_1$. In both cases, if the manipulator tells the truth then $A_2$ is eliminated in the first round, hence $A_3$ ends up winning; under the proposed manipulation, $A_1$ is eliminated in the first round, and $A_2$ ends up winning. Hence, the manipulation is indeed beneficial.

By Lemma 2.2(b), we know each of the two pivotal scenarios happens with probability $\sim (1/2)\sqrt{1/\pi(1/2^{M-1}+\epsilon)N}$. Therefore, the total probability that the manipulator is pivotal is twice this quantity. We have thus shown

$$\sigma_N^{STV} \gtrsim \sqrt{\frac{1}{\pi\left(\frac{1}{2^{M-1}}+\epsilon\right)N}}.$$

Taking $\epsilon \longrightarrow 0$ gives the result. □

# F Analysis of the pair-or-plurality voting system

**Proof of Lemma 3.7:** For readability, we will refer to the candidates as $A, B, C$, with the understanding that this does not necessarily represent the tie-breaking order

Pareto efficiency is immediate: if all voters rank $A$ above $B$, then $B$ is not viable and so cannot win. Hence we focus on monotonicity.

Consider a profile $P$ at which some voter reports preference $ABC$. We need only consider what happens when this voter changes her preference by transposing the winner $f(P)$ with the candidate ranked immediately above him.

If $f(P) = C$, and the voter changes his preference to $ACB$, this cannot change the set of viable candidates, nor can it cause $C$ to lose in a majority vote against another candidate given that $C$ previously won this pairwise contest. This leaves only the case in which all three candidates are viable; in this case, the change can only increase $C$'s score and decrease $B$'s (while leaving $A$'s unchanged), so that $C$ remains the winner.

It remains to consider the case in which $f(P) = B$, and the voter changes his preference to $BAC$. Let $P'$ be the resulting profile. The change cannot affect the set of viable candidates except by making $A$ inviable. If this happens, $A$ gets exactly $K$ first-place votes at $P$. Suppose that $f(P') = C$ (otherwise $f(P') = B$ and we are done). Then, $A, B, C$ are all viable at $P$, while only $B$ and $C$ are viable at $P'$.

We claim that $B$ and $C$ both have at least $L$ first-place votes at $P$. If $B$ has less than $L$ first-place votes, then $C$ has at least $N + 1 - K - L > (N + 1)/2$ first-place votes and so gets at least half the total points, giving $f(P) = C$, a contradiction. If $C$ has less than $L$ first-place votes, then $B$ likewise has more than $(N + 1)/2$ first-place votes and so $f(P') = B$.

Hence, at $P$, all the points from voters ranking $B$ first go to $B$, and all the points from voters ranking $C$ first go to $C$. Since there are $K$ voters ranking $A$ first, their points go to $B$ and $C$ in the same quantities as rank $B$ or $C$ second, respectively. So the outcome is effectively determined by a pairwise vote between $B$ and $C$ — exactly the same as at $P'$. Thus $f(P) = f(P')$, a contradiction.

Thus we can assume that the same set of candidates is viable at $P$ as at $P'$. Since the change from $ABC$ to $BAC$ can only improve $B$'s standing in a pairwise majority vote, we only need to concern ourselves with the case where all three candidates are viable at both $P$ and $P'$.

Let us consider then the effect of changing $ABC$ to $BAC$ on each candidate's score. Let $s_A(A), s_A(B), s_A(C)$ be the number of points awarded to $A, B, C$, respectively, from the voters ranking $A$ first. Let us consider the effect of removing an $ABC$ vote on the vector $s_A = (s_A(A), s_A(B), s_A(C))$. If this leaves at least $L$ total voters with $A$ as their first-place vote, the net change in $s_A$ is $(-1, 0, 0)$. Otherwise, $s_A(A)$ is changed by $-L/(L-K)$ and $s_A(C)$ is changed by $\leq L/2(L-K)$, so $s_A(B)$ is changed by $\geq -1 - (-L/(L-K)) - (L/2(L-K)) = (2K-L)/2(L-K)$. In short, the net change in $s_A$ is of the form

$$\Delta s_A \quad = \quad (-1,0,0) \quad \text{or} \quad \left(-\frac{L}{L-K}, \geq \frac{2K-L}{2(L-K)}, \leq \frac{L}{2(L-K)}\right). \tag{F.1}$$

Now consider the effect of adding a $BAC$ vote on the corresponding vector $s_B = (s_B(A), s_B(B), s_B(C))$ of points from the voters ranking $B$ first. If there are initially at least $L$ such voters, the net change is $(0, 1, 0)$; otherwise, $s_B(B)$ changes by $L/(L-K)$, $s_B(A)$ changes by at most $(L-2K)/2(L-K)$, and $s_B(C)$ changes by at most 0. So the net change in $s_B$ is

$$\Delta s_B \quad = \quad (0,1,0) \quad \text{or} \quad \left(\leq \frac{L-2K}{L-K}, \frac{L}{L-K}, \leq 0\right). \tag{F.2}$$

Finally, when one voter's preference changes from $ABC$ to $BAC$, the net effect on the scores of the three candidates is given by the vector sum $\Delta s = \Delta s_A + \Delta s_B$. From (F.1), $\Delta s_A(B) \geq \Delta S_A(A)$, and from (F.2), $\Delta s_B(B) \geq \Delta S_B(A)$; thus $\Delta s(B) \geq \Delta s(A)$. From (F.1), $\Delta s_A(B) \geq \Delta s_A(C) - 1$, and from (F.2), $\Delta s_B(B) \geq \Delta s_B(C) + 1$; thus $\Delta s(B) \geq \Delta s(C)$. We conclude that the net change in $B$'s score from $P$ to $P'$ is at least as large as the net change in $A$'s score or $C$'s score. Since $B$ was the winner at the original profile $P$, then, $B$ again wins at $P'$. So we have $f(P')$ in this case as well, as required.

□

The proof of Proposition 3.8 will make use of the following two lemmas.

**Lemma F.1** *Let $x_N, y_N$ be a sequences of positive integers with $(y_N - x_N)/N > \epsilon$,*

*where $\epsilon > 0$ is some constant; $y_N < N - x_N$; and $x_N \to \infty$ as $N \to \infty$. Also let $b$ be a fixed positive integer. Then*

$$\max_{a,\alpha} \left[ b \cdot \sum_{\substack{x_N < x < y_N \\ x \equiv a \pmod b}} \mathbf{P} \left( \begin{matrix} x \\ N-x \end{matrix} \middle| N; \begin{matrix} \alpha \\ 1-\alpha \end{matrix} \right) \right] \sim 1. \tag{F.3}$$

(Here the maximum is taken for a fixed $N$, over all choices of integers $a$ and probabilities $\alpha$.)

**Proof:** The sum of $\mathbf{P}(x, N-x \mid N; \alpha, 1-\alpha)$ over *all* $x$ congruent to $a$ mod $b$, without the restriction $x_N < x < N - x_N$, is equal to

$$\frac{1}{b} \sum_{i=0}^{b-1} \zeta^{-ai} (\alpha \zeta^i + (1-\alpha))^N \tag{F.4}$$

where $\zeta$ is a primitive complex $b$-th root of 1. (This can be checked by expanding using the binomial theorem and cancelling terms where the powers of $\zeta$ sum to 0, as in the proof of Lemma C.3.) Fix any positive integers $K, L$. For $\alpha \leq K/N$, then $\sum_{x<L} \mathbf{P}(x, N-x \mid N; \alpha, 1-\alpha)$ is asymptotically bounded below, for $N \to \infty$, by the probability that a Poisson variable with parameter $K$ takes on value at most $L$. In particular, for any fixed $K$, choose $L$ large enough so that the latter Poisson probability is greater than $1 - 1/b$, then we certainly have $\max_{\alpha \leq K/N} \sum_{x > x_N} \mathbf{P}(x, N - x \mid N; \alpha, 1-\alpha) < 1/b$ for $N$ large enough, since eventually $x_N > L$. Hence, for any fixed $K$, we can restrict attention to $\alpha > K/N$, and likewise $\alpha < 1 - K/N$. In this case, the term in the sum (F.4) corresponding to $i = 0$ is equal to 1, and all other terms are bounded above in absolute value uniformly by $e^{-\lambda K}$ for some constant $\lambda > 0$ that depends only on $b$. So for each $K$, the left side of (F.4) is $\leq (1 + (b-1)e^{-\lambda K})/b$ for $N$ large enough and all $\alpha \in (K/N, 1 - K/N)$. Then, the maximum on the left side of (F.3) is also bounded above by $1 + (b-1)e^{-\lambda K}$ for all $N$ large enough. By choosing $K$ arbitrarily large, we see that this maximum is $\lesssim 1$.

To see that it is $\gtrsim 1$, simply take $\alpha = (x_N + y_N)/2N$ and apply (F.4) to obtain $\sum_{x \equiv a \pmod b} \mathbf{P}(x, N-x \mid N; \alpha, 1-\alpha) \to 1/b$, and note that the probability of realizing

a value $x < x_N$ or $x > y_N$ is $\stackrel{e}{\sim} 1$ by Lemma 2.2(a).

□

**Lemma F.2** *Let $S$ be a set of profiles $(x_1, \ldots, x_r)$ each of which satisfies $x_1, x_2 \geq K$, for some integer $K$. Then for any distribution $\alpha$,*

$$\sum_{(x_1,\ldots,x_r)\in S} \left[ \mathbf{P}\left( \begin{array}{c|c} x_1 & \\ x_2 & \\ x_3 & N;\ \alpha \\ \vdots & \\ x_r & \end{array} \right) - \mathbf{P}\left( \begin{array}{c|c} x_1+1 & \\ x_2-1 & \\ x_3 & N;\ \alpha \\ \vdots & \\ x_r & \end{array} \right) \right] \leq \frac{e^{1/12}}{\sqrt{\pi K}}.$$

**Proof:** We first prove the result for $r = 2$. In this case, taking $N$ and $\alpha$ as fixed, the expression $h(x) = \mathbf{P}(x, N-x \mid N; \alpha_1, \alpha_2)$ is unimodal as a function of $x$. Calling its maximum $x^*$, the specified difference is negative for $x < x^*$ and nonnegative for $x \geq x^*$, so the sum in the lemma statement is maximized when $S$ is the set of pairs whose $x_1$-values are $x^*, x^*+1, \ldots, x^*-K$. In this case, the sum of differences telescopes and the sum is simply $h(x^*) - h(x^*-K+1) \leq h(x^*)$. It follows from Lemmas 2.3 and C.7 that $h(x^*) \leq e^{1/12}/\sqrt{\pi K}$. This completes the proof in the case $r = 2$.

For the general case, let $S'$ be the set of all values of the $(r-1)$-tuple $x' = (x_1 + x_2, x_3, \ldots, x_r)$ for $(x_1, \ldots, x_r) \in S$. For each such $x' \in S'$, let $S_{x'}$ be the set of pairs $(x_1, x_2)$ such that $(x_1, x_2, \ldots, x_r) \in S$.

By Lemma C.2, we can rewrite the sum in the lemma statement as as

$$\sum_{(x_1,\ldots,x_r)\in S} \mathbf{P}\left( \begin{array}{c|c} x_1+x_2 & \\ x_3 & N; \\ \vdots & \\ x_r & \end{array} \begin{array}{c} \alpha_1+\alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_r \end{array} \right) \times$$
$$\left[ \mathbf{P}\left( \begin{array}{c|c} x_1 & \\ x_2 & x_{1+2};\ \end{array} \begin{array}{c} \beta_1 \\ \beta_2 \end{array} \right) - \mathbf{P}\left( \begin{array}{c|c} x_1+1 & \\ x_2-1 & x_{1+2};\ \end{array} \begin{array}{c} \beta_1 \\ \beta_2 \end{array} \right) \right]$$

(with $\beta_1 = \alpha_1/(\alpha_1+\alpha_2)$ and $\beta_2$ similarly)

$$= \sum_{x'=(x_{1+2},x_3,\ldots,x_r)\in S'} \mathbf{P}\left(\begin{array}{c|c} x' & N; \end{array}\begin{array}{c} \alpha_1+\alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_r \end{array}\right) \times$$
$$\left(\sum_{(x_1,x_2)\in S_{x'}} \left[\mathbf{P}\left(\begin{array}{cc|c} x_1 & & \beta_1 \\ x_2 & x_{1+2}; & \beta_2 \end{array}\right) - \mathbf{P}\left(\begin{array}{cc|c} x_1+1 & & \beta_1 \\ x_2-1 & x_{1+2}; & \beta_2 \end{array}\right)\right]\right).$$

By the $r = 2$ case, the expression in square brackets is at most $e^{1/12}/\sqrt{\pi K}$, so the whole sum is

$$\leq \frac{e^{1/12}}{\sqrt{\pi K}} \sum_{x'=(x_{1+2},x_3,\ldots,x_r)\in S'} \mathbf{P}\left(\begin{array}{c|c} x' & N; \end{array}\begin{array}{c} \alpha_1+\alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_r \end{array}\right).$$

Since the sum of probabilities is at most 1, we are done.

□

**Proof of Proposition 3.8:** We will prove the following claim: If $\lambda$ is an integer such that $L/K \geq \lambda$ for each (sufficiently large) $N$, and $K \to \infty$ as $N \to \infty$, then the pair-or-plurality voting rule satisfies

$$\sigma_N^{POP} \lesssim \left[\frac{1}{2} + \frac{1}{\lambda - 1}\right] \sqrt{\frac{12}{\pi N\left(4 - \frac{1}{4\lambda^2}\right)}}.$$

The desired bound will then follow by taking $\lambda \to \infty$.

Our proof will make frequent use of the following observation: At any profile where all three candidates are viable, if a candidate $A_i$ wins, then $A_i$ must have score at least $(N+1)/3$. For each $A_j \neq A_i$, the voters ranking $A_j$ first can contribute at most $K$ points to $A_i$, so there must be at least $(N+1)/3 - 2K > L$ voters ranking $A_i$ first. In particular, all the points from these voters are awarded to $A_i$; and even if

we change one of their votes, the other such voters still award all their points to $A_i$.

Henceforth, as in the proof of Proposition 3.3, we will notate the manipulator's true preference as $ABC$ for readability; this does not necessarily correspond to the tie-breaking order.

We first restrict the set of manipulations that need to be considered. With reference to (2.3), we have either $\mathcal{C}^+ = \{A\}$ or $\mathcal{C}^+ = \{A, B\}$. In the first case, the manipulator wishes to maximize the probability of $A$ winning; by monotonicity (Lemma 3.7) the optimal manipulation ranks $A$ first, so we need only consider the manipulation $ACB$. In the second case, the manipulator wishes to minimize the probability of $C$ winning; again by monotonicity, this is best done by ranking $C$ last, so we need only consider the manipulation $BAC$. So we need to show that in each of these two cases, the probability that the manipulation succeeds (under the worst-case belief $\phi$) is $\lesssim [1/2+1/(\lambda-1)]\sqrt{12/\pi N(4-1/4\lambda^2)}$. (We will henceforth denote this expression by $AUB(N)$, for "asymptotic upper bound.")

First we consider the case of a manipulation from $ABC$ to $ACB$. For any realization of the opponent-profile, this manipulation cannot change the set of viable candidates. If only one or two candidates are viable, the manipulation has no effect on whether $A$ wins or not, since it does not change the result of a pairwise vote between $A$ and either of the other candidates. Suppose all three candidates are viable. The manipulation cannot improve the outcome from $C$ to $A$, by monotonicity; so we need only consider whether it can improve the outcome from $B$ to $A$. Consider any opponent-profile $P$ at which $f(P, ACB) = A$. By our initial observation, the voters ranking $A$ first assign all their points to $A$, at both $(P, ABC)$ and $(P, ACB)$. So the manipulation from $ABC$ to $ACB$ actually can have no effect on any candidate's score and thus no effect on the outcome.

This leaves us to consider manipulations from $ABC$ to $BAC$. The manipulation cannot improve the outcome from $C$ to $A$, again by monotonicity, so we need only show that the probability that it improves the outcome from $C$ to $B$ is $\lesssim AUB(N)$. Let $S$ be the set of all opponent-profiles $P$ such that $f(ABC, P) = C$

and $f(BAC, P) = B$. We want to show that

$$\sum_{P \in S} \mathbf{P}(P \mid N; \phi) \lesssim AUB(N),$$

where $\phi$ is taken to be the belief that maximizes this sum.

If the manipulation does not change the set of viable candidates, and one or two candidates are viable, then as in the previous case, manipulation cannot improve the outcome from $C$ to $B$. If the manipulation does change the set of viable candidates, then it can only make $A$ inviable: $P$ must be such that all three candidates are viable at $(ABC, P)$ but only $B, C$ are viable at $(BAC, P)$. Then $P$ contains exactly $K - 1$ first-place votes for $A$. For $B$ to win at $(BAC, P)$, then $B$ must have at least $(N + 1) - (K - 1) > L + 1$ first-place votes. For $C$ to win at $(ABC, P)$, it must also have more than $L$ first-place votes, by our initial observation. Hence, at both $(ABC, P)$ and $(BAC, P)$, all the voters ranking $B$ or $C$ first award all their points to their first-choice candidate. But at $(ABC, P)$, since $A$ has exactly $K$ first-place votes, these voters each award one point to their second-choice candidate, so the outcome is given by a pairwise vote between $B$ and $C$ — exactly as at $(BAC, P)$. So the winner is the same at both profiles.

Consequently, for every $P \in S$, all three candidates are viable at $(ABC, P)$ and at $(BAC, P)$. Since $f(ABC, P) = C$ and $f(BAC, P) = B$, by our initial observation, $P$ contains at least $L$ first-place votes for $B$ and for $C$. Let $w, x, y, z$, respectively, denote the number of $ABC$ votes, the number of $ACB$ votes, the number of first-place votes for $B$, and the number of first-place votes for $C$ in profile $P$.

Let $S_A$ be the set of pairs $(w, x)$ such that there exist $y$ and $z$ with $(w, x, y, z) \in S$. For each such $(w, x)$ there exists at most one such pair $(y, z)$, by monotonicity; regard $y$ and $z$ then as functions of $(w, x)$. We can write the desired probability as

$$\sum_{(w,x,y,z) \in S} \mathbf{P}\left( \begin{array}{c|cc} w & & \alpha_w \\ x & N; & \alpha_x \\ y+z & & \alpha_y + \alpha_z \end{array} \right) \mathbf{P}\left( \begin{array}{c|cc} y & y+z; & \alpha_y/(\alpha_y + \alpha_z) \\ z & & \alpha_z/(\alpha_y + \alpha_z) \end{array} \right)$$

$$= \sum_{(w,x)\in S_A} \mathbf{P}\left(\begin{array}{c|cc} w & & \alpha_w \\ x & N; & \alpha_x \\ N-(w+x) & & 1-(\alpha_w+\alpha_x) \end{array}\right) \times \tag{F.5}$$
$$\mathbf{P}\left(\begin{array}{c|cc} y(w,x) & & \beta \\ z(w,x) & N-(w+x); & 1-\beta \end{array}\right)$$

with $\beta = \alpha_y/(\alpha_y + \alpha_z)$.

Moreover, since the absolute difference between points awarded to $B$ and points awarded to $C$ from the voters who rank $A$ first is always at most $K$, we must have $|y - z| \leq K + 1$. We can use this to obtain a uniform upper bound, across all $(w, x) \in S_A$, for the second factor in (F.5). The bound we will use is

$$\max_{\substack{\beta;y,z \\ y+z\geq(2N-1)/3 \\ |y-z|\leq K+1}} \mathbf{P}\left(\begin{array}{c|cc} y & & \beta \\ z & y+z; & 1-\beta \end{array}\right) \lesssim \sqrt{\frac{12}{\pi N(4-9\kappa^2)}} \tag{F.6}$$

where $\kappa$ is a number such that $K/N \leq \kappa$ for each $N$. (Here the asymptotics are with respect to $N$ as usual.) To see that (F.6) holds, fix $s \geq (2N-1)/3$ and consider maximizing with respect to $y, z, \beta$, subject to the conditions $y + z = s$ and $y, z \leq (2+3\kappa')s/4$, where $\kappa'$ is a number such that $(K+1)/(2N-1) \leq \kappa'/2$. Note that the condition $y, z \leq (2+3\kappa')s/4$ is implied by $y+z = s$ and $|y-z| \leq K+1 \leq 3\kappa' s/2$, so these new conditions are a weakening of the original conditions for the maximization in (F.6). The maximum is given by $\beta = y/s$ (by Lemma 2.3), and given this, the maximum over $y, z$ is achieved by making $y$ as large as possible (from Lemma 2.4), i.e. $y = \lfloor (2+3\kappa')s/4 \rfloor$. From Lemma (2.1) we get

$$\mathbf{P}\left(\begin{array}{c|cc} \lfloor (2+3\kappa')s/4 \rfloor & & \beta \\ s - \lfloor (2+3\kappa')s/4 \rfloor & s; & 1-\beta \end{array}\right) \sim \sqrt{\frac{8}{\pi s(4-9\kappa'^2)}}.$$

Plugging in $s \geq 2(N-1)/3$, and noticing that $\kappa'$ can be made arbitrarily close to $\kappa$ for $N$ large, gives us the upper bound in (F.6).

Now, $L/K \geq \lambda$ gives $(K+1)/N \lesssim 1/6\lambda$, so we can take $\kappa$ arbitrarily close

to $1/6\lambda$ in (F.6). To apply (F.6) to our problem we also need to know that the inequality $y + z \geq (2N - 1)/3$ holds for all $(w, x, y, z) \in S$. But if this is false, then $w + x > N/3 > L$, so all the voters ranking $A$ first award all their points to $A$, and the winner is simply determined by plurality rule at both the profiles $(ABC, P)$ and $(BAC, P)$. Since $f(ABC, P) = C$, $C$ has at least $(N + 1)/3$ first-place votes; since $f(BAC, P) = B$, $B$ has at least $(N - 2)/3$ votes in $P$, giving $y + z \geq (2N - 1)/3$ after all.

Combining with our previous observation $|y - z| \leq K + 1$ for $(w, x, y, z) \in S$, we see that the second probability factor in (F.5) is indeed a value of the maximand in (F.6), so the bound (F.6) applies with an appropriate value of $\kappa$. To determine such a value, notice that $L/K \geq \lambda$ gives $K/N \leq 1/6\lambda$, so we can take $\kappa$ arbitrarily close to $1/6\lambda$ in (F.6). Using this bound in (F.5) leads to

$$\sum_{P \in S} \mathbf{P}(P \mid N; \phi) \lesssim \tag{F.7}$$
$$\left[ \sum_{(w,x) \in S_A} \mathbf{P} \left( \begin{array}{c|cc} w & & \alpha_w \\ x & N; & \alpha_x \\ N - (w + x) & & 1 - \alpha_{w+x} \end{array} \right) \right] \sqrt{\frac{12}{\pi N \left(4 - \frac{1}{4\lambda^2}\right)}}.$$

This means we just need to focus henceforth on the first factor on the right side of (F.7). If we can show that

$$\sum_{(w,x) \in S_A} \mathbf{P} \left( \begin{array}{c|cc} w & & \alpha_w \\ x & N; & \alpha_x \\ N - (w + x) & & 1 - \alpha_{w+x} \end{array} \right) \lesssim \frac{1}{2} + \frac{1}{\lambda - 1}, \tag{F.8}$$

then the upper bound in the proposition will follow from (F.7), and the proof will be complete.

To further save on notation, we will henceforth write $\mathbf{P}(w, x)$ rather than write out $\mathbf{P}(w, x, N - (w + x) \mid N; \alpha_w, \alpha_x, 1 - \alpha_{w+x})$; the extra arguments will be implied.

From now on we will assume that ties between $B$ and $C$ are broken in favor of $B$. (The case where they are broken in favor of $C$ is essentially identical.)

Break the set of pairs of nonnegative integers $(w, x)$, with $w + x \geq K$, into four regions:

- $R_1$: $w + x < L$ and $w \geq x(1 - 2K/L) + K$.

  For $(w, x)$ in this region, the scores of $B, C$ associated with the true profile and the manipulated profile are

$$(ABC, P): \quad \frac{K(L - w - x - 1)}{L - K} + y \text{ for } B, \qquad z \text{ for } C$$

$$(BAC, P): \quad \frac{K(L - w - x)}{L - K} + y + 1 \text{ for } B, \qquad z \text{ for } C.$$

  Thus, for $(w, x)$ in this region, the manipulation increases $B$'s score by $L/(L-K)$ and leaves $C$'s score unaffected.

- $R_2$: $w + x < L$ and $x \geq (w+1)(1-2K/L)+K$. In this case, the scores of $B, C$ are

$$(ABC, P): \quad y \text{ for } B, \quad \frac{K(L - w - x - 1)}{L - K} + z \text{ for } C$$

$$(BAC, P): \quad y + 1 \text{ for } B, \quad \frac{K(L - w - x)}{L - K} + z \text{ for } C.$$

  Thus, for $(w, x)$ in this region, manipulation increases $B$'s score by 1 and $C$'s score by $K/(L - K)$.

- $R_3$: $w + x < L$, $w < x(1 - 2K/L) + K$ and $x < (w+1)(1 - 2K/L) + K$. In this case, the scores of $B, C$ are

$$(ABC, P): \; w+1-\frac{(w + x + 1 - K)L}{2(L - K)}+y \text{ for } B, \; x-\frac{(w + x + 1 - K)L}{2(L - K)}+z \text{ for } C$$

$$(BAC, P): \quad w-\frac{(w + x - K)}{2(L - K)}+y+1 \text{ for } B, \quad x-\frac{(w + x - K)L}{2(L - K)}+z \text{ for } C.$$

  Thus, for $(w, x)$ in this region, manipulation has no effect on the difference between $B$'s and $C$'s scores, so it cannot change the winner from $C$ to $B$: $S_A \cap R_3 = \emptyset$.

- $R_4$: $w + x \geq L$. In this case, the scores are

$$(ABC, P): \qquad y \text{ for } B, \qquad z \text{ for } C$$

$$(BAC, P): \qquad y + 1 \text{ for } B, \qquad z \text{ for } C.$$

  Thus, for $(w, x)$ in this region, manipulation increases $B$'s score by 1 and leaves $C$'s score unaffected.

Let $T$ be the set of all pairs $(w, x)$ such that $w + x \geq K$ and $N - w - x$ is odd.

We know that the maximum probability of drawing $(w, x) \in T$ is $\sim 1/2$, by Lemma F.1. Our strategy will be to show that the probability of $(w, x) \in S_A$ and the probability of $(w, x) \in T$ are close.

We begin by showing that the probability of $(w, x) \in R_1 \cap S_A$ and the probability of $(w, x) \in R_1 \cap T$ are close.

First consider the probability that

$$(w, x) \in R_1 \cap S_A \qquad \text{and} \qquad (w + 1, x) \in R_1 \cap S_A. \tag{F.9}$$

Notice that if $\lfloor K(L - w - x - 1)/(L - K) \rfloor = \lfloor K(L - w - x)/(L - K)$, then (F.9) cannot occur. Indeed, $(w, x) \in R_1 \cap S_A$ means that for suitable choices of $(y, z)$,

$$\frac{K(L - w - x - 1)}{L - K} + y < z \qquad \text{but} \qquad \frac{K(L - w - x)}{L - K} + y + 1 \geq z$$

which means that $\lfloor K(L - w - x)/(L - K) \rfloor + y = z - 1$. Then $\lfloor K(L - w - x)/(L - K) \rfloor + (N - w - x)$ must be odd (since $N - w - x = y + z$). Likewise, $(w+1, x) \in R_1 \cap S_A$ requires $\lfloor K(L - w - x - 1)/(L - K) \rfloor + (N - w - 1 - x)$ to be odd. But these expressions cannot both be odd if $\lfloor K(L - w - x - 1)/(L - K) \rfloor = \lfloor K(L - w - x)/(L - K) \rfloor$.

Let

$$V = \left\{ v \;\middle|\; \left\lfloor \frac{K(L - v - 1)}{L - K} \right\rfloor < \left\lfloor \frac{K(L - v)}{L - K} \right\rfloor \right\}.$$

Thus, the probability that (F.9) arises is at most the probability that $w + x \in V$, given that $(w + x, N - (w + x))$ is drawn from $\mathbf{M}(N; \alpha_w + \alpha_x, 1 - (\alpha_w + \alpha_x))$. Write

$\alpha_v = \alpha_w + \alpha_x$.

Since $(L-K)/K > \lambda - 1$, it follows that of any $\lambda$ consecutive integers, at most one can be in $V$.

Now, we claim that for any set $V$ with this property, there exists $a$ such that

$$\sum_{v \in V} \mathbf{P}\left( \begin{matrix} v \\ N-v \end{matrix} \middle| N; \begin{matrix} \alpha_v \\ 1-\alpha_v \end{matrix} \right) \leq \sum_{\substack{K \leq v \leq N-K \\ v \equiv a \mod \lambda - 1}} \mathbf{P}\left( \begin{matrix} v \\ N-v \end{matrix} \middle| N; \begin{matrix} \alpha_v \\ 1-\alpha_v \end{matrix} \right). \tag{F.10}$$

Indeed, choose $a$ to be the value of $v \in V$ for which $bP(v, N-v \mid N; \alpha_v, 1-\alpha_v)$ is maximized. Since this latter expression is unimodal in $v$, for any two consecutive elements of $V$ that differ by more than $\lambda - 1$, either the lower element can be increased by 1 or the higher element can be decreased by 1 in such a way that the expression on the left side of (F.10) is increased. We thus replace $V$ by a new set for which the left-hand side of (F.10) is higher than before. This operation cannot be repeated forever; when it terminates, it must be that every two consecutive elements of the current set differ by $\lambda - 1$. The resulting set clearly satisfies (F.10), and so the original set $V$ did as well.

By Lemma F.1, the right-hand side of (F.10) is $\lesssim 1/(\lambda - 1)$. Hence, the same holds for the probability of (F.9), since $w + x \in V$ is a necessary condition for (F.9).

Notice that

$$\begin{aligned} \sum_{(w,x) \in R_1 \cap S_A} \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix} \leq \sum_{(w,x) \in R_1 \cap T} \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix} + \\ \sum_{\substack{(w,x) \in R_1 \cap S_A \setminus T \\ (w+1,x) \in R_1 \cap T \setminus S_A}} \left[ \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix} - \mathbf{P}\begin{pmatrix} w+1 \\ x \end{pmatrix} \right] + \\ \sum_{\substack{(w,x) \in R_1 \cap S_A \setminus T \\ (w+1,x) \in R_1 \cap S_A}} \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix} + \\ \sum_{\substack{(w,x) \in R_1 \cap S_A \setminus T \\ (w+1,x) \notin R_1 \cap T}} \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix}. \end{aligned}$$

The second sum on the right-hand side is at most $e^{1/12}/\sqrt{2\pi K}$, by Lemma F.2 (notice that $(w,x) \in R_1$ ensures $w \geq K$). The third sum consists of those pairs $(w,x)$ satisfying (F.9), which have a total probability $\lesssim 1/(\lambda-1)$, by the preceding argument. For the fourth sum, notice that if $(w,x) \in R_1 \setminus T$ then $(w+1,x) \in T$, so the only pairs counted by this sum are those for which $(w+1,x) \notin R_1$ — that is, $w+1+x=L$. By Lemma C.7, these pairs have a total probability $\leq e^{1/12}/\sqrt{\pi L}$.

The bounds $e^{1/12}/\sqrt{\pi K}$, $e^{1/12}/\sqrt{2\pi L}$ both go to 0 as $N \to \infty$, and thus we get

$$\sum_{(w,x)\in R_1\cap S_A} \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix} - \sum_{(w,x)\in R_1\cap T} \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix} \lesssim \frac{1}{\lambda-1}. \tag{F.11}$$

This takes care of $R_1 \cap S_A$ for now. Next let us perform a similar analysis for pairs $(w,x) \in R_2 \cap S_A$.

Consider the possibility that

$$(w,x) \in R_2 \cap S_A \qquad \text{and} \qquad (w,x+1) \in R_2 \cap S_A. \tag{F.12}$$

If $(w,x) \in R_2 \cap S_A$, then for suitable choices of $(y,z)$,

$$y < \frac{K(L-w-x-1)}{L-K} + z \qquad \text{and} \qquad y+1 \geq \frac{K(L-w-x)}{L-K} + z.$$

This means that

$$\left\lceil \frac{K(L-w-x-1)}{L-K} \right\rceil = \left\lceil \frac{K(L-w-x)}{L-K} \right\rceil = y-z+1.$$

Hence, $\lceil K(L-w-x)/(L-K)\rceil + (N-w-x)$ must be odd; and if $(w,x+1) \in R_2 \cap S_A$, then $\lceil K(L-w-x)/(L-K)\rceil + (N-w-x-1)$ must be odd. These quantities cannot both be odd, however. So we see that (F.12) can never occur.

Thus, we can perform an analysis for the probability of $(w,x) \in R_2 \cap S_A$ that is entirely analogous to what was done for $(w,x) \in R_1 \cap S_A$, but our life is now simplified by the fact that (F.12) has probability zero (unlike its counterpart (F.9)). The result

is

$$\sum_{(w,x)\in R_2\cap S_A} \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix} - \sum_{(w,x)in R_2\cap T} \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix} \leq \frac{e^{1/12}}{\sqrt{\pi K}} + \frac{e^{1/12}}{\sqrt{\pi L}} \to 0. \tag{F.13}$$

Next we turn to $R_3$. Since $R_3 \cap S_A = \emptyset$, we simply have

$$\sum_{(w,x)\in R_3\cap S_A} \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix} = 0 \leq \sum_{(w,x)\in R_3\cap T} \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix}. \tag{F.14}$$

Finally, we claim that $R_4 \cap S_A \subseteq T$. Check: if $(w,x) \in R_4 \cap S_A$, then $y < z$ but $y + 1 \geq z$, so integrality implies $y = z - 1$, and hence $N - w - x = y + z$ is odd. So it is immediate that

$$\sum_{(w,x)\in R_4\cap S_A} \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix} \leq \sum_{(w,x)\in R_4\cap T} \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix}. \tag{F.15}$$

Finally, using the fact that every $(w,x) \in S_A$ must lie in one of the regions $R_1, \ldots, R_4$, we can combine (F.11), (F.13), (F.14), (F.15):

$$\begin{aligned}
\sum_{(w,x)\in S_A} \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix} &= \sum_{i=1}^{4} \sum_{(w,x)\in R_i\cap S_A} \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix} \\
&\lesssim \sum_{i=1}^{4} \sum_{(w,x)\in R_i\cap T} \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix} + \frac{1}{\lambda - 1} \\
&\leq \sum_{(w,x)\in T} \mathbf{P}\begin{pmatrix} w \\ x \end{pmatrix} + \frac{1}{\lambda - 1} \\
&\lesssim \frac{1}{2} + \frac{1}{\lambda - 1}.
\end{aligned}$$

Thus, we have proven (F.8). Combining with (F.7), as previously mentioned, gives the result.

□

# G Proofs of lower bounds

The proofs of the results from Section 4 are in this appendix (except for results that are proven in the main text). We present the proofs in the same order that they are sketched in the text.

**Proof of Theorem 4.5:** For any two candidates $A, B$, let $K^*(A;B)$ be the maximum number $K$ such that $f(K\ A, N+1-K\ B) \neq A$. By unanimity, $K^*(A;B) < N+1$.

Let us first assume that there are three candidates — which we may, by relabeling, call $A, B, C$ — such that

$$f(1\ A, K\ B, N-K\ C) \neq A \qquad \text{for all} \qquad K. \tag{G.1}$$

Note that this also implies $f(K\ B, N+1-K\ C) \neq A$ for all $K$ (otherwise, change one of the $B$ or $C$ votes to $A$, and monotonicity implies that (G.1) is violated).

Write $K^*$ for $K^*(B;C)$. Also write $\widetilde{K}^*$ for the maximum value of $K$ such that $f(1\ A, K\ B, N-K\ C) \neq B$.

We have

$$f(K\ B, N+1-K\ C) = B \qquad \text{for all} \qquad K > K^* \tag{G.2}$$

by definition, and

$$f(K\ B, N+1-K\ C) \neq B \qquad \text{for all} \qquad K < K^* \tag{G.3}$$

since otherwise monotonicity would imply $f(K^*\ B, N+1-K^*\ C) = B$, a contradiction. By similar arguments,

$$f(1\ A, K\ B, N-K\ C) = B \qquad \text{for all} \qquad K > \widetilde{K}^*; \tag{G.4}$$

$$f(1\ A, K\ B, N-K\ C) \neq B \qquad \text{for all} \qquad K < \widetilde{K}^*. \tag{G.5}$$

Notice that $\widetilde{K}^* \geq K^* - 1$, since otherwise $f(1\ A, \widetilde{K}^*+1\ B, N-\widetilde{K}^*\ C) = B$ and

$f(K^*\ B, N+1-K^*\ C) \neq B$ would violate monotonicity. We split into three cases:

(i) Suppose $\widetilde{K}^* = K^* - 1$. Let the manipulator's true preference be any ordering with $A$ ranked first and $B$ last; let the proposed manipulation be a vote for $C$; and let the manipulator's belief be $\phi = (\alpha_B\ B, (1-\alpha_B)\ C)$ with $\alpha_B = K^*/N$. Since the other voters all vote for $B$ or $C$, (G.2)-(G.5) imply that the manipulator cannot affect whether $B$ wins, unless the realized opponent-profile is $(K^*\ B, N-K^*\ C)$, in which case a vote for $A$ leads to $B$ winning and a vote for $C$ leads to $B$ losing. So considering the definition (2.3) of susceptibility with $\mathcal{C}^+ = \mathcal{C} \setminus \{B\}$, we have

$$\sigma \geq \mathbf{P}(K^*\ B, N-K^*\ C \mid N; \phi),$$

which is $\geq \sigma_N^*$ by Lemma 2.4.

(ii) Suppose $\widetilde{K}^* = K^*$. Let the manipulator's true preference be any ordering with $A$ ranked first and $B$ second; let the proposed manipulation be a vote for $B$; and again let the manipulator's belief be $\phi = (\alpha_B\ B, (1-\alpha_B)\ C)$ with $\alpha_B = K^*/N$. From (G.1) and the observation following it, no matter whether the manipulator votes for $A$ or $B$, $A$ cannot win. Again, (G.2)-(G.5) imply that the manipulator cannot affect whether or not $B$ wins, unless the realized opponent-profile is $(K^*\ B, N-K^*\ C)$ in which case a vote for $A$ leads to $B$ losing and a vote for $B$ leads to $B$ winning. So considering (2.3) with $\mathcal{C}^+ = \{A, B\}$, we again have

$$\sigma \geq \mathbf{P}(K^*\ B, N-K^*\ C \mid N; \phi) \geq \sigma_N^*.$$

(iii) Suppose $\widetilde{K}^* > K^*$. Let the manipulator's true preference be any ordering with $C$ ranked first and $B$ last; let the proposed manipulation be a vote for $A$; and let the belief be $\phi = (\alpha_B\ B, (1-\alpha_B)\ C)$ with $\alpha_B = (K^*+1)/N$. Once again, the manipulator cannot affect whether or not $B$ wins, unless the opponent-profile is $(K\ B, N-K\ C)$ for some $K$ with $K^* < K \leq \widetilde{K}^*$, in which case a vote for $C$ leads to $B$ winning and a vote for $A$ leads to $B$ losing. Considering (2.3) with

$\mathcal{C}^+ = \mathcal{C} \setminus \{B\}$, we again have

$$\sigma \geq \sum_{K=K^*+1}^{\widetilde{K}^*} \mathbf{P}(K\ B, N-K\ C \mid N; \phi) \geq \mathbf{P}(K^*+1\ B, N-K^*-1\ C \mid N; \phi) \geq \sigma_N^*.$$

This completes the proof of the inequality in case $A, B, C$ can be chosen so that (G.1) holds.

Now suppose no such $A, B, C$ exist. Choose $A, B, C$ so that $f(1\ A, K\ B, N - K\ C) = A$ for $K$ as large as possible. By assumption, there also exists $K'$ such that $f(1\ C, K'\ B, N - K'\ A) = C$, and by maximality $K' \leq K$. If $K < N$, then monotonicity implies $f(N - K'\ A, K'\ B, 1\ C) = A$, a contradiction. Therefore $K = N$, so that $f(1\ A, N\ B) = A$. Again by assumption, there exists $K''$ such that $f(1\ B, K''\ C, N - K''\ A) = B$. If $K'' < N$ then monotonicity implies $f(1\ A, N\ B) = B$, a contradiction. So $K'' = N$, or $f(1\ B, N\ C) = B$. By monotonicity again, $f(N\ B, 1\ C) = B$.

Suppose the manipulator's true preference ranks $C$ first and $B$ last; let the proposed manipulation be a vote for $A$, and let the belief be that everyone else votes for $B$ with probability 1. Then a truthful vote for $C$ leads to $B$ winning, while manipulation leads to $A$ winning, hence (taking $\mathcal{C}^+ = \mathcal{C} \setminus \{B\}$) we have susceptibility $\sigma = 1$.

This proves that the inequality $\sigma \geq \sigma_N^*$ always holds.

It remains to study the equality case. This proof roughly follows the above case analysis but requires further splitting into subcases. We wish to show that the inequality is strict if $f$ is not a majority rule. So there is a profile at which strictly more than half the voters vote for some candidate — say $C$ — but some other candidate wins — say $B$. We may assume $B$ and $C$ are chosen so as to maximize the number of voters voting for $C$ with $B$ winning.

By monotonicity, $B$ still wins when all the non-$C$ votes are replaced by $B$'s, and it follows that $K^*(B; C) \leq (N-2)/2$. Let $A$ be an arbtirary candidate distinct from $B$ and $C$. Define $\widetilde{K}^*$ as before. We review the cases from the preceding analysis, making amendments as needed. Note that assumption (G.1) was only used in case

(ii).

- In case (i), the same argument as before applies. Since $K^* < (N-1)/2$, Lemma 2.4 implies that the inequality at the end of case (i) holds strictly.

- In case (ii), if (G.1) holds, then the analysis goes through as before and again the final inequality holds strictly.

  Suppose (G.1) fails. Then we have $f(1\ A, K\ B, N-K\ C) = B$ whenever $K > K^*$, and $f(1\ A, K\ B, N-K\ C) = C$ whenever $K < K^*$ because the extremal choice of $B$ and $C$ implies that $C$ wins whenever at least $N+1-K^*$ voters vote for $C$. Hence, the failure of (G.1) can only happen for $K = K^*$: $f(1\ A, K^*\ B, N-K^*\ C) = A$. By monotonicity, we then have

$$f(J\ A, K\ B, N+1-J-K\ C) = A \qquad \text{for all} \qquad K \leq K^*,\ J+K-1 \geq K^*. \tag{G.6}$$

  And the extremal property of $B$ and $C$ implies that

$$f(J\ A, K\ B, N+1-J-K\ C) = C \qquad \text{whenever} \qquad J+K-1 < K^*. \tag{G.7}$$

  If $K^* \geq 1$ then (G.6) and (G.7) imply that we can use the triple $(B, A, C)$ instead of $(A, B, C)$: this triple has the same value of $K^*$, but falls into case (i), from which the proof is complete.

  Finally suppose $K^* = 0$. Then we have $f(1\ A, 1\ B, N-1\ C) = B$; $f(1\ A, N\ C) = A$ (and by monotonicity $f(K\ A, N+1-K\ C) = A$ for all $K \geq 1$); $f(1\ B, N\ C) = B$; and $f(N+1\ C) = C$. Let the manipulator have true preference ranking $C$ first, $B$ second, and $A$ last; let the proposed manipulation be a vote for $B$, and let the belief $\phi$ be $1/N\ A, (N-1)/N\ C$. If the realized opponent profile is that all others vote for $C$, then truthful voting leads to $C$ winning, while manipulating leads to $B$ winning. For any other possible opponent-profile, telling the truth leads to $A$ winning, and at least when the opponent-profile is $1\ A, N-1\ C$, manipulation leads to $B$ winning instead. It follows by taking $\mathcal{C}^+ = \mathcal{C} \setminus \{A\}$

that

$$\sigma \geq \mathbf{P}(1\ A, N-1\ C \mid N; \phi) > \sigma_N^*.$$

- In case (iii), if $K^* \leq (N-4)/2$, then the final inequality in case (iii) becomes strict, again by Lemma 2.4. So we may assume $K^* > (N-4)/2 \geq 0$.

  If $f(1\ A, K^*\ B, N-K^*\ C) = A$, we again have (G.6) and (G.7), so that just as in case (ii) above, we can replace the triple $(A, B, C)$ by $(B, A, C)$, and end up in case (i), for which the proof has been completed. (Note that this uses the assumption $K^* > 0$.)

  Finally, suppose $f(1\ A, K^*\ B, N-K^*\ C) \neq A$. We also have $f(1\ A, K^*\ B, N-K^*\ C) \neq B$ by the assumption of case (iii).

  As before, the extremal property of $B$ and $C$ implies $f(1\ A, K\ B, N-K\ C) = C$ for $K < K^*$. In this case, consider the same preferences, belief, and proposed manipulation as in the original analysis for case (ii). If the realized opponent-profile is $(K\ B, N-K\ C)$ for $K < K^*$, then $C$ wins regardless of whether the manipulator votes for $A$ or $B$. Otherwise, a vote for $B$ will ensure that $B$ wins, while a vote for $A$ will fail to ensure an outcome in $\mathcal{C}^+ = \{A, B\}$ if the realized opponent-profile is $(K^*\ B, N-K^*\ C)$. Hence (2.3) with $\mathcal{C}^+ = \{A, B\}$ gives

$$\sigma \geq \mathbf{P}(K^*\ B, N-K^*\ C \mid N; \phi) > \sigma_N^*.$$

This shows that $\sigma > \sigma_N^*$ in every possible case.

□

Next we proceed to the proof of Theorem 4.4. This proof makes reference to proof techniques from Theorem 4.7, which was given in the main text.

**Proof of Lemma 4.10:** For each $K = 0, \ldots, \overline{K}$, let $J(K)$ be the highest value such that $f(P_{J,K}) = A_j$, or $J(K) = \underline{J} - 1$ if no such value exists. By (i) and (iii), $f(P_{J,K}) = A_j$ for $J \leq J(K)$ and $= A_i$ for $J > J(K)$. Also (iii) ensures that $J(K-1) \leq J(K) + 1$ (whenever these quantities are defined).

Choose integer values $0 = K_0, K_1, K_2, \ldots, K_r = \overline{K}$, where any two successive $K_i$

differ by at most $40\sqrt{\widetilde{N}}/\kappa$ and with $r \leq \sqrt{\widetilde{N}}\kappa/20$. Certainly this can be done, as long as $N$ is sufficiently large.

Now, by (iv), $J(0) = \overline{J}$, while by (v), $J(\overline{K}) = \underline{J} - 1$. Therefore

$$J(0) - J(\overline{K}) > \overline{J} - \underline{J} > \kappa\widetilde{N}.$$

Therefore, there exists some $i \in \{1, \ldots, r\}$ such that

$$J(K_{i-1}) - J(K_i) > \frac{\kappa\widetilde{N}}{r} > 20\sqrt{\widetilde{N}}.$$

Put

$$\gamma = \frac{J(K_{i-1}) + J(K_i)}{2\widetilde{N}},$$

$$\delta_1 = \frac{K_{i-1} + \sqrt{2\widetilde{N}}}{\widetilde{N}}, \qquad \phi_1 = \begin{pmatrix} \gamma & \succ \\ \delta_1 & \succ' \\ 1-\gamma-\delta_1 & \succ'' \end{pmatrix},$$

$$\delta_2 = \min\left\{\frac{K_i - \sqrt{2\widetilde{N}}}{\widetilde{N}}, 1-\gamma\right\}, \qquad \phi_2 = \begin{pmatrix} \gamma & \succ \\ \delta_2 & \succ' \\ 1-\gamma-\delta_2 & \succ'' \end{pmatrix}.$$

It is straightforward to check that $\phi_1$ and $\phi_2$ are legitimate probability distributions (that is, all entries are nonnegative); the only nontrivial part is $K_i - \sqrt{2\widetilde{N}} \geq 0$ which follows from $K_i - K_{i-1} \geq J(K_{i-1}) - J(K_i) > 20\sqrt{\widetilde{N}}$.

We will show that

$$\overline{f}_{A_j}(\phi_1) > 3/4 \tag{G.8}$$

and

$$\overline{f}_{A_i}(\phi_2) > 3/4. \tag{G.9}$$

Suppose that the $\widetilde{N}$-profile $P = (x \succ, y \succ', z \succ'')$ is drawn according to $IID(\phi_1)$. If the inequalities

$$x \geq \underline{J} \tag{G.10}$$

$$y \geq K_{i-1} \tag{G.11}$$

$$x + y \leq K_{i-1} + J(K_{i-1}) \tag{G.12}$$

are satisfied, then we must have $f(P) = A_j$, using $f(J(K_{i-1}) \ \succ, K_{i-1} \ \succ', \widetilde{N} - J(K_{i-1}) - K_{i-1} \ \succ'') = A_j$ and the monotonicity relation (iii). Notice also that if

$$x \leq (3J(K_{i-1}) + J(K_i))/4 \tag{G.13}$$

$$y \leq K_{i-1} + 4\sqrt{\widetilde{N}} \tag{G.14}$$

are satisfied, then (G.12) will automatically hold.

Now we apply the same Chebyshev argument as in the proof of Lemma 4.2. We have $(x, \widetilde{N}-x) \sim \mathbf{M}(\widetilde{N}; \gamma, 1-\gamma)$, so (G.10) and (G.13) are satisfied unless $|x-E[x]| \geq (J(K_{i-1}) - J(K_i))/4$, which happens with probability

$$\Pr\left(|x - E[x]| \geq \frac{J(K_{i-1}) - J(K_i)}{4}\right) \leq \frac{Var(x)}{\left(\frac{J(K_{i-1})-J(K_i)}{4}\right)^2} \leq \frac{\widetilde{N}/4}{(5\sqrt{\widetilde{N}})^2} = \frac{1}{100}.$$

Likewise, $(y, \widetilde{N} - y) \sim \mathbf{M}(\widetilde{N}; \delta_1, 1 - \delta_1)$, to (G.11) and (G.14) are satisfied unless $|y - E[y]| \geq \sqrt{2\widetilde{N}}$, which happens with probability

$$\Pr(|y - E[y]| \geq \sqrt{2\widetilde{N}}) \leq \frac{Var(y)}{(\sqrt{2\widetilde{N}})^2} \leq \frac{\widetilde{N}/4}{2\widetilde{N}} = \frac{1}{8}.$$

We conclude that (G.10), (G.11), (G.13), (G.14) are all satisfied — and hence $f(P) = A_j$ — with probability at least $1 - 1/00 - 1/8 > 3/4$. This gives (G.8).

Similarly, suppose that the $\widetilde{N}$-profile $P = (x \ \succ, y \ \succ', z \ \succ'')$ is drawn according to $IID(\phi_2)$. If the inequalities

$$x \leq \overline{J} \tag{G.15}$$

$$y \leq K_i \tag{G.16}$$

$$x + y > K_i + J(K_i) \tag{G.17}$$

are satisfied, then we must have $f(P) = A_i$, using $f(J(K_i)+1 \ \succ, K_i \ \succ', \widetilde{N} - J(K_i) - K_i - 1 \ \succ'') = A_i$ and the monotonicity condition (iii). (Note that we cannot have $J(K_i) + K_i = \widetilde{N}$, because then $f(J(K_i) \ \succ, K_i \ \succ', 0 \ \succ'') = A_j$, together with $f(\widetilde{N} - \overline{K} \ \succ, \overline{K} \ \succ', 0 \ \succ'') = A_i$ from (v), would give a contradiction to (iii).)

Notice also that if

$$x \geq (J(K_{i-1}) + 3J(K_i))/4 \tag{G.18}$$

$$y > K_i - 4\sqrt{N} \tag{G.19}$$

are satisfied, then (G.17) will automatically hold.

If $\delta_2 = (K_i - \sqrt{2\widetilde{N}})/\widetilde{N}$, then exactly the same Chebyshev arguments as before give that (G.15), (G.16), (G.18), (G.19) are all satisfied — and hence $f(P) = A_i$ — with probability greater than 3/4. Otherwise, we necessarily have $x + y = \widetilde{N}$ so that (G.17) is always satisfied, and then the same arguments show that (G.15), (G.16) are both satisfied with probability greater than 3/4. In either case, then, we get (G.9).

Now that (G.8) and (G.9) are proven, we use Lemma 4.9 to complete the argument. Notice that $\phi_2 - \phi_1 = \Delta(\succ' - \succ'')$, where

$$0 \leq \Delta < \frac{K_i}{\widetilde{N}} - \frac{K_{i-1}}{\widetilde{N}} < \frac{40}{\kappa\sqrt{\widetilde{N}}}.$$

By (ii), preferences $\succ'$ and $\succ''$ rank $A_i$ and $A_j$ in the same way. If they both rank $A_i$ above $A_j$, then let $\mathcal{C}^+$ be the set of candidates weakly preferred to $A_i$ under $\succ''$. Lemma 4.9(a) gives

$$\sum_{A \in \mathcal{C}^+} \overline{f}_A(\phi_2) - \sum_{A \in \mathcal{C}^+} \overline{f}_A(\phi_1) \leq \widetilde{N}\Delta\sigma. \tag{G.20}$$

By (G.8) and (G.9), the left-hand side of (G.20) is at least $3/4 - (1 - 3/4) = 1/2$, so

$$\frac{1}{2} \leq \widetilde{N}\Delta\sigma \leq \sqrt{\widetilde{N}}\frac{40}{\kappa}\sigma.$$

If $\succ'$ and $\succ''$ both rank $A_j$ above $A_i$, then let $\mathcal{C}^+$ be the set of candidates weakly

preferred to $A_j$ under $\succ'$. Lemma 4.9(a) gives

$$\sum_{A\in\mathcal{C}^+} \overline{f}_A(\phi_1) - \sum_{A\in\mathcal{C}^+} \overline{f}_A(\phi_2) \leq \widetilde{N}\Delta\sigma$$

and we again arrive at $1/2 \leq \sqrt{\widetilde{N}}(40/\kappa)\sigma$. Thus in either case we have

$$\sigma \geq \frac{\kappa}{80}\widetilde{N}^{-1/2}$$

which is the promised result. □

**Proof of Theorem 4.4:** We first suppose there are just three candidates, $\mathcal{C} = \{A, B, C\}$. For every $K$, we have $f(K\ ABC, \widetilde{N} - K\ BCA) \in \{A, B\}$ by Pareto efficiency (and moreover it is $B$ when $K = 0$ and $A$ when $K = \widetilde{N}$). Moreover, by monotonicity, if this expression equals $A$ for some $K$ then it also equals $A$ for all higher $K$. So, writing

$$K_{AB} = \max\left\{K \;\middle|\; f\begin{pmatrix} K\ ABC \\ \widetilde{N} - K\ BCA \end{pmatrix} = B\right\},$$

we have $f(K\ ABC, \widetilde{N} - K\ BCA) = B$ if $K \leq K_{AB}$ and $A$ if $K > K_{AB}$. Likewise define

$$K_{BC} = \max\left\{K \;\middle|\; f\begin{pmatrix} K\ BCA \\ \widetilde{N} - K\ CAB \end{pmatrix} = C\right\},$$

$$K_{CA} = \max\left\{K \;\middle|\; f\begin{pmatrix} K\ CAB \\ \widetilde{N} - K\ ABC \end{pmatrix} = A\right\},$$

$$K_{CB} = \max\left\{K \mid \;\middle|\; f\begin{pmatrix} K\ CBA \\ \widetilde{N} - K\ BAC \end{pmatrix} = B\right\},$$

$$K_{BA} = \max\left\{K \;\middle|\; f\begin{pmatrix} K\ BAC \\ \widetilde{N} - K\ ACB \end{pmatrix} = A\right\},$$

$$K_{AC} = \max\left\{ K \;\middle|\; f\begin{pmatrix} K\ ACB \\ \widetilde{N} - K\ CBA \end{pmatrix} = C \right\}.$$

We now have two cases.

(i) $K_{AB} + K_{BC} + K_{CA} + K_{CB} + K_{BA} + K_{AC} > 7\widetilde{N}/2$.

In this case one of the three quantities $K_{AB} + K_{BA}, K_{BC} + K_{CB}, K_{CA} + K_{AC}$ is greater than $7\widetilde{N}/6$. Without loss of generality we will assume $K_{CA} + K_{AC} > 7\widetilde{N}/6$, which is the case shown in Figure 4.4.

Let

$$K^* = \max\left\{ K \;\middle|\; f\begin{pmatrix} K\ CAB \\ \widetilde{N} - K\ ACB \end{pmatrix} = A \right\}.$$

As before, $f(K\ CAB, \widetilde{N} - K\ ACB) = A$ for $K \leq K^*$ and $= C$ for $K > K^*$. Then one of the following two inequalities must hold:

$$K_{CA} - K^* > \frac{\widetilde{N}}{12}; \qquad K_{AC} - (N - K^*) > \frac{\widetilde{N}}{12}.$$

We assume henceforth that the first inequality holds (otherwise, the argument is the same with $A$ and $C$ reversed).

Now we apply Lemma 4.10 with

$$\succ = CAB, \qquad \succ' = ACB, \qquad \succ'' = ABC,$$

$$\underline{J} = K^* + 1, \qquad \overline{J} = K_{CA}, \qquad \kappa = \frac{1}{13}, \qquad \overline{K} = \widetilde{N} - (K^* + 1),$$

$$A_i = C, \qquad A_j = A.$$

The condition $\overline{J} - \underline{J} > \kappa\widetilde{N}$ is evidently satsified (as long as $N$ is large), so we need to verify conditions (i)-(v) of the lemma. (i) follows from Pareto efficiency. (ii) is immediate. (iii) follows from monotonicity. (iv) is the definition of $K_{CA}$ (and our monotonicity observation earlier). (v) is the definition of $K^*$. Hence, the lemma applies, and $\sigma$ is bounded by a constant times $N^{-1/2}$. This takes

care of case (i).

(ii) $K_{AB} + K_{BC} + K_{CA} + K_{CB} + K_{BA} + K_{AC} \leq 7\widetilde{N}/2$.

In this case one of the quantities $K_{AB} + K_{BC} + K_{CA}$, $K_{CB} + K_{BA} + K_{AC}$ is at most $7\widetilde{N}/4$. Without loss of generality we will assume

$$K_{AB} + K_{BC} + K_{CA} \leq \frac{7\widetilde{N}}{4}. \tag{G.21}$$

We can now focus our attention on the $ABC - BCA - CAB$ simplex.

We will also assume for now the inequalities

$$K_{AB} + K_{BC}, K_{BC} + K_{CA}, K_{CA} + K_{AB} \geq \frac{89}{90}\widetilde{N}. \tag{G.22}$$

Afterwards we will come back to address the (easier) case where one of these inequalities is violated.

An outline of our argument is illustrated in Figure G.1. In the top-left panel, the dots marked on the edges of the simplex are the profiles ($K_{AB}$ $ABC, \widetilde{N} - K_{AB}$ $BCA$), ($K_{BC}$ $BCA, \widetilde{N} - K_{BC}$ $CAB$), and ($K_{CA}$ $CAB, \widetilde{N} - K_{CA}$ $ABC$). The assumption $K_{AB}+K_{BC}+K_{CA} \leq 7\widetilde{N}/4$ ensures that the downward-pointing triangle in the figure has side length at least $\widetilde{N}/4$. Consider the profile at the center of the triangle, and without loss of generality assume that the winner there is $A$. Then consider the smaller downward-pointing triangle (shown in the bottom two panels). Using monotonicity we can show that at each profile in the smaller triangle, $f$ must choose either $A$ or $B$. If $f$ chooses $A$ at the center of the smaller triangle, then consider the shaded trapezoid in the bottom-left panel of Figure G.1. By monotonicity arguments, $f$ chooses either $A$ or $C$ at each profile in the trapezoid, and chooses $A$ near the left edge and $C$ at the right edge. Then we can apply Lemma 4.10 to this trapezoid. If instead $f$ chooses $B$ at the center of the smaller triangle, then we consider the parallelogram shown in the bottom-right panel, and similarly apply Lemma 4.10.

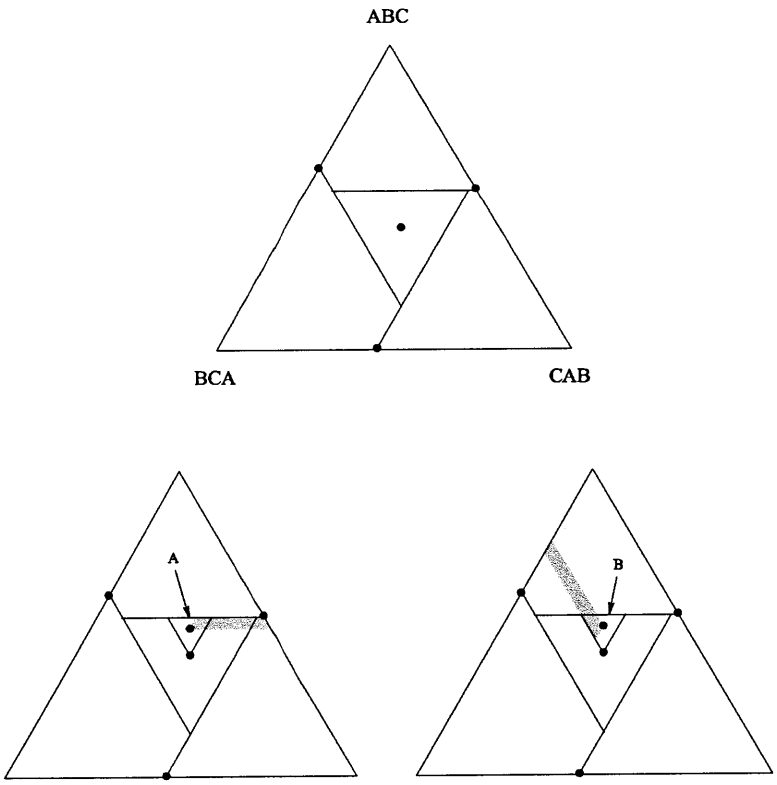


Figure G.1: Proof of Theorem 4.4 (case (ii))

Now we begin the proof properly. Let $P_0$ be a profile with

$$P_0 = \begin{pmatrix} x_0\ ABC \\ y_0\ BCA \\ z_0\ CAB \end{pmatrix} \approx \begin{pmatrix} (\widetilde{N} + K_{AB} + K_{BC} - 2K_{CA})/3\ ABC \\ (\widetilde{N} + K_{BC} + K_{CA} - 2K_{AB})/3\ BCA \\ (\widetilde{N} + K_{CA} + K_{AB} - 2K_{BC})/3\ CAB \end{pmatrix},$$

where the approximation means that we add or subtract at most 1 to each component to ensure $x_0, y_0, z_0$ are integers. Inequality (G.21), together with (G.22), ensure that $x_0, y_0, z_0$ are all positive. We have $f(P_0) = A$, $B$, or $C$. Without loss of generality, suppose henceforth that $f(P_0) = A$.

Now take

$$T = \left\{ \begin{pmatrix} \widetilde{N} - s - t\ ABC \\ s\ BCA \\ t\ CAB \end{pmatrix} \middle| \ s \leq y_0, t \leq z_0, s + t \geq K_{CA} \right\}.$$

Note that any profile $P \in T$ for which $t = z_0$ is obtained from $P_0$ by changing some $BCA$ votes to $ABC$, so by monotonicity $f(P) = A$. Consequently, we cannot have $f(P) = C$ for any $P \in T$: if $f(\widetilde{N} - s - t\ ABC, s\ BCA, t\ CAB) = C$, then by monotonicity $f(\widetilde{N} - s - z_0\ ABC, s\ BCA, z_0\ CAB) = C$, but this profile is also in $T$ and we showed that $A$ must win there, a contradiction. Hence, $f(P) \in \{A, B\}$ for all $P \in T$.

Let $P_1$ be a profile with

$$P_1 = \begin{pmatrix} x_1\ ABC \\ y_1\ BCA \\ z_1\ CAB \end{pmatrix} \approx \begin{pmatrix} (7\widetilde{N} + K_{AB} + K_{BC} - 8K_{CA})/9\ ABC \\ (\widetilde{N} - 5K_{AB} + 4K_{BC} + 4K_{CA})/9\ BCA \\ (\widetilde{N} + 4K_{AB} - 5K_{BC} + 4K_{CA})/9\ CAB \end{pmatrix}.$$

This profile is the "center of the smaller triangle" in Figure G.1. Again, one can verify that all components are positive. Moreover, $P_1 \in T$: all of the relevant inequalities reduce (up to negligible rounding error) to $K_{AB} + K_{BC} + K_{CA} \leq 2\widetilde{N}$, which is true by (G.21). Therefore, $f(P_1) \in \{A, B\}$. We have two cases.

If $f(P_1) = A$, then we will apply Lemma 4.10 with

$$\succ = ABC, \qquad \succ' = BCA, \qquad \succ'' = CAB,$$

$$\underline{J} = x_1, \qquad \overline{J} = \widetilde{N} - K_{CA} - 4, \qquad \kappa = \frac{1}{40}, \qquad \overline{K} = y_1 - (\widetilde{N} - K_{CA} - 4 - x_1),$$

$$A_i = A, \qquad A_j = C.$$

The required inequality $\overline{J} - \underline{J} > \kappa\widetilde{N}$ follows directly from (G.21). We proceed to verify conditions (i)-(v) of the lemma.

To verify condition (i), suppose for contradiction that $f(J\ ABC, K\ BCA, \widetilde{N} - J - K\ CAB) = B$ for some $\underline{J} \leq J \leq \overline{J}$ and $0 \leq K \leq \overline{K}$. By monotonicity, $f(J\ ABC, y_1\ BCA, \widetilde{N} - J - y_1\ CAB) = B$ also. (Note that $K \leq \overline{K} \leq y_1$; and this profile is well-defined since (G.21) and (G.22) imply $J + y_1 < \widetilde{N}$). But since $J \geq x_1$, $f(P_1) = A$ and monotonicity imply $f(J\ ABC, y_1\ BCA, \widetilde{N} - J - y_1\ CAB) = A$, a contradiction. Thus condition (i) of Lemma 4.10 holds.

Condition (ii) is immediate. (iii) follows from monotonicity given (i): if $f(P_{J,K}) = A$ then $f(P_{J+1,K-1}) = A$ by monotonicity, and $f(P_{J+1,K})$ cannot equal $C$ because then monotonicity would require $f(P_{J,K}) = C$, so $f(P_{J+1,k}) = A$ instead. (iv) follows from the definition of $K_{CA}$. And (v) holds because each of the relevant profiles $P_{J,\overline{K}}$ lies in the set $T$ (checking the relevant linear inequalities is straightforward), hence $f(P_{J,\overline{K}}) = A$ or $B$; since we have already ruled out $B$ with condition (i), we must have $f(P_{J,\overline{K}}) = A$ for each $J$, and condition (v) is satisfied. This checks all the conditions to apply Lemma 4.10, and we conclude that $\sigma$ is bounded below by a constant times $N^{-1/2}$.

If on the other hand $f(P_1) = B$, then we will apply Lemma 4.10 with

$$\succ = BCA, \qquad \succ' = CAB, \qquad \succ'' = ABC,$$

$$\underline{J} = y_1, \qquad \overline{J} = y_0 - 4, \qquad \kappa = \frac{1}{40}, \qquad \overline{K} = z_1,$$

$$A_i = B, \qquad A_j = A.$$

Again, the requirement $\overline{J} - \underline{J} > \kappa\widetilde{N}$ follows from (G.21), so we proceed to verify conditions (i)-(v).

If $f(J\ BCA,\ K\ CAB,\ \widetilde{N} - J - K\ ABC) = C$ for some $(J, K)$, then by monotonicity we also have $f(J\ BCA, z_1\ CAB, \widetilde{N} - J - z_1\ ABC) = C$. ((G.21) and (G.22) ensure this is a valid profile.) But this profile lies in $T$, so we should have $f(J\ BCA, z_1\ CAB, \widetilde{N} - J - z_1\ ABC) \in \{A, B\}$, a contradiction. This shows that condition (i) is satisfied. Condition (ii) is immediate. (iii) follows from monotonicity given (i): if $f(P_{J,K}) = B$, then $f(P_{J+1,K-1}) = B$ by monotonicity, and we cannot have $f(P_{J+1,K}) = A$ since then monotonicity would imply $f(P_{J,K}) = A$ as well, so we must have $f(P_{J+1,K}) = B$. (iv) follows from $\widetilde{N} - \overline{J} > K_{AB}$ (which in turn follows from (G.21)). Finally, $f(P_{\underline{J},\overline{K}}) = f(P_1) = B$, so $f(P_{J,\overline{K}}) = B$ for all $J$ (by condition (iii)), verifying (v). So we have checked all the conditions, and Lemma G.21 applies. We again conclude that $\sigma$ is bounded below by a constant times $N^{-1/2}$.

This completes the proof of case (ii) of the theorem as long as (G.22) is satisfied. It remains to address the case where (G.22) is violated. Without loss of generality, we assume that

$$K_{CA} + K_{AB} < \frac{89}{90}\widetilde{N}.$$

Then we can apply Lemma 4.10 with

$$\succ = ABC, \qquad \succ' = BCA, \qquad \succ'' = CAB,$$

$$\underline{J} = K_{AB} + 1, \qquad \overline{J} = \widetilde{N} - K_{CA} - 1, \qquad \kappa = \frac{1}{100}, \qquad \overline{K} = \widetilde{N} - K_{AB} - 1,$$

$$A_i = A, \qquad A_j = C.$$

It is clear that $\overline{J} - \underline{J} > \kappa\widetilde{N}$ as long as $N$ is large, so we check (i)-(v).

For (i), suppose $f(J\ ABC, K\ BCA, \widetilde{N} - J - K\ CAB) = B$ for some $J, K$ with $J \geq K_{AB} + 1$. By monotonicity, $f(J\ ABC, \widetilde{N} - J\ BCA) = B$ also.

This contradicts $J > K_{AB}$. Then (i) follows. (ii) is immediate. (iii) holds by monotonicity: if $f(P_{J,K}) = A$, then $f(P_{J+1,K-1}) = A$ by monotonicity directly; and $f(P_{J+1,K}) = C$ would imply $f(P_{J,K}) = C$ by monotonicity, a contradiction, so from (i) we must have $f(P_{J+1,K}) = A$ instead. (iv) follows from the definition of $K_{CA}$, and (v) follows from the definition of $K_{AB}$. Thus all the conditions hold and once again Lemma 4.10 assures us that $\sigma$ is bounded below by a constant times $N^{-1/2}$.

This completes the analysis of cases (i) and (ii). We have had to apply Lemma 4.10 with only finitely many values of $\kappa$, so if we simply let $c$ be the smallest of the corresponding values of $c(\kappa)$, then we have $\sigma \geq cN^{-1/2}$ in every subcase (as always, assuming $N$ is sufficiently large).

Finally, the foregoing analysis assumed that $\mathcal{C}$ consisted of just three candidates. If there are more than three candidates, then let $A, B, C$ be any three of them, and restrict attention to $(N+1)$-profiles (and beliefs) at which each voter ranks $A, B, C$ higher than any other candidate, with the remaining candidates all ranked according to some fixed order. By Pareto efficiency, only $A, B$, or $C$ can win at any such profile. Then, all of the preceding analysis carries through directly, with the preferences $ABC$ replaced by $ABC\ldots$, $BCA$ replaced by $BCA\ldots$, and so forth.

□

Next, we round out Subsection 4.5 by supplying the proof of Theorem 4.2.

**Proof of Theorem 4.2:** Let $c_1$ be the constant given by Lemma C.8. Take $A, B, C$ to be any three different candidates. We consider two possibilities.

(i) Suppose there exists some $K$ such that $f(K\ B, N+1-K\ C) \notin \{B, C\}$. Let $S \subseteq \{1, \ldots, N\}$ be the set of all such values. Let $\alpha$ be the value given by Lemma C.8 for the set $S$. Put $\phi = (\alpha\ B, 1-\alpha\ C)$. The conclusion of the lemma can be written as

$$\Pr_{IID(\phi)}(f(C,P) \notin \{B,C\}) - \Pr_{IID(\phi)}(f(B,P) \notin \{B,C\}) \geq \frac{c_1}{N},$$

where the probabilities are over opponent-profiles $P$ drawn according to $IID(\phi)$; or equivalently,

$$\Pr_{IID(\phi)}(f(B,P) \in \{B,C\}) - \Pr_{IID(\phi)}(f(C,P) \in \{B,C\}) \geq \frac{c_1}{N}. \qquad \text{(G.23)}$$

If the manipulator's true preference ranks $C$ first and $B$ second, then consider the manipulation to reporting $C$, with the top-set $\mathcal{C}^+ = \{B,C\}$. The left side of (G.23) is $\leq \sigma$, by (2.3). So we get $\sigma \geq c_1/N$ in this case.

(ii) Suppose that $f(K\ B, N+1-K\ C) \in \{B,C\}$ for all $K$. Assume that $\sigma < 1/(N+1)$ (otherwise we are done). We will first show that there exists exactly one value of $K$ such that $f(1\ A, K\ B, N-K\ C) \notin \{B,C\}$.

For any $\alpha \in [0,1]$, consider (2.3) for a manipulator with true preference $B \ldots C$, considering a manipulation to $A$, with belief $\phi = (\alpha\ B, 1-\alpha\ C)$ and $\mathcal{C}^+ = \mathcal{C} \setminus \{C\}$. We get

$$\begin{aligned}
&\sum_{K=0}^{N} \mathbf{P}\left(\begin{array}{c|cc} K & N; & \alpha \\ N-K & & 1-\alpha \end{array}\right) \times \\
&\quad \left[\mathbf{I}\left(f\begin{pmatrix} 1\ A \\ K\ B \\ N-K\ C \end{pmatrix} \in \mathcal{C}^+\right) - \mathbf{I}\left(f\begin{pmatrix} K+1\ B \\ N-K\ C \end{pmatrix} \in \mathcal{C}^+\right)\right] \\
&\qquad \leq \sigma \\
&\qquad < \frac{1}{N+1}.
\end{aligned}$$

Now integrate over $\alpha$ from 0 to 1, using the well-known identity $\int_0^1 \binom{N}{K}\alpha^K(1-\alpha)^{N-K}\,d\alpha = 1/(N+1)$. (The identity can be proven by showing that the integral is equal at two successive values of $K$, since the difference of the integrals at $K$

and $K+1$ is $\frac{1}{N+1}\binom{N+1}{K+1}\alpha^{K+1}(1-\alpha)^{N-K}|_0^1 = 0$.) This gives

$$\sum_{K=0}^{N} \frac{1}{N+1}\left[\mathbf{I}\left(f\begin{pmatrix} 1\ A \\ K\ B \\ N-K\ C \end{pmatrix} \in \mathcal{C}^+\right) - \mathbf{I}\left(f\begin{pmatrix} K+1\ B \\ N-K\ C \end{pmatrix} \in \mathcal{C}^+\right)\right] < \frac{1}{N+1}.$$

Applying (2.3) for a manipulator with true preference $A \ldots C$, considering a manipulation to $B$, gives the same inequality with the left-hand side negated. Hence, after multiplying through by $N+1$, we get

$$\left|\sum_{K=0}^{N} \mathbf{I}\left(f\begin{pmatrix} 1\ A \\ K\ B \\ N-K\ C \end{pmatrix} \neq C\right) - \sum_{K=0}^{N} \mathbf{I}\left(f\begin{pmatrix} K+1\ B \\ N-K\ C \end{pmatrix} \neq C\right)\right| < 1.$$

The left side is an integer, so it must be zero.

Subtracting both of the sums from $N+1$, we get the simpler equation

$$\sum_{K=0}^{N} \mathbf{I}\left(f\begin{pmatrix} 1\ A \\ K\ B \\ N-K\ C \end{pmatrix} = C\right) - \sum_{K=0}^{N} \mathbf{I}\left(f\begin{pmatrix} K+1\ B \\ N-K\ C \end{pmatrix} = C\right) = 0.$$

Thus, the number of profiles with one $A$ vote and all other votes $B$ or $C$, such that $C$ wins, equals the number of profiles with all $B$ or $C$ votes, and at least one $B$ vote, such that $C$ wins. This is in turn equal to the total number of profiles with all $B$ or $C$ votes, such that $C$ wins, minus one (since $f(N+1\ C) = C$ by unanimity).

By the same argument with $B$ and $C$ reversed, we see that the number of profiles with one $A$ vote and all other votes $B$ or $C$, such that $B$ wins, equals the total number of profiles with all $B$ or $C$ votes, such that $B$ wins, minus one.

Adding these two quantities, and using the fact that $f(K\ B, N+1-K\ C) \in$

$\{B, C\}$ for all $K$ by assumption, we see that there are exactly $N$ profiles of the form $(1\ A, K\ B, N-K\ C)$ at which either $B$ or $C$ wins. Hence, there is exactly one profile of this form at which some other candidate wins, as claimed.

Let $K^*$ be the unique value for which $f(1\ A, K^*\ B, N-K^*\ C) \notin \{B, C\}$.

Now let $K_B$ be the minimum value such that $f(K_B\ B, N-K_B\ C) = B$. Let $K_C$ be the maximum value such that $f(K_C\ B, N-K_C\ C) = C$. If $f$ were simple over $B, C$, we would have $K_B = K_C + 1$; but we no longer have this assumption. Instead, we will show that $K_B$ and $K_C$ are both close to $K^*$, and therefore close to each other; this in turn will allow us to repeat the argument from the proof of Theorem 4.7.

Let $S = \{K \mid f(K\ B, N+1-K\ C) = C\}$. Let $\alpha_C \geq K_C/N$ be the value given by Lemma C.8 for this set. Assume that $\sigma < c_1/3N$ (otherwise we are done).

Consider a manipulator with belief $\phi_C = (\alpha_C\ B, 1-\alpha_C\ C)$, preference $A \ldots B$, manipulating to $C$. We will write $\mathbf{P}[K]$ rather than $\mathbf{P}(K, N-K \mid N; \phi_C)$ to save on notation. The manipulation cannot decrease the probability of $B$ by more than $\sigma$, hence

$$\sum_{K=0}^{N} \mathbf{P}[K] \left[ \mathbf{I}\left( f\begin{pmatrix} 1\ A \\ K\ B \\ N-K\ C \end{pmatrix} = B \right) - \mathbf{I}\left( f\begin{pmatrix} K\ B \\ N+1-K\ C \end{pmatrix} = B \right) \right] \leq \frac{c_1}{3N}. \tag{G.24}$$

Similarly, a manipulator with the same belief and preference $A \ldots C$, manipulating to $B$, cannot decrease the probability of $C$ by more than $\sigma$, hence

$$\sum_{K=0}^{N} \mathbf{P}[K] \left[ \mathbf{I}\left( f\begin{pmatrix} 1\ A \\ K\ B \\ N-K\ C \end{pmatrix} = C \right) - \mathbf{I}\left( f\begin{pmatrix} K+1\ B \\ N-K\ C \end{pmatrix} \right) = C \right] \leq \frac{c_1}{3N}. \tag{G.25}$$

Now add (G.24) and (G.25). Notice that the $f(1\ A, K\ B, N-K\ C)$ terms add

up to cover each possible value of $K$ exactly once, except for $K = K^*$. Thus we get

$$1 - \mathbf{P}[K^*] - \sum_{K=0}^{N} \mathbf{P}[K] \times \left[ \mathbf{I}\left( f\begin{pmatrix} K & B \\ N+1-K & C \end{pmatrix} = B \right) + \mathbf{I}\left( f\begin{pmatrix} K+1 & B \\ N+1-K & C \end{pmatrix} = C \right) \right] \leq \frac{2c_1}{3N}. \tag{G.26}$$

But the remaining terms on the left side come under control because

$$\begin{aligned} &\sum_{K=0}^{N} \mathbf{P}[K] \left[ \mathbf{I}\left( f\begin{pmatrix} K & B \\ N+1-K & C \end{pmatrix} = B \right) + \mathbf{I}\left( f\begin{pmatrix} K+1 & B \\ N-K & C \end{pmatrix} = C \right) \right] \\ &= \sum_{K=0}^{N} \mathbf{P}[K] \left[ 1 - \mathbf{I}\left( f\begin{pmatrix} K & B \\ N+1-K & C \end{pmatrix} = C \right) + \mathbf{I}\left( f\begin{pmatrix} K+1 & B \\ N-K & C \end{pmatrix} = C \right) \right] \\ &= 1 + \sum_{K=0}^{N} [\mathbf{P}[K-1] - \mathbf{P}[K]]\, \mathbf{I}\left( f\begin{pmatrix} K & B \\ N+1-K & C \end{pmatrix} = C \right) \\ &\leq 1 - \frac{c_1}{N} \end{aligned}$$

where the second equality comes from reindexing the sum, and the final inequality comes from Lemma C.8.

Combining with (G.26) gives

$$1 - \mathbf{P}[K^*] - \left[1 - \frac{c_1}{N}\right] \leq \frac{2c_1}{3N}$$

or, finally,

$$\mathbf{P}\left( \begin{matrix} K^* \\ N-K^* \end{matrix} \,\middle|\, N; \begin{matrix} \alpha_C \\ 1-\alpha_c \end{matrix} \right) \geq \frac{c_1}{3N}. \tag{G.27}$$

Now combining (G.27) with Lemma C.4 gives

$$\frac{c_1}{3N} \leq e^{-N\frac{(\alpha_C - K^*/N)^2}{2}}$$

from which

$$\left|\alpha_C - \frac{K^*}{N}\right| \leq \sqrt{\frac{\ln N - \ln(c_1/3)}{N}}.$$

As long as $N$ is sufficiently large, the right-hand side is $\leq N^{-1/3}$. So we can conclude

$$K_C \leq \alpha_C N \leq K^* + N^{2/3}.$$

Now, exactly the same argument with the roles of $B$ and $C$ reversed leads to the conclusion that

$$K_B \geq K^* - N^{2/3}.$$

Therefore, we have

$$K_C - K_B \leq 2N^{2/3}. \tag{G.28}$$

This is the assertion that $K_B$ and $K_C$ are close to each other, as promised. (Notice also from the definitions that $K_B \leq K_C + 1$.)

From here, we will assume that $f$ has susceptibility $\sigma < 1/\widetilde{N}$ and obtain a contradiction, following the same steps as for Theorem 4.7. As long as $N$ is large enough, we may assume that $K_B \leq 2\widetilde{N}/3$ (otherwise $K_C \geq \widetilde{N}/3$, so just switch $B$ and $C$). Let

$$\phi_1 = (\alpha_1\ B, 1 - \alpha_1\ C) \qquad \text{with} \qquad \alpha_1 = \min\{\frac{K_C + \sqrt{2\widetilde{N}}}{\widetilde{N}}, 1\}.$$

Whenever more than $K_C$ voters vote for $B$ and the rest vote for $C$, $B$ wins; so the same Chebyshev argument as before gives $\overline{f}(\phi_1) = (\gamma_1\ B, 1 - \gamma_1\ C)$ where $\gamma_1 \geq 7/8$. Let

$$\phi_2 = (\alpha_2\ B, 1 - \alpha_2\ C) \qquad \text{with} \qquad \alpha_2 = \max\{\frac{K_B - \sqrt{2\widetilde{N}}}{\widetilde{N}}, 0\},$$

and obtain $\overline{f}(\phi_2) = (\gamma_2\ B, 1 - \gamma_2\ C)$ where $\gamma_2 \leq 1/8$.

Write $\phi_1 - \phi_2 = \Delta(B - C)$. On account of (G.28), we have

$$\Delta = \alpha_1 - \alpha_2 \leq 2\sqrt{2}\widetilde{N}^{-1/2} + 2\widetilde{N}^{-1/3} \leq 3\widetilde{N}^{-1/3}$$

as long as $N$ is large. Again taking $c_0$ to be the constant from Lemma 4.9, we have

$$c_0\widetilde{N}\Delta\sigma < 3c_0\widetilde{N}^{-1/3} < \frac{1}{8}$$

as long as $N$ is large. Exactly as before, we now define $\phi_3 = \phi_1 + \Delta(A - B)$ and $\phi_4 = \phi_1 + \Delta(A - C)$, and apply Lemma 4.9 to each of the pairs connected by thick lines in Figure 4.5, obtaining constraints on the values of $\overline{f}(\phi_3)$ and $\overline{f}(\phi_4)$ until we reach a contradiction.

□

Finally, we give proofs of the ingredients for Theorem 4.1. We begin with Lemma 4.11.

**Proof of Lemma 4.11:** Let $v$ denote the four-way difference on the left-hand side of (4.7).

Put

$$w_1 = \overline{f}\begin{pmatrix} \alpha\ \succ_1 \\ \beta\ \succ_3 \\ \gamma\ \phi \end{pmatrix} - \overline{f}\begin{pmatrix} \alpha\ \succ_2 \\ \beta\ \succ_3 \\ \gamma\ \phi \end{pmatrix},$$

$$w_2 = \overline{f}\begin{pmatrix} \alpha\ \succ_1 \\ \beta\ \succ_4 \\ \gamma\ \phi \end{pmatrix} - \overline{f}\begin{pmatrix} \alpha\ \succ_2 \\ \beta\ \succ_4 \\ \gamma\ \phi \end{pmatrix}.$$

Apply Lemma 4.9(b) twice to the difference represented by $w_1$: once letting $\mathcal{C}'$ be the set of candidates $A \neq A_i, A_j$ such that $(w_1)_A \geq 0$, and once letting $\mathcal{C}'$ be the set of candidates $A \neq A_i, A_j$ such that $(w_1)_A < 0$. We obtain

$$\sum_{A \neq A_i, A_j} |(w_1)_A| \leq 2c_0\widetilde{N}\alpha\sigma \leq 2c_0\widetilde{N}\sigma.$$

Likewise,

$$\sum_{A \neq A_i, A_j} |(w_2)_A| \leq 2c_0 \widetilde{N}\sigma.$$

Then, since $v = w_1 - w_2$, we get

$$\sum_{A \neq A_i, A_j} |v_A| \leq 4c_0 \widetilde{N}\sigma. \tag{G.29}$$

Now put

$$w_3 = \overline{f}\begin{pmatrix} \alpha \succ_1 \\ \beta \succ_3 \\ \gamma\ \phi \end{pmatrix} - \overline{f}\begin{pmatrix} \alpha \succ_1 \\ \beta \succ_4 \\ \gamma\ \phi \end{pmatrix},$$

$$w_4 = \overline{f}\begin{pmatrix} \alpha \succ_2 \\ \beta \succ_3 \\ \gamma\ \phi \end{pmatrix} - \overline{f}\begin{pmatrix} \alpha \succ_2 \\ \beta \succ_4 \\ \gamma\ \phi \end{pmatrix}.$$

Using $v = w_3 - w_4$, analogous computations give

$$\sum_{A \neq A_k, A_l} |v_A| \leq 4c_0 \widetilde{N}\sigma. \tag{G.30}$$

Now if $\{A_i, A_j\}$ is disjoint from $\{A_k, A_l\}$, then (G.29) and (G.30) immediately lead us to $\sum_{A \in C} |v_A| \leq 8c_0 \widetilde{N}\sigma$ which is stronger than (4.7). Otherwise, $\{A_i, A_j\}$ and $\{A_k, A_l\}$ have one element in common — say $A_i$ — in which case (G.29) and (G.30) give $\sum_{A \neq A_i} |v_A| \leq 8c_0 \widetilde{N}\sigma$. Since the sum of the components of $v$ is zero, we also have $|v_{A_i}| \leq 8c_0 \widetilde{N}\sigma$, and (4.7) follows. □

We now prove the three main lemmas that combine to give the theorem.

**Proof of Lemma 4.12:** Suppose the conclusion does not hold. Then the same reasoning as in case (i) of Theorem 4.7 gives a distribution $\phi$ such that

$$\Pr_{IID(\phi)}(f(CAB, P) = C) - \Pr_{IID(\phi)}(f(CBA, P) = C) \geq \frac{c_1}{N}.$$

If we consider a manipulator with true preference $CBA$, manipulating to $CAB$, with

top-set $\mathcal{C}^+ = \{C\}$, then this gives us $\sigma \geq c_1/N$, contradicting the given. $\square$

**Proof of Lemma 4.13:** Define the following vectors in $\mathbb{R}^M$:

$$v_1 = \overline{f}\begin{pmatrix} x\ ABC\ldots \\ y\ BAC\ldots \\ z\ BAC\ldots \end{pmatrix} - \overline{f}\begin{pmatrix} x\ ABC\ldots \\ y\ ABC\ldots \\ z\ BAC\ldots \end{pmatrix}$$

(we write e.g. $(x\ ABC\ldots, y\ BAC\ldots, z\ BAC\ldots)$ rather than $(x\ ABC\ldots, y + z\ BAC\ldots)$ to aid readability; no confusion should result)

$$v_2 = \overline{f}\begin{pmatrix} x\ ACB\ldots \\ y\ BAC\ldots \\ z\ BAC\ldots \end{pmatrix} - \overline{f}\begin{pmatrix} x\ ACB\ldots \\ y\ ABC\ldots \\ z\ BAC\ldots \end{pmatrix}$$

$$v_3 = \overline{f}\begin{pmatrix} x\ CAB\ldots \\ y\ BAC\ldots \\ z\ BAC\ldots \end{pmatrix} - \overline{f}\begin{pmatrix} x\ CAB\ldots \\ y\ ABC\ldots \\ z\ BAC\ldots \end{pmatrix}$$

$$v_4 = \overline{f}\begin{pmatrix} x\ CAB\ldots \\ y\ BAC\ldots \\ z\ BCA\ldots \end{pmatrix} - \overline{f}\begin{pmatrix} x\ CAB\ldots \\ y\ ABC\ldots \\ z\ BCA\ldots \end{pmatrix}$$

$$v_5 = \overline{f}\begin{pmatrix} x\ CAB\ldots \\ y\ BAC\ldots \\ z\ CBA\ldots \end{pmatrix} - \overline{f}\begin{pmatrix} x\ CAB\ldots \\ y\ ABC\ldots \\ z\ CBA\ldots \end{pmatrix}$$

By applying Lemma 4.11 repeatedly, we get

$$|v_1 - v_2| \leq 16c_0\widetilde{N}\sigma; \qquad |v_2 - v_3| \leq 16c_0\widetilde{N}\sigma;$$

$$|v_3 - v_4| \leq 16c_0\widetilde{N}\sigma; \qquad |v_4 - v_5| \leq 16c_0\widetilde{N}\sigma.$$

Adding these and using the triangle inequality gives

$$|v_1 - v_5| \leq 64 c_0 \widetilde{N} \sigma.$$

Next, define

$$v_1' = \overline{f}\begin{pmatrix} x'\ ABC\ldots \\ y\ BAC\ldots \\ z'\ BAC\ldots \end{pmatrix} - \overline{f}\begin{pmatrix} x'\ ABC\ldots \\ y\ ABC\ldots \\ z'\ BAC\ldots \end{pmatrix}$$

$$v_5' = \overline{f}\begin{pmatrix} x'\ CAB\ldots \\ y\ BAC\ldots \\ z'\ CBA\ldots \end{pmatrix} - \overline{f}\begin{pmatrix} x'\ CAB\ldots \\ y\ ABC\ldots \\ z'\ CBA\ldots \end{pmatrix}.$$

Then the above reasoning also gives

$$|v_1' - v_5'| \leq 64 c_0 \widetilde{N} \sigma,$$

and hence we obtain

$$|(v_1 - v_1') - (v_5 - v_5')| \leq 128 c_0 \widetilde{N} \sigma. \tag{G.31}$$

Now define

$$w_1 = \overline{f}\begin{pmatrix} x\ CAB\ldots \\ y\ BAC\ldots \\ z\ CBA\ldots \end{pmatrix} - \overline{f}\begin{pmatrix} x'\ CAB\ldots \\ y\ BAC\ldots \\ z'\ CBA\ldots \end{pmatrix}$$

$$w_2 = \overline{f}\begin{pmatrix} x\ CAB\ldots \\ y\ BCA\ldots \\ z\ CBA\ldots \end{pmatrix} - \overline{f}\begin{pmatrix} x'\ CAB\ldots \\ y\ BCA\ldots \\ z'\ CBA\ldots \end{pmatrix}$$

$$w_3 = \overline{f}\begin{pmatrix} x\ CAB\ldots \\ y\ CBA\ldots \\ z\ CBA\ldots \end{pmatrix} - \overline{f}\begin{pmatrix} x'\ CAB\ldots \\ y\ CBA\ldots \\ z'\ CBA\ldots \end{pmatrix}.$$

Then Lemma 4.11 gives

$$|w_1 - w_2| \leq 16c_0\widetilde{N}\sigma; \qquad |w_2 - w_3| \leq 16c_0\widetilde{N}\sigma,$$

so by the triangle inequality,

$$|w_1 - w_3| \leq 32c_0\widetilde{N}\sigma.$$

However, $w_3 = 0$, because our assumption (4.8) implies that both $\overline{f}(\cdots)$ values in the definition of $w_3$ are just $C$ with probability 1. Thus we actually have

$$|w_1| \leq 32c_0\widetilde{N}\sigma. \tag{G.32}$$

Similarly define

$$w_4 = \overline{f}\begin{pmatrix} x\,CAB\ldots \\ y\,ABC\ldots \\ z\,CBA\ldots \end{pmatrix} - \overline{f}\begin{pmatrix} x'\,CAB\ldots \\ y\,ABC\ldots \\ z'\,CBA\ldots \end{pmatrix}$$

$$w_5 = \overline{f}\begin{pmatrix} x\,CAB\ldots \\ y\,ACB\ldots \\ z\,CBA\ldots \end{pmatrix} - \overline{f}\begin{pmatrix} x'\,CAB\ldots \\ y\,ACB\ldots \\ z'\,CBA\ldots \end{pmatrix}$$

$$w_6 = \overline{f}\begin{pmatrix} x\,CAB\ldots \\ y\,CAB\ldots \\ z\,CBA\ldots \end{pmatrix} - \overline{f}\begin{pmatrix} x'\,CAB\ldots \\ y\,CAB\ldots \\ z'\,CBA\ldots \end{pmatrix}.$$

Lemma 4.11 gives

$$|w_4 - w_5| \leq 16c_0\widetilde{N}\sigma,$$

$$|w_5 - w_6| \leq 16c_0\widetilde{N}\sigma,$$

and as before we actually have $w_6 = 0$, so we conclude

$$|w_4| \leq 32c_0\widetilde{N}\sigma. \tag{G.33}$$

Notice now that $v_5 - v_5' = w_1 - w_4$, so (G.32) and (G.33) give us

$$|v_5 - v_5'| \leq 64c_0\widetilde{N}\sigma,$$

and combining this with (G.31) we obtain

$$|v_1 - v_1'| \leq 192c_0\widetilde{N}\sigma. \tag{G.34}$$

This is exactly what we sought to prove. □

**Proof of Lemma 4.14:** We proceed by considering the behavior of $f$ near the endpoints of the $ABC\ldots - BAC\ldots$ edge, showing that $\overline{f}$ cannot be very close to linearity.

Given $f$, let $M$ denote the supremum of the left-hand side of (4.11), over all choices of $x, y, z, x', z'$. We consider two cases.

(i) There is some $K \leq \sqrt{\widetilde{N}}/2$ such that $f(K\ BAC\ldots, \widetilde{N} - K\ ABC\ldots) \neq A$. In this case, as long as $N$ is sufficiently large, we have

$$\begin{aligned}
\overline{f}_A\left(\begin{array}{c} 1 - K/\widetilde{N}\ ABC\ldots \\ K/\widetilde{N}\ BAC\ldots \end{array}\right) &\leq 1 - \mathbf{P}\left(\begin{array}{c|c} \widetilde{N} - K & 1 - K/\widetilde{N} \\ K & K/\widetilde{N} \end{array}\ \middle|\ \widetilde{N};\right) \\
&\leq 1 - \sigma_N^* \\
&< 1 - \frac{1}{\sqrt{2\widetilde{N}}}
\end{aligned}$$

by Lemma 2.4 and the asymptotic behavior of $\sigma_N^*$. Therefore by taking $x = 1 - K/\widetilde{N}$, $y = K/\widetilde{N}$, $z = 0$, and noting $f(x + y\ ABC\ldots, z\ BAC\ldots) = A$ by weak unanimity, we get

$$(v_1)_A \leq -\frac{1}{\sqrt{2\widetilde{N}}}.$$

It follows that for any choices of $x', z' \geq 0$ with $x' + z' = 1 - K/\widetilde{N}$,

$$(v_1')_A \leq -\frac{1}{\sqrt{2\widetilde{N}}} + M.$$

In particular, for any positive integer $r \leq \lfloor 2\sqrt{\widetilde{N}} \rfloor$, we may take $x' = 1 - rK/\widetilde{N}, z = (r-1)K\widetilde{N}$ to obtain

$$\overline{f}_A \left( \begin{array}{c} (1 - rK/\widetilde{N})\ ABC \ldots \\ rK/\widetilde{N}\ BAC \ldots \end{array} \right) - \overline{f}_A \left( \begin{array}{c} (1 - (r-1)K/\widetilde{N})\ ABC \ldots \\ (r-1)K/\widetilde{N}\ BAC \ldots \end{array} \right) \leq -\frac{1}{\sqrt{2\widetilde{N}}} + M.$$

If we apply this inequality for $r = 1, 2, \ldots, \overline{r} = \lfloor 2\sqrt{\widetilde{N}} \rfloor$ and telescope, we obtain

$$\overline{f}_A \left( \begin{array}{c} (1 - \overline{r}K/\widetilde{N})\ ABC \ldots \\ \overline{r}K/\widetilde{N}\ BAC \ldots \end{array} \right) - \overline{f}_A \left( \begin{array}{c} 1\ ABC \ldots \\ 0\ BAC \ldots \end{array} \right) \leq \overline{r} \left( -\frac{1}{\sqrt{2\widetilde{N}}} + M \right).$$

The left side cannot be lower than $0 - 1 = -1$, so

$$-1 \leq \overline{r} \left( -\frac{1}{\sqrt{2\widetilde{N}}} + M \right)$$

which leads to

$$M \geq \frac{1}{\sqrt{2\widetilde{N}}} - \frac{1}{\overline{r}} \sim \left( \frac{\sqrt{2}-1}{2} \right) \cdot \frac{1}{\sqrt{\widetilde{N}}}.$$

(ii) For all $K \leq \sqrt{\widetilde{N}}/2$, $f(K\ BAC \ldots, \widetilde{N} - K\ ABC \ldots) = A$. Then apply Lemma C.9 with $c = 1/6$ to conclude that if an $\widetilde{N}$-profile $P$ is drawn $IID(\alpha\ BAC \ldots, 1 - \alpha\ ABC \ldots)$ for any $\alpha \leq 1/6\sqrt{\widetilde{N}}$, then the probability that $f(P) \neq A$ is at most $1/\widetilde{N}$, as long as $N$ is sufficiently large.

Let $s$ be an integer with $6\sqrt{\widetilde{N}} < s < 7\sqrt{\widetilde{N}}$. Then taking $x = 1 - 1/s$, $y = 1/s$, $z = 0$, and again using $f(x + y\ ABC \ldots, z\ BAC \ldots) = A$ by weak unanimity,

we get

$$(v_1)_A \geq -\frac{1}{\widetilde{N}}.$$

So for any choices of $x', z' \geq 0$ with $x' + z' = 1 - 1/s$, we have

$$(v_1')_A \geq -\frac{1}{\widetilde{N}} - M.$$

In particular, for any $r = 0, \ldots, s-1$, we can take $x' = (s-1-r)/s$ and $z' = r/s$ to obtain

$$\overline{f}_A\left(\begin{array}{c} 1-(r+1)/s\ ABC\ldots \\ (r+1)/s\ BAC\ldots \end{array}\right) - \overline{f}_A\left(\begin{array}{c} 1-r/s\ ABC\ldots \\ r/s\ BAC\ldots \end{array}\right) \geq -\frac{1}{\widetilde{N}} - M.$$

Summing for $r = 0, \ldots, s-1$ and telescoping gives

$$\overline{f}_A\left(\begin{array}{c} 0\ ABC\ldots \\ 1\ BAC\ldots \end{array}\right) - \overline{f}_A\left(\begin{array}{c} 1\ ABC\ldots \\ 0\ BAC\ldots \end{array}\right) \geq s\left(-\frac{1}{\widetilde{N}} - M\right).$$

Using weak unanimity, the left side equals $0 - 1 = -1$, so

$$-1 \geq s\left(-\frac{1}{\widetilde{N}} - M\right)$$

from which

$$M \geq \frac{1}{s} - \frac{1}{\widetilde{N}} \gtrsim \frac{1}{7\sqrt{\widetilde{N}}}.$$

In both cases (i) and (ii), we showed that $M$ was bounded below by a function of $\widetilde{N}$ that is asymptotically equal to a constant times $1/\sqrt{\widetilde{N}}$, which is exactly what the lemma claims.

□

Now we give the proof of Theorem 4.6. Essentially, we just need to replace Lemma 4.14 with a corresponding statement giving a sharper bound for simple rules:

**Lemma G.1** *There exists some absolute constant $c_3$, independent of $N$, with the following property: As long as $N$ is large enough, for any $f$ that is simple over $A$ and*

*B, there exist some nonnegative* $x, y, z, x', z'$ *with*

$$\left| \left( \overline{f} \begin{pmatrix} x\ ABC\ldots \\ y+z\ BAC\ldots \end{pmatrix} - \overline{f} \begin{pmatrix} x+y\ ABC\ldots \\ z\ BAC\ldots \end{pmatrix} \right) - \left( \overline{f} \begin{pmatrix} x'\ ABC\ldots \\ y+z'\ BAC\ldots \end{pmatrix} - \overline{f} \begin{pmatrix} x'+y\ ABC\ldots \\ z'\ BAC\ldots \end{pmatrix} \right) \right| \geq c_3. \tag{G.35}$$

**Proof:** Let $K^*$ be the threshold such that $f(K\ ABC\ldots, \widetilde{N} - K\ BAC\ldots) = A$ iff $K \geq K^*$. Just as in the proof of Theorem 4.2, assume that $K^* \leq \widetilde{N}/2$ (otherwise switch $A$ and $B$), and put

$$\phi_1 = (\alpha_1\ ABC\ldots, 1-\alpha_1\ BAC\ldots) \qquad \text{with} \qquad \alpha_1 = \frac{K^* + \sqrt{2\widetilde{N}}}{\widetilde{N}},$$

$$\phi_2 = (\alpha_2\ ABC\ldots, 1-\alpha_2\ BAC\ldots) \qquad \text{with} \qquad \alpha_2 = \max\left\{ \frac{K^* - \sqrt{2\widetilde{N}}}{\widetilde{N}}, 0 \right\}.$$

By simplicity, $\overline{f}(\phi_1), \overline{f}(\phi_2)$ both put positive weight only on $A$ and $B$, and by the same Chebyshev argument as in Theorem 4.2, $\overline{f}(\phi_1)$ puts probability at least 7/8 on $A$, while $\overline{f}(\phi_2)$ puts probability at most 1/8 on $A$.

Next put

$$\phi_3 = (\alpha_3\ ABC\ldots, 1-\alpha_3\ BAC\ldots) \qquad \text{with} \qquad \alpha_3 = 2\alpha_1 - \alpha_2.$$

Since $\alpha_3 > \alpha_1$, the same Chebyshev argument gives that $\overline{f}(\phi_3)$ puts probability at least 7/8 on $A$ (and the remaining probability on $B$). We now have

$$|\overline{f}(\phi_1) - \overline{f}(\phi_2)| \geq 3/2,$$

$$|\overline{f}(\phi_3) - \overline{f}(\phi_1)| \leq 1/4.$$

Now take

$$x = \alpha_1, \qquad y = \alpha_2 - \alpha_1, \qquad z = 1 - \alpha_2,$$

$$x' = \alpha_3, \qquad z' = 1 - \alpha_1.$$

The expression on the left side of (G.35) reduces to

$$|(\overline{f}(\phi_1) - \overline{f}(\phi_2)) - (\overline{f}(\phi_3) - \overline{f}(\phi_1))| \geq \frac{3}{2} - \frac{1}{4} = \frac{5}{4}$$

which proves the lemma. □

**Proof of Theorem 4.6:** As usual, it suffices to assume $N$ is large enough so that Lemma G.1 applies. Assume $A, B, C$ are chosen so that $f$ is simple over $A$ and $B$. Let $c_0, c_1, c_3$ be as in Lemmas 4.12, 4.13, G.1. Either $\sigma \geq c_1/N$, and we are done; or else Lemma 4.12 applies, in which case (4.10) and (by simplicity) (G.35) apply; combining these gives $\sigma \geq c_3/192c_0\widetilde{N}$. □

# H Construction for quickly-decaying susceptibility

We provide here the construction of a tops-only voting rule that attains susceptibility on the order of $N^{-\kappa}$ with $\kappa > 1/2$, as required by Theorem 4.3. The actual construction is more elaborate than the approximate random dictatorship sketched in the main paper, so we first give a more detailed overview.

The main idea behind the construction is to subdivide the simplex of vote profiles into *blocks* as illustrated in Figure H.1. Within each block, we then assign winners $A_1, \ldots, A_M$ to the various profiles, in proportions that correspond to the position of the block in the vote simplex.

More specifically, in order to avoid creating especially large opportunities for manipulation near the edge of the vote simplex, we need to focus on *viable* candidates at each vote profile, as in the construction of the pair-or-plurality system in Subsection 3.3. Roughly speaking, each candidate needs to get more than some threshold number of votes to be considered viable; the threshold will be taken to be (asymptotically) some constant $\lambda$ times $N$. Then, for each set $\mathcal{C}'$ of candidates, we consider the

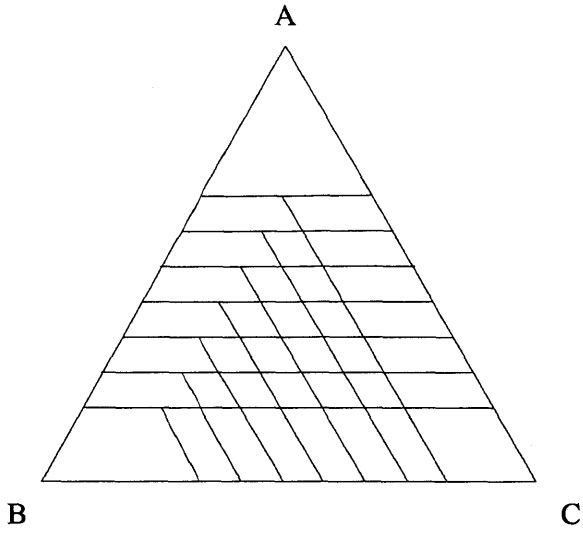


Figure H.1: Sketch of the construction for Theorem 4.3

subspace of vote profiles in which the viable candidates are precisely the members of $C'$, and carve up this set of profiles into blocks, depending on how many votes each viable candidate receives. All blocks have equal size $S$ along each of the dimensions corresponding to a viable candidate.

For any given block, we define a weight for each viable candidate by subtracting $\lambda N$ from her vote total. We then assign each viable candidate to some profiles within the block, so that the fraction of profiles assigned to a given candidate is approximately proportional to her weight. Within the block, we use Lemma C.11 to determine exactly which profiles are assigned to each candidate so that the difference between a candidate's relative probability and her weight is kept small.

Consider now the susceptibility of a voting rule defined in this way, with blocks of size $S$. When the manipulator changes his vote, this affects the distribution over realized vote profiles in two ways: it changes the distribution over blocks, and it changes the distribution over profiles within each block. By considering the distribution within each block, we show that the distribution over winning candidates equals the distribution is pinned down by the distribution over blocks to within order $S^d N^{-(d-2)/2}$ (ignoring constant factors). Here $d$ is the value used in applying Lemma C.11. We also show that the change across blocks affects the distribution over candidates on the

order of $S^{-1/d}N^{-1/2}$. Hence, our construction gives an upper bound for susceptibility that is approximately on the same order as $\max\{S^dN^{-(d-2)/2}, S^{-1/d}N^{-1/2}\}$. In order to achieve the fastest possible rate of decline in susceptibility as $N \to \infty$, we choose $d = 6$ and $S \approx N^{9/37}$, with the resulting rate of decline $N^{-20/37}$. We will henceforth use these numbers for concreteness.[9]

**Proof of Theorem 4.3:** We first give the exact construction of the voting system. Fix constants $\lambda, \mu$ with $0 < \lambda < 1/M$ and $0 < \mu < 1 - M\lambda$. Also fix $\underline{\alpha}$ with $0 < \underline{\alpha} < \min\{\lambda/3, \mu/3, (1 - M\lambda - \mu)/(M + 1)\}$.

For each value of $N$, choose integers $S_N, L_N, R_N$ such that

- $S_N = 2^{6h}$ for some integer $h$ and $N^{9/37} \lesssim S_N \lesssim 2^6 \cdot N^{9/37}$;
- $L_N \sim \lambda N$;
- $R_N S_N \sim \mu N$.

We will henceforth refer to these as $S, L, R$, with the dependence on $N$ implicit. Verbally, we refer to them as the *block size*, *viability lower bound*, and *number of blocks* (in each dimension).

A *block label* is a sequence consisting of $M-1$ or fewer (possibly zero) nonnegative integers, whose sum is at most $R$.

Given a profile $P = (x_1\ A_1, \ldots, x_M\ A_M)$, we compute a corresponding block label $BL(P)$ by the following algorithm:

1. For each $i = 1, \ldots, M$, if $x_i < L$, put $\Lambda_i = 0$. Otherwise, put $\Lambda_i = \lfloor (x_i - L)/S \rfloor + 1$.

2. Let $t$ be the smallest index such that $\Lambda_1 + \ldots + \Lambda_t > R$. Notice that $t$ must exist, since

$$\frac{\Lambda_1 + \cdots + \Lambda_M}{R} \geq \frac{x_1 + \cdots + x_M - ML}{RS} \sim \frac{(1 - M\lambda)N}{RS} \sim \frac{1 - M\lambda}{\mu} > 1.$$

[9]It is possible to achieve faster rates of convergence through minor improvements on the construction, but we do not bother doing so here, since we have not found a construction showing that the exponent $-1$ in Theorem 4.2 is tight.

Then define $BL(P)$ to be the $(t-1)$-term sequence $(\Lambda_1, \ldots, \Lambda_{t-1})$.

Given a block label $\Lambda$, we define the corresponding *block* as $BL^{-1}(\Lambda)$, obtained by inverting the above procedure. For each candidate $A_i$ we construct a lower bound $\underline{x}_i$ and an upper bound $\overline{x}_i$ on the number of votes: If $\Lambda = (\Lambda_1, \ldots, \Lambda_{t-1})$, then

- if $i \leq t-1$ and $\Lambda_i = 0$, put $\underline{x}_i = 0$ and $\overline{x}_i = L - 1$;
- if $i \leq t-1$ and $\Lambda_i > 0$, put $\underline{x}_i = S(\Lambda_i - 1) + L$ and $\overline{x}_i = S\Lambda_i + L - 1$;
- for $i = t$, put $\underline{x}_i = S(R - \sum_j \Lambda_j) + L$ and $\overline{x}_i = N + 1$;
- for $i > t$, put $\underline{x}_i = 0$ and $\overline{x}_i = N + 1$.

Then one readily checks that $BL^{-1}(\Lambda)$ is the set of all $(N+1)$-profiles of votes $(x_1, \ldots, x_M)$ such that $\underline{x}_i \leq x_i \leq \overline{x}_i$ for all $i$.

We also define *weights* $W_i(\Lambda)$, for each block label $\Lambda$ and each candidate $A_i$: if $\Lambda = (\Lambda_1, \ldots, \Lambda_{t-1})$ then

- for $i \leq t-1$, $W_i(\Lambda) = \Lambda_i/(R+1)$;
- for $i = t$, $W_i(\Lambda) = 1 - \sum_j \Lambda_j/(R+1)$;
- for $i > t$, $W_i(\Lambda) = 0$.

Thus we always have $\sum_i W_i(\Lambda) = 1$.

We further modify these weights by rounding down to integer multiples of $1/2^h$: for each $i < M$, define the *rounded weight* $\widetilde{W}_i(\Lambda) = \lfloor 2^h W_i(\Lambda) \rfloor / 2^h$, and put $\widetilde{W}_M(\Lambda) = 1 - \sum_{i=1}^{M-1} \widetilde{W}_i(\Lambda)$.

Let $S = 2^{6h}$, and let $Z = \{0, 1, \ldots, 2^{6h} - 1\}$ be partitioned into $2^h$ subsets $Z_0, \ldots, Z_{2^h-1}$ according to Lemma C.11. For each block label $\Lambda$, we let $g_\Lambda$ be any function from $\{0, 1, \ldots, 2^h - 1\} \to \mathcal{C}$ such that $|g_\Lambda^{-1}(A_i)| = 2^h \widetilde{W}_i(\Lambda)$ for each candidate $A_i$. Thus, the proportion of values of $y$ on which $g_\Lambda$ takes the value $A_i$ equals the rounded weight of $A_i$.

Finally, we are ready to define the voting rule $f$. Given a profile of votes, $P = (x_1\ A_1, \ldots, x_M\ A_M)$, we define $f(P)$ as follows:

- Let $(\Lambda_1, \ldots, \Lambda_{t-1}) = \Lambda = BL(P)$ be the block label.

- If every term $\Lambda_i$ is zero, then let $f(P) = A_t$.

- Otherwise, consider the smallest $i$ such that $\Lambda_i > 0$. Let

$$\widehat{x}_i = (x_i - L) - S\left\lfloor \frac{x_i - L}{S} \right\rfloor.$$

Then $\widehat{x}_i$ is an element of $Z$. So $\widehat{x}_i \in Z_y$ for exactly one $y$. Put $f(P) = g_\Lambda(y)$.

This defines the voting rule. The statement of Theorem 4.3 promised that it would be Pareto efficient and tops-only. Tops-onliness is clear from the construction, so we should check Pareto efficiency. Evidently we must check that $f(P)$ is always a candidate who gets at least one vote in profile $P$. If every term $\Lambda_i$ of the block label $BL(P)$ is zero, then $\Lambda_t > R > 1$ so that $x_t > 0$. Otherwise, notice that whenever $A_i$ is a candidate with $\Lambda_i = 0$, then $W_i(\Lambda) = 0$, and so $g_\Lambda(y) \neq A_i$ for all $y$. Consequently we cannot have $f(P) = A_i$ for any such $i$. Thus, $f(P)$ must be a candidate $A_i$ for whom $\Lambda_i > 0$, implying $x_i > 0$.

Our remaining task is to prove the susceptibility bound. The proof of the bound is based on two claims. Let $\epsilon$ be an arbitrary small positive constant.

**Claim I.** There is a constant $c_I$ such that the following holds. For all distributions $\phi \in \Delta(\mathcal{C})$, all candidates $A_i, A_j$,

$$\left| Pr_\phi(f(A_j, P) = A_i) - \sum_\Lambda Pr_\phi((A_j, P) \in BL^{-1}(\Lambda))\widetilde{W}_i(\Lambda) \right| < c_I N^{-(20/37-\epsilon)}.$$

Here the $Pr_\phi(\cdots)$ expressions refer to probabilities concerning the profile $(A_j, P)$, given that $P$ is formed by having each of the $N$ other voters drawn independently from $\phi$.

**Claim II.** There is a constant $c_{II}$ such that the following holds. For all distribu-

tions $\phi \in \Delta(\mathcal{C})$, and all candidates $A_i, A_j, A_k$,

$$\left| \sum_{\Lambda} Pr_{\phi}((A_j, P) \in BL^{-1}(\Lambda))\widetilde{W}_i(\Lambda) - \sum_{\Lambda} Pr_{\phi}((A_k, P) \in BL^{-1}(\Lambda))\widetilde{W}_i(\Lambda) \right| < c_{II} N^{-(20/37-\epsilon)}. \quad \text{(H.1)}$$

We shall prove these two claims, then show how this quickly completes the proof of the theorem.

**Proof of Claim I.** We rewrite the expression inside the absolute value as

$$\sum_{\Lambda} \left[ Pr_{\phi}((A_j, P) \in BL^{-1}(\Lambda) \text{ and } f(A_j, P) = A_i) - Pr_{\phi}((A_j, P) \in BL^{-1}(\Lambda)\widetilde{W}_i(\Lambda)) \right].$$

For each block label $\Lambda$ consisting of zeroes, the relevant difference is zero. (If $t-1$ is the length of $\Lambda$, then $W_t(\Lambda) = 1$, while $w_i(\Lambda) = 0$ for $i \neq t$; and $f$ takes the value $A_t$ throughout $BL^{-1}(\Lambda)$.) So we can restrict to the sum over $\Lambda$ having a nonzero component.

For each candidate $A_k$, $k < M$, let $\Theta_{k,t}$ be the set of all block labels $\Lambda$ with length $t-1$ such that $\Lambda_l = 0$ for all $l < k$, but $\Lambda_k > 0$. It suffices to show that there is a constant $c'$, independent of $\phi$, such that

$$\left| \sum_{\Lambda \in \Theta_{k,t}} \left[ Pr_{\phi}((A_j, P) \in BL^{-1}(\Lambda) \cap f^{-1}(A_i)) - Pr_{\phi}((A_j, P) \in BL^{-1}(\Lambda))\widetilde{W}_i(\Lambda) \right] \right| < c' N^{-(20/37-\epsilon)}. \quad \text{(H.2)}$$

First consider any distribution $\phi$ such that $\phi_t < \underline{\alpha}$. If $P \sim IID(\phi)$, then the number of votes received by candidate $A_t$ in $P$ has expectation $\phi_t N \leq \underline{\alpha} N$ and variance $\phi_t(1-\phi_t)N \leq \underline{\alpha} N$, so by Chebyshev, the probability that $A_t$'s vote count is at least $2\underline{\alpha} N$ is $\leq 1/\underline{\alpha} N$. Consequently, the probability that $(A_j, P)$ gives $A_t$ at least $2\underline{\alpha} N + 1$ votes is $\leq M/\underline{\alpha} N$. Notice that at every profile in any block $\Lambda \in \Theta_{k,t}$, we

must have $\Lambda_t \geq 1$ from which $x_t \geq L > 2\underline{\alpha}N + 1$. Thus

$$\sum_{\Lambda \in \Theta_{k,t}} Pr_\phi((A_j, P) \in BL^{-1}(\Lambda)) \leq M/\underline{\alpha}N. \tag{H.3}$$

But both probabilities on the left side of (H.2) are bounded above by the sum in (H.3), hence (H.2) holds in this case (with the appropriate choice of $c'$).

Similarly, consider any distribution $\phi$ such that $\phi_k < \underline{\alpha}$. Because every block $\Lambda \in \Theta_{k,t}$ must have $\Lambda_k \geq 1$, from which any profile in such a block must have $x_k \geq L > 2\underline{\alpha}N + 1$, we can follow the same argument to show that (H.2) is satisfied again.

This means we can henceforth restrict to distributions $\phi$ such that

$$\phi_t \geq \underline{\alpha} \qquad \text{and} \qquad \phi_k \geq \underline{\alpha}.$$

For any $\phi$, let $l$ be the highest index such that $\phi_l \geq \underline{\alpha}$; thus $l \geq t$.

Consider any block $\Lambda = (\Lambda_1, \ldots, \Lambda_{t-1}) \in \Theta_{k,t}$. For each $s = 1, \ldots, M$, define bounds $\underline{x}_s, \overline{x}_s$ as in the computation of $BL^{-1}(\Lambda)$ above. Consider any given values $x_s$, with $\underline{x}_s \leq x_s \leq \overline{x}_s$, for each $s \neq k, l$; write $x_{-kl}$ for the vector of such values. Define $[x_{-kl}]$ to be the set of all profiles having the specified number of votes for each candidate $A_s$, $s \neq k, l$.

We further break down the left-hand side of (H.2) by summing over different values of $x_{-kl}$. Define notations

$$\Pi_1(\Lambda, x_{-kl}) = Pr_\phi\left((A_j, P) \in BL^{-1}(\Lambda) \cap [x_{-kl}] \cap f^{-1}(A_i)\right),$$

$$\Pi_2(\Lambda, x_{-kl}) = Pr_\phi\left((A_j, P) \in BL^{-1}(\Lambda) \cap [x_{-kl}]\right) \cdot \widetilde{W}_i(\Lambda).$$

Then in the left-hand side of (H.2), the first expression is $\sum_{x_{-kl}} \Pi_1(\Lambda, x_{-kl})$ (where the sum is over all possible vectors $x_{-kl}$), and the second expression is $\sum_{x_{-kl}} \Pi_2(\lambda, x_{-kl})$.

Thus, (H.2) is equivalent to

$$\left| \sum_{\substack{\Lambda \in \Theta_{k,t} \\ x_{-kl}}} [\Pi_1(\Lambda, x_{-kl}) - \Pi_2(\Lambda, x_{-kl})] \right| < c' N^{-(20/37-\epsilon)}. \tag{H.4}$$

We prove this by breaking into several cases depending on the choice of $\Lambda$ and $x_{-kl}$. We first deal with cases that have low probability, so that their contribution to the sums of $\Pi_1, \Pi_2$ in (H.4) is small; and then we can deal with the substantive case where we actually make use of the elaborate construction behind $f$ within each block.

(i) First, for each $s \neq k, l$, consider choices of $x_{-kl}$ that have $|x_s - \phi_s N| > \underline{\alpha} N/M$. The probability that $(A_j, P)$ gives candidate $A_s$ such a number of votes is at most the probability that $P$ gives $A_s$ a number of votes within $\underline{\alpha} N/2M$ of $\phi_s N$. By the usual Chebyshev argument, this probability is $\leq 4M^2/\underline{\alpha} N$.

Since $\sum_\Lambda \Pi_1(\Lambda, x_{-kl}) \leq Pr_\phi((A_j, P) \in [x_{-kl}])$, the sum of $\Pi_1(\Lambda, x_{-kl})$ over all $\Lambda$ and all $x_{-kl}$ with $|x_s - \phi_s N| > \underline{\alpha} N/M$ is at most $4M^2/\underline{\alpha} N$. Similarly, the same holds for $\Pi_2$. Thus, all the pairs $(\Lambda, x_{-kl})$ for which $|x_s - \phi_s N| > \underline{\alpha} N/M$ make a total contribution to the left side of (H.4) that is bounded above by a constant times $N^{-1}$.

(ii) Next, consider choices of $\Lambda$ that have $|(S(\Lambda_k - 1) + L) - \phi_k N| > \underline{\alpha} N/M$. If $(A_j, P)$ is in such a block $BL^{-1}(\Lambda)$, then the number of votes for candidate $A_k$ is between $\underline{x}_k = S(\Lambda_k - 1) + L$ and $\overline{x}_k = S\Lambda_k + L - 1$. For $N$ sufficiently large, this means that the number of votes for $A_k$ in $P$ is more than $(\underline{\alpha}/2M)N$ away from $\phi_k N$. Again, this occurs with probability $\leq 4M^2/\underline{\alpha} N$.

Since $\sum_{x_{-kl}} \Pi_1(\Lambda, x_{-kl}) \leq Pr_\phi((A_j, P) \in BL^{-1}(\Lambda))$, and similarly for $\Pi_2$, the pairs $(\Lambda, x_{-kl})$ for which $|(S(\Lambda_k - 1) + L) - \phi_k N| > \underline{\alpha} N/M$ make a total contribution to the left side of (H.4) that is bounded above by a constant times $N^{-1}$.

From this and the previous bullet point, we see that in proving (H.4) it suffices

to restrict attention to pairs $(\Lambda, x_{-kl})$ for which

$$|x_s - \phi_s N| \leq \underline{\alpha} N/M \qquad \text{for all } s \neq k, l; \tag{H.5}$$

$$|(S(\Lambda_k - 1) + L) - \phi_k N| \leq \underline{\alpha} N/M. \tag{H.6}$$

That is, the contribution of all other pairs to the sum in (H.4) is negligible.

(iii) We will show that (H.5) and (H.6) imply

$$|\Pi_1(\Lambda, x_{-kl}) - \Pi_2(\Lambda, x_{-kl})| \leq c'' N^{-(48/37-\epsilon)} Pr_\phi((A_j, P) \in [x_{-kl}]) \tag{H.7}$$

where $c''$ is a constant not depending on $\phi$, $N$, or $x_{-kl}$.

We proceed by first showing that $BL^{-1}(\Lambda) \cap [x_{-kl}]$ contains exactly $\overline{x}_k - \underline{x}_k + 1$ profiles. That is, for every choice of $x_k$ with $\underline{x}_k \leq x_k \leq \overline{x}_k$, there is exactly one choice of $x_l$ such that the profile $(x_k, x_l, x_{-kl})$ is in $BL^{-1}(\Lambda) \cap [x_{-kl}]$. The relevant choice of $x_l$ would of course be $x_l = x_{k+l} - x_k$, where $x_{k+l} = N + 1 - \sum_{s \neq k,l} x_s$, so we just need to check that this value of $x_l$ always lies between the bounds $\underline{x}_l$ and $\overline{x}_l$.

There are two cases for the lower bound:

- If $l > t$, then $\underline{x}_l = 0$. We have

$$\begin{aligned}
x_l &= N + 1 - \sum_{s \neq k,l} x_s - x_k \\
&\geq N + 1 - \sum_{s \neq k,l} (\phi_s N + \underline{\alpha} N/M) - \overline{x}_k \\
&\geq (\phi_k + \phi_l) N - (M - 2)\underline{\alpha} N/M - S\Lambda_k - L \\
&\geq (\phi_k + \phi_l) N - (M - 2)\underline{\alpha} N/M - (S - L + \phi_k N + \underline{\alpha}/MN) - L \\
&= \phi_l N - (M - 1)\underline{\alpha} N/M - S \\
&\geq \underline{\alpha} N/M - S \\
&\geq 0
\end{aligned}$$

as long as $N$ is sufficiently large.

- If $l = t$, then $\underline{x}_l = S(R - \sum_{s<t} \Lambda_s) + L$. We also know, by definition of $l$, that $\phi_s < \underline{\alpha}$ for each $s > t$, so (H.5) implies $x_s \leq \underline{\alpha}(1 + 1/M)N$ for each such $s$.

  Let $\nu$ be the constant $1 - \underline{\alpha}(M+1) - (M\lambda + \mu) > 0$. We have

$$\begin{aligned}
x_l &= N + 1 - \sum_{s \neq t} x_s \\
&\geq N + 1 - \sum_{s<t} \overline{x}_s - \sum_{s>t} \underline{\alpha}(1 + 1/M)N \\
&\geq N + 1 - \sum_{s<t} (S\Lambda_s + L - 1) - \underline{\alpha}(M+1)N \\
&\geq N(1 - \underline{\alpha}(M+1)) - S\sum_{s<t} \Lambda_s - (M-1)L \\
&= (M\lambda + \mu + \nu)N - S\sum_{s<t} \Lambda_s - ML + L \\
&\geq ML + RS - S\sum_{s<t} \Lambda_s - ML + L \\
&\qquad\qquad \text{(when } N \text{ is sufficiently large)} \\
&= S(R - \sum_{s<t} \Lambda_s) + L.
\end{aligned}$$

  Thus the lower bound is satisfied when $l = t$ as well.

As for the upper bound, in both cases, $\overline{x}_l = N + 1$. Then

$$x_l = N + 1 - \sum_{s \neq k,l} x_s - x_k \leq N + 1$$

so the upper bound is always satisfied.

Thus $BL^{-1}(\Lambda) \cap [x_{-kl}]$ contains exactly $\overline{x}_k - \underline{x}_k + 1 = S$ profiles.

For each profile $(x_k, x_l, x_{-kl})$, we will explicitly write out the probability of this profile being the realized value of $(A_j, P)$. Specifically, let $(x_k^-, x_l^-, x_{-kl}^-)$ be identical to $(x_k, x_l, x_{-kl})$ except that the $j$-component has been decreased by 1. (There is no more succinct way to write this without breaking into cases

dependinding whether $j = k$, $j = l$, or neither.) Likewise put $x^-_{k+l} = x^-_k + x^-_l$, and $\phi_{k+l} = \phi_k + \phi_l$ and $\phi_{-kl}$ for the vector of other components of $\phi$. Then the probability of achieving $(x_k, x_l, x_{-kl})$ is

$$\mathbf{P}\left(\begin{array}{c} x^-_k \\ x^-_l \\ x^-_{kl} \end{array} \middle| N;\ \phi\right) = \mathbf{P}\left(\begin{array}{c} x^-_{k+l} \\ x^-_{-kl} \end{array} \middle| N;\ \begin{array}{c} \phi_{k+l} \\ \phi^-_{-kl} \end{array}\right) \mathbf{P}\left(\begin{array}{c} x^-_k \\ x^-_l \end{array} \middle| x^-_{k+l};\ \begin{array}{c} \phi_k/\phi_{k+l} \\ \phi_l/\phi_{k+l} \end{array}\right)$$

by Lemma C.2.

For succinctness, write

$$\beta = \mathbf{P}\left(\begin{array}{c} x^-_{k+l} \\ x^-_{-kl} \end{array} \middle| N;\ \begin{array}{c} \phi_{k+l} \\ \phi^-_{-kl} \end{array}\right)$$

for the first factor, which is independent of $x_k$, and

$$\widehat{\mathbf{P}}(x^-_k) = \mathbf{P}\left(\begin{array}{c} x^-_k \\ x^-_{k+l} - x^-_k \end{array} \middle| x^-_{k+l};\ \begin{array}{c} \phi_k/\phi_{k+l} \\ \phi_l/\phi_{k+l} \end{array}\right)$$

for the second factor.

Let $\underline{x}^-_k = \underline{x}_k$ or $\underline{x}_k - 1$ (depending whether $j = k$ or $j \neq k$). Then the possible values of $x^-_k$ corresponding to profiles $(x_k, x_l, x_{-kl}) \in BL^{-1}(\Lambda) \cap [x_{-kl}]$ are exactly the numbers $\underline{x}^-_k + z$, for $z \in Z$. (Recall we defined $Z = \{0, 1, \ldots, S-1\}$.)

Now, we have $\phi_k/\phi_{k+l} \geq \phi_k \geq \underline{\alpha}$, and likewise $\phi_l/\phi_{k+l} \geq \underline{\alpha}$. We also have $x^-_{k+l} \geq \phi_{k+l}N - \underline{\alpha}N - 1$ (using (H.5)) $\geq \underline{\alpha}N - 1$, a lower bound that grows linearly in $N$.

Consequently, we can apply Lemma C.11, with $d = 6$. As long as $N$ is greater than some absolute threshold $N_0$, we have the inequality for any two values $y, y' \in \{0, 1, \ldots, 2^h - 1\}$:

$$\left| \sum_{z \in Z_y} \widehat{\mathbf{P}}(\underline{x}^-_k + z) - \sum_{z \in Z_{y'}} \widehat{\mathbf{P}}(\underline{x}^-_k + z) \right| \leq 2^{41h} h N^{-6\left(\frac{1}{2} - \frac{\epsilon}{18}\right)}. \tag{H.8}$$

This inequality is the key step in the proof of Claim I; it was for this reason that we needed to use the sets $Z_y$ in constructing $f$.

Since $h \leq \ln(N) \leq N^{\epsilon/3}$ for large $N$, and

$$2^{41h} = S^{41/6} \leq (constant) \cdot N^{123/74},$$

we can simplify the right side of (H.8) to give

$$\left| \sum_{z \in Z_y} \widehat{\mathbf{P}}(\underline{x}_k^- + z) - \sum_{z \in Z_{y'}} \widehat{\mathbf{P}}(\underline{x}_k^- + z) \right| \leq (constant) \cdot N^{-(\frac{99}{74} - \epsilon)}.$$

Next, sum over all choices of $y' \in \{0, 1, \ldots, 2^h - 1\}$ and use the triangle inequality. Since $2^h = S^{1/6} \gtrsim N^{3/74}$, we obtain

$$\left| 2^h \sum_{z \in Z_y} \widehat{\mathbf{P}}(\underline{x}_k^- + z) - \sum_{z=0}^{2^{6h}-1} \widehat{\mathbf{P}}(\underline{x}_k^- + z) \right| \leq (constant) \cdot N^{-(\frac{96}{74} - \epsilon)}.$$

Sum again over all $y$ with $g_\Lambda(y) = A_i$, and then also divide by $2^h$. The right-hand side has been multiplied by $\widetilde{W}_i(\Lambda)/2^h \leq 1$, and so we get

$$\left| \sum_{\substack{y \in g_\Lambda^{-1}(A_i) \\ z \in Z_y}} \widehat{\mathbf{P}}(\underline{x}_k^- + z) - \widetilde{W}_i(\Lambda) \sum_{z=0}^{2^{6h}-1} \widehat{\mathbf{P}}(\underline{x}_k^- + z) \right| \leq (constant) \cdot N^{-(\frac{48}{37} - \epsilon)} \quad \text{(H.9)}$$

(after simplifying the exponent on the right side).

Now we return to the definitions of $\Pi_1$ and $\Pi_2$. Notice that $\Pi_1(\Lambda, x_{-kl})$ is the sum of the probabilities of profiles $(x_k, x_l, x_{-kl}) \in BL^{-1}(\Lambda) \cap [x_{-kl}]$ on which $f$ takes the value $A_i$. Writing $\widehat{f}(x_k)$ for $f(x_k, x_{k+l} - k, x_{-kl})$, we have

$$\Pi_1(\Lambda, x_{-kl}) \;=\; \sum_{z : \widehat{f}(\underline{x}_k + z) = A_i} \beta \widehat{\mathbf{P}}(\underline{x}_k^- + z).$$

Moreover, by assumption $\Lambda_k > 0$, while $\Lambda_l = 0$ for all $l < k$. Therefore, the construction of $f$ on the block $BL^{-1}(\Lambda)$ implies that $\widehat{f}(\underline{x}_k + z) = A_i$ if and only if $z \in Z_y$ for some $y$ such that $g_\Lambda(y) = A_i$. That is,

$$\Pi_1(\Lambda, x_{-kl}) = \sum_{\substack{y \in g_\Lambda^{-1}(A_i) \\ z \in Z_y}} \beta \widehat{\mathbf{P}}(\underline{x}_k^- + z).$$

Meanwhile, $\Pi_2(\Lambda, x_{-kl})$ is the sum of the probabilities of all profiles in $BL^{-1}(\Lambda) \cap [x_{-kl}]$, regardless of the corresponding values of $f$, multiplied by $\widetilde{W}_i(\Lambda)$. This can be written as

$$\Pi_2(\Lambda, x_{-kl}) = \left( \sum_{z=0}^{2^{6h}-1} \beta \widehat{\mathbf{P}}(\underline{x}_k^- + z) \right) \cdot \widetilde{W}_i(\Lambda).$$

Now we see that multiplying (H.9) by $\beta$ gives

$$|\Pi_1(\Lambda, x_{-kl}) - \Pi_2(\Lambda, x_{-kl})| \leq (constant) \cdot \beta \cdot N^{-\left(\frac{48}{37} - \epsilon\right)}.$$

Since $\beta = Pr_\phi((A_j, P) \in [x_{-kl}]$, we see that this is exactly (H.7), as promised.

This completes the main goal of item (iii). Before leaving this case, however, let us consider what happens when we hold fixed $x_{-kl}$ and sum over $\Lambda$. If $BL^{-1}(\Lambda) \cap [x_{kl}] = \emptyset$, then $\Pi_1(\Lambda, x_{-kl}) = \Pi_2(\Lambda, x_{-kl}) = 0$, so these choices of $\Lambda$ will contribute nothing to the sum on the left-hand side of (H.4). How many block labels $\Lambda$ make a nonzero contribution, i.e. satisfy $BL^{-1}(\Lambda) \cap [x_{kl}] \neq \emptyset$? Suppose $\Lambda$ is such a block label, with length $t-1$. For each $s \leq t-1$ except for $s = k$, the value of $\Lambda_s$ is uniquely determined by the constraint $\underline{x}_s \leq x_s \leq \overline{x}^s$. (Recall that $l \geq t$.) This determines every component of $\Lambda$ except for $\Lambda_k$, and so we get at most $R+1$ such block labels.

Now we are ready to complete the proof of (H.2). Consider the sum

$$\sum_{\substack{\Lambda\in\Theta_{k,t}\\ x_{-kl}}} [\Pi_1(\Lambda, x_{-kl}) - \Pi_2(\Lambda, x_{-kl})]$$

on the left side of (H.2). Each term of the sum is indexed by a pair $(\Lambda, x_{-kl})$. Again, we can consider only terms with $BL^{-1}(\Lambda) \cap [x_{-kl}] \neq \emptyset$, because the other terms are all zero.

All the terms for which $x_{-kl}$ violates (H.5) have a total sum whose absolute value is bounded by a constant times $N^{-1}$ (this was case (i)). All the terms for which $\Lambda$ violates (H.6) have a sum that is again bounded by a constant times $N^{-1}$ (this was case (ii)). For the remaining terms, we apply case (iii). Consider any $x_{-kl}$ satisfying (H.5). Sum over all $\Lambda$ that satisfy (H.6). Using (H.7), and our previous observation that at most $R+1$ choices of $\Lambda$ make a nonzero contribution to the left-hand side, we get

$$\begin{aligned}\left|\sum_{\substack{\Lambda\in\Theta_{k,t}\\ \Lambda \text{ satisfies } (H.6)}} [\Pi_1(\Lambda, x_{-kl}) - \Pi_2(\Lambda, x_{-kl})]\right| \\ &\leq (constant)\cdot N^{-\left(\frac{48}{37}-\epsilon\right)} Pr_\phi((A_j, P)\in[x_{-kl}])\cdot(R+1)\\ &\leq (constant)\cdot N^{-\left(\frac{20}{37}-\epsilon\right)} Pr_\phi((A_j, P)\in[x_{-kl}])\end{aligned}$$

since $R+1 \leq (constant)\cdot N^{28/37}$. Summing over all choices of $x_{-kl}$, and using the obvious fact that

$$\sum_{x_{-kl} \text{ satisfies } (H.5)} Pr_\phi((A_j, P)\in[x_{-kl}]) \leq 1,$$

we obtain

$$\left|\sum_{\substack{\Lambda\in\Theta_{k,t} \text{ satisfying } (H.6)\\ x_{-kl} \text{ satisfying } (H.5)}} [\Pi_1(\Lambda, x_{-kl}) - \Pi_2(\Lambda, x_{-kl})]\right| \leq (constant)\cdot N^{-\left(\frac{20}{37}-\epsilon\right)}.$$

These three cases together cover every possible pair $(\Lambda, x_{-kl})$. So, adding them

together, we obtain (H.4). We already saw that (H.4) was equivalent to (H.2), so we have proven (H.2) and the proof of Claim I is complete.

**Proof of Claim II.** Rewrite the asserted bound as a sum over all $N$-profiles $P$:

$$\left|\sum_P Pr_\phi(P)\widetilde{W}_i(BL(A_j,P)) - \sum_P Pr_\phi(P)\widetilde{W}_i(BL(A_k,P))\right| < c_{II}N^{-(20/37-\epsilon)}$$

or equivalently

$$\left|\sum_P \mathbf{P}(P \mid N;\phi)\left[\widetilde{W}_i(BL(A_j,P)) - \widetilde{W}_i(BL(A_k,P))\right]\right| < c_{II}N^{-(20/37-\epsilon)}. \qquad \text{(H.10)}$$

Notice that the $P$ term on the left side can only be nonzero if $(A_j,P)$ and $(A_k,P)$ are in different blocks. In fact, it is necessary not only that these two terms be in different blocks but that these blocks have different rounded weights for $A_i$. We will bound the left side of (H.10) by bounding both the probability of drawing a $P$ for which $BL(A_j,P)$ and $BL(A_k,P)$ have different rounded weights for $A_i$, and the amount by which these rounded weights can differ.

Specifically, we will show

$$Pr_{IID(\phi)}\big(\widetilde{W}_i(BL(A_j,P)) \neq \widetilde{W}_i(BL(A_k,P))\big) < (constant)\cdot N^{-(1/2-\epsilon)} \qquad \text{(H.11)}$$

and

$$\left|\widetilde{W}_i(BL(A_j,P)) - \widetilde{W}_i(BL(A_k,P))\right| < (constant)\cdot N^{-3/74} \qquad \text{for each } P. \qquad \text{(H.12)}$$

First, we prove (H.11). Without loss of generality we may assume $j<k$.

Define $\Lambda_1,\ldots,\Lambda_M$ from the profile $(A_j,P)$ following the block label algorithm, and put $\Lambda=(\Lambda_1,\ldots,\Lambda_{t-1})=BL(A_j,P)$. Similarly define $\Lambda'_1,\ldots,\Lambda'_M$ from $(A_k,P)$, and put $\Lambda'=(\Lambda'_1,\ldots,\Lambda'_{t'-1})=BL(A_k,P)$. Notice that $\Lambda_s=\Lambda'_s$ for each $s$, except possibly if $s=j$ or $s=k$, in which case we may have $\Lambda'_j=\Lambda_j-1$ or $\Lambda'_k=\Lambda_k+1$, respectively.

We consider all the cases in which $\widetilde{W}_i(\Lambda)\neq\widetilde{W}_i(\Lambda')$. There are several possibilities,

depending whether the lengths $t, t'$ are different or equal.

(a) It may be that $t < t'$.

(b) It may be that $t > t'$.

If $t = t'$, then we must have $j < t$ and $\widetilde{W}_j(\Lambda) \neq \widetilde{W}_j(\Lambda')$, or else $k < t$ and $\widetilde{W}_k(\Lambda) \neq \widetilde{W}_k(\Lambda')$: otherwise, $\widetilde{W}_s(\Lambda) = \widetilde{W}_s(\Lambda')$ for all $s$. Thus we have just two remaining possibilities:

(c) $j < t$ and $\widetilde{W}_j(\Lambda) \neq \widetilde{W}_j(\Lambda')$.

(d) $k < t$ and $\widetilde{W}_k(\Lambda) \neq \widetilde{W}_k(\Lambda')$.

We will deal with each of these cases in turn, and show that the probability of each one is bounded above by a constant times $N^{-(1/2-\epsilon)}$.

(a) If $t < t'$, then $\Lambda_1 + \cdots + \Lambda_t > R$ but $\Lambda'_1 + \cdots + \Lambda'_t \leq R$. This can only happen if $j < t$, $\Lambda'_j = \Lambda_j - 1$ and $\Lambda_1 + \cdots + \Lambda_t = R + 1$. We will estimate the probability of these latter two equalities jointly occurring, for any *fixed* value of $t > j$.

Write $(A_j, P) = (x_1\ A_1, \ldots, x_M\ A_M)$ as usual. To have $\Lambda'_j = \Lambda_j - 1$ we must have $x_j = L + \Lambda_j S$ exactly. We claim that we need only worry about values of $x_j$ that are within $2N^{1/2}\sqrt{\ln N}$ of $\phi_j N$. Indeed, using Lemma C.4, the probability of realizing any given value of $x_j$ outside this range is at most

$$e^{-N \cdot \frac{(2N^{-1/2}\sqrt{\ln N})^2}{2}} = e^{-2\ln N} = N^{-2},$$

so the total probablity of realizing all such $x_j$ is at most $N^{-1}$.

We may also assume that $\underline{\alpha} \leq \phi_j \leq 1 - \underline{\alpha}$. For if $\phi_j < \underline{\alpha}$ and $x_j \leq \phi_j N + 2N^{1/2}\sqrt{\ln N}$, then $x_j < L$ (as long as $N$ is large enough); and if $\phi_j > 1 - \underline{\alpha}$ and $x_j \geq \phi_j N - 2N^{1/2}\sqrt{\ln N}$, then $x_j > L + RS \geq L + \Lambda_j S$ (again for large $N$).

The number of possible values of $x_j = L + \Lambda_j S$ that are within $2N^{1/2}\sqrt{\ln N}$ of $\phi_j N$ is at most a constant times $N^{1/2}\sqrt{\ln N}/S < N^{1/2+\epsilon}/S$. Moreover, for each such value, the probability of realizing it is at most a constant times $N^{-1/2}$,

by Lemma C.7. (Note that $x_j \leq L \sim \lambda N$ and $N - x_j \leq N - (L + RS) \sim (1 - \lambda - \mu)N$.)

Therefore,

$$Pr_{IID(\phi)}(x_j = L + \Lambda_j S) \leq (constant) \cdot N^{\epsilon}/S.$$

Now, *conditional* on the value of $x_j = L + \Lambda_j S$, the remaining terms $x_{-j}$ are distributed multinomially (by Lemma C.2). What is the probability that $\Lambda_1 + \cdots + \Lambda_t = R + 1$?

As long as $\Lambda_j < R + 1$, we are looking for the probability, under the specified multinomial distribution, that

$$\sum_{\substack{1 \leq s \leq t \\ s \neq j}} \Lambda_s = R + 1 - \Lambda_j.$$

Consider any realization of the profile for which this occurs. If we let $\Gamma$ be the set of indices $s$ ($1 \leq s \leq t$, $s \neq j$) such that $\Lambda_s > 0$, then we also have $\sum_{s \in \Gamma} \Lambda_s = R + 1 - \Lambda_j > 0$.

Consider *any* possible choice of the nonempty set $\Gamma$ not containing $j$, and estimate the probability that $\sum_{s \in \Gamma} \Lambda_s = R + 1 - \Lambda_j$, conditional on the value of $x_j = L + \Lambda_j S$. Since $|x_s - L - S(\Lambda_s - 1)| \leq S$ for each $s \in \Gamma$, the desired event can happen only if

$$\left| \sum_{s \in \Gamma} x_s - (|\Gamma| L + S(R + 1 - \Lambda_j - |\Gamma|)) \right| \leq |\Gamma| \cdot S.$$

This requires that the sum $\sum_{s \in \Gamma} x_s$ — which is binomially distributed — should lie between the lower bound

$$|\Gamma| L + S(R + 1 - \Lambda_j - |\Gamma|) - |\Gamma| S$$

and the upper bound

$$|\Gamma|L + S(R + 1 - \Lambda_j - |\Gamma|) + |\Gamma|S.$$

The lower bound is at least

$$|\Gamma|L + S(1 - 2M) \geq \frac{\lambda}{2}N$$

when $N$ is large, and the upper bound is at most

$$|\Gamma|L + S(R + 1 + M) \leq ML + RS + (M + 1)S \leq \frac{1 + \lambda + \mu}{2}N$$

when $N$ is large. Therefore, each realization of $\sum_{s\in\Gamma} x_s$ has probability bounded by a constant times $N^{-1/2}$ by Lemma C.7, and so their total probability is at most

$$(2|\Gamma| \cdot S + 1) \cdot (constant) \cdot N^{-1/2}.$$

Summing over all possible sets $\Gamma$ (there are certainly at most $2^{M-1}$ possibilities), we see that

$$Pr_{IID(\phi)}\left(\sum_{s\in\Gamma} \Lambda_s = R + 1 - \Lambda_j \text{ for some set } \Gamma, j \notin \Gamma \mid x_j = L + \Lambda_j S\right) \leq (constant) \cdot S \cdot N^{-1/2} \qquad \text{(H.13)}$$

for each fixed choice of $\Lambda_j < R + 1$.

Therefore,

$$Pr_{IID(\phi)}(\Lambda_1 + \cdots + \Lambda_t = R + 1 \mid x_j = L + \Lambda_j S) \leq (constant) \cdot N^{-1/2}$$

for each fixed choice of $\Lambda_j < R + 1$.

Finally,

$$
\begin{aligned}
&Pr_{IID(\phi)}(\Lambda'_j = \Lambda_j + 1 \text{ and } \Lambda_1 + \cdots + \Lambda_t = R+1) \\
&\qquad \le \sum_{\Lambda_j} Pr_{IID(\phi)}(x_j = L + \Lambda_j S \text{ and } \Lambda_1 + \cdots + \Lambda_t = R+1) \\
&\qquad \le \Bigg( \sum_{\Lambda_j < R+1} Pr_{IID(\phi)}(x_j = L + \Lambda_j S) \times \\
&\qquad Pr_{IID(\phi)}(\Lambda_1 + \cdots + \Lambda_t = R+1 \mid x_j = L + \Lambda_j S) \Bigg) \\
&\qquad\qquad + Pr_{IID(\phi)}(x_j = L + (R+1)S) \\
&\qquad \le \left( \sum_{\Lambda_j < R+1} Pr_{IID(\phi)}(x_j = L + \Lambda_j S) \cdot (constant) \cdot N^{-1/2} S \right) \\
&\qquad\qquad + (constant) \cdot N^{-1/2} \\
&\qquad \le \left( \sum_{\Lambda_j} Pr_{IID(\phi)}(x_j = L + \Lambda_j S) \right) \cdot (constant) \cdot N^{-1/2} S \\
&\qquad\qquad + (constant) \cdot N^{-1/2} \\
&\qquad \le (constant) \cdot (N^{\epsilon}/S) \cdot N^{-1/2} S + (constant) \cdot N^{-1/2} \\
&\qquad \le (constant) \cdot N^{-(1/2-\epsilon)}.
\end{aligned}
$$

This shows that the total probability of case (a) is at most a constant times $N^{-(1/2-\epsilon)}$.

(b) If $t > t'$, then $\Lambda_1 + \cdots + \Lambda_{t'} \le R$ but $\Lambda'_1 + \cdots + \Lambda'_{t'} > R$. This can only happen if $k < t$, $\Lambda'_k = \Lambda_k + 1$ and $\Lambda'_1 + \cdots + \Lambda'_{t'} = R+1$. From here we proceed exactly as in case (a), with $\Lambda$ and $\Lambda'$ interchanged, and with the role of $j$ played instead by $k$. We thus see that the probability of case (b) is also at most a constant times $N^{-(1/2-\epsilon)}$.

(c) Suppose $j < t$. If $\widetilde{W}_j(\Lambda) \neq \widetilde{W}_j(\Lambda')$, it must certainly happen that $W_j(\Lambda) \neq W_j(\Lambda')$, which requires $\Lambda_j \neq \Lambda'_j$ (since $j < t$). As in (a), this requires $\Lambda'_j = \Lambda_j - 1$ and $x_j = L + \Lambda_j S$ exactly. Also as in (a), we need only worry about values of $x_j$ that are within $2N^{1/2}\sqrt{\ln N}$ of $\phi_j N$, because the total probability of all other

values of $x_j$ is at most $N^{-1}$. Note that

$$|x_j - \phi_j N| \leq 2N^{1/2}\sqrt{\ln N}$$

is equivalent to

$$\left|\Lambda_j - \frac{\phi_j N - L}{S}\right| \leq \frac{2N^{1/2}\sqrt{\ln N}}{S}. \tag{H.14}$$

However, $\Lambda'_j = \Lambda_j - 1$ implies $W_j(\Lambda') = W_j(\Lambda) - 1/(R+1)$, and therefore

$$2^h W_j(\Lambda') = 2^h W_j(\Lambda) - \frac{2^h}{R+1}.$$

For the rounded weights to differ, $\widetilde{W}_j(\Lambda') \neq \widetilde{W}_j(\Lambda)$ or equivalently

$$\lfloor 2^h W_j(\Lambda) \rfloor \neq \lfloor 2^h W_j(\Lambda) \rfloor,$$

it must be that

$$K \leq 2^h W_j(\Lambda) < K + \frac{2^h}{R+1}$$

for some integer $K$. Writing this in terms of $\Lambda_j$, we have

$$K \leq \frac{2^h}{R+1}\Lambda_j < K + \frac{2^h}{R+1}.$$

Now, for each integer $K$, we get exactly one choice of $\Lambda_j$ that satisfies this. Moreover, the difference between two successive such values of $\Lambda_j$ is at least

$$\lfloor \frac{1}{2^h/(R+1)} \rfloor \geq \lfloor (constant) \cdot \frac{N/S}{S^{1/6}} \rfloor = \lfloor (constant) \cdot N^{53/74} \rfloor.$$

For $N$ sufficiently large, this is bigger than the width of the window in (H.14), since the latter is

$$2\frac{2N^{1/2}\sqrt{\ln N}}{S} \lesssim N^{1/2}.$$

Therefore, for $N$ sufficiently large, there is only one possible value of $\Lambda_j$ — say $\Lambda_j^*$ — that falls in the window (H.14) and allows $\widetilde{W}_j(\Lambda) \neq \widetilde{W}_j(\Lambda')$.

We also know that a realization of this case requires $\Lambda_j > 0$ (since $\Lambda'_j = \Lambda_j - 1$), and $\Lambda_j \leq R$ (since $j < t$). Thus, using the same arguments as in case (a), $(L + \Lambda^*_j S)/N$ is bounded strictly between 0 and 1, and so the probability of realizing $x_j = L + \Lambda^*_j S$ is bounded by a constant times $N^{-1/2}$.

In summary: for case (c), to happen, either (H.14) must be violated, which happens with probability at most $N^{-1}$; or we must have $x_j = L + \Lambda^*_j S$ for a specific value $0 < \Lambda^*_j \leq R$ (although this value may depend on $\phi$), which happens with probability at most a constant times $N^{-1/2}$. This shows that the total probability of case (c) is at most a constant times $N^{-1/2}$.

(d) For this case to happen, we must have $\Lambda_k = \Lambda'_k - 1$. From here we proceed exactly as in (c), with the roles of $\Lambda$ and $\Lambda'$ interchanged, and the role of $j$ played by $k$.

This covers all four cases (a)-(d), completing the proof of (H.11).

Next we prove (H.12). We retain the notation $\Lambda, \Lambda'$, and so forth from the proof of (H.11). We regard $j, k, P$ as fixed, and prove that (H.12) holds for every possible choice of $i = 1, \ldots, M$,

Suppose $i < \min\{t, t'\}$. We have three cases:

- If $i \neq j, k$, then $W_i(\Lambda) = \Lambda_i = \Lambda'_i = W_i(\Lambda')$ and so $\widetilde{W}_i(\Lambda) = \widetilde{W}_i(\Lambda')$.
- If $i = j$, then either $\Lambda_i = \Lambda'_i$ and so $\widetilde{W}_i(\Lambda) = \widetilde{W}_i(\Lambda')$ again, or else $\Lambda'_i = \Lambda_i - 1$. In the latter case,

$$2^h W_i(\Lambda) - 2^h W_i(\Lambda') = \frac{2^h}{R+1} < (constant) \cdot \frac{S^{1/6}}{N/S} < 1$$

for large enough $N$; hence

$$0 \leq \lfloor 2^h W_i(\Lambda) \rfloor - \lfloor 2^h W_i(\Lambda') \rfloor \leq 1$$

and so

$$\widetilde{W}_i(\Lambda) - \widetilde{W}_i(\Lambda') = \frac{\lfloor 2^h W_i(\Lambda) \rfloor - \lfloor 2^h W_i(\Lambda')}{2^h}$$

lies between 0 and

$$\frac{1}{2^h} = \frac{1}{S^{1/6}} < (constant) \cdot N^{-3/74}.$$

- If $i = k$, then either $\Lambda_i = \Lambda'_i$ or $\Lambda_i = \Lambda'_i - 1$, and we proceed as in the previous case.

Suppose $i > \max\{t, t'\}$. Then we have $W_i(\Lambda) = 0 = W_i(\Lambda')$, so $\widetilde{W}_i(\Lambda) = \widetilde{W}_i(\Lambda')$.

If $t = t'$, then we have shown that (H.12) holds for every $i \neq t$. By the identity

$$\sum_{i=1}^{M} \left( \widetilde{W}_i(BL(A_j, P)) - \widetilde{W}_i(BL(A_k, P)) \right) = 1 - 1 = 0 \tag{H.15}$$

we conclude that (H.12) holds for $i = t$ as well.

This leaves us only to deal with the case $t \neq t'$. Suppose that $t < t'$. As in case (a) of the proof of (H.11), this implies $j < t$, $\Lambda'_j = \Lambda_j - 1$ and $\Lambda_1 + \cdots + \Lambda_t = R + 1$, whereas $\Lambda'_s = \Lambda_s$ for every $s \leq t$, $s \neq j$. Hence $\Lambda'_1 + \cdots + \Lambda'_t = R$, and then $\Lambda'_s = 0$ for all $t < s < t'$ (because otherwise we would have $\Lambda'_1 + \cdots + \Lambda'_s > R$ contradicting the minimality of $t'$).

The above analysis showed that (H.11) holds for every $i < t$ and $i > t'$. Moreover, since $\Lambda_1 + \cdots + \Lambda_t = R + 1$, we have

$$W_t(\Lambda) = 1 - \frac{\sum_{s<t} \Lambda_j}{R+1} = \frac{\Lambda_t}{R+1},$$

while also $W_t(\Lambda') = \Lambda'_t/(R+1)$; and so the same logic used for the case $i < t$ shows that (H.11) holds for $i = t$ also. And if $t < i < t'$ then $W_i(\Lambda) = 0 = W_i(\Lambda')$. Thus, (H.11) holds for every $i \neq t'$. By (H.15), it holds for $i = t'$ as well.

This covers the case $t < t'$. The case $t > t'$ is identical, with the roles of $\Lambda$ and $\Lambda'$ interchanged and $k$ in place of $j$.

This completes the proof of (H.12). Now we can prove (H.10). Let $\Omega$ be the set

of all $N$-profiles $P$ for which $\widetilde{W}_i(BL(A_j,P)) \neq \widetilde{W}_i(BL(A_k,P))$. We have

$$\begin{aligned}
&\left|\sum_P \mathbf{P}(P \mid N;\phi)\left[\widetilde{W}_i(BL(A_j,P)) - \widetilde{W}_i(BL(A_k,P))\right]\right| \\
&\qquad = \left|\sum_{P\in\Omega} \mathbf{P}(P \mid N;\phi)\left[\widetilde{W}_i(BL(A_j,P)) - \widetilde{W}_i(BL(A_k,P))\right]\right| \\
&\qquad \leq Pr_{IID(\phi)}(P \in \Omega)\cdot(constant)\cdot N^{-\frac{3}{74}} \\
&\qquad\qquad \text{by (H.12)} \\
&\qquad \leq (constant)\cdot N^{-\left(\frac{1}{2}-\epsilon\right)-\frac{3}{74}} \\
&\qquad\qquad \text{by (H.11).}
\end{aligned}$$

This gives (H.10), and so Claim II is proven.

**Completion of Proof of Theorem 4.3.** Suppose the manipulator has belief $\phi$ and considers a change in his vote from $A_j$ to $A_k$. We show that this manipulation can change the probability of any candidate $A_i$ winning by (asymptotically) no more than a constant times $N^{-(20/37-\epsilon)}$. We have

$$\left|Pr_\phi(f(A_j,P)=A_i) - \sum_\Lambda Pr_\phi(A_j,P)\in BL^{-1}(\Lambda))\widetilde{W}_i(\Lambda)\right| \lesssim c_I N^{-(20/37-\epsilon)}$$

by Claim I;

$$\begin{aligned}
&\left|\sum_\Lambda Pr_\phi((A_j,P)\in BL^{-1}(\Lambda))\widetilde{W}_i(\Lambda) - \sum_\Lambda Pr_\phi((A_k,P)\in BL^{-1}(\Lambda))\widetilde{W}_i(\Lambda)\right| \\
&\qquad\qquad \lesssim \quad c_{II}N^{-(20/37-\epsilon)}
\end{aligned}$$

by Claim II;

$$\left|\sum_\Lambda Pr_\phi((A_k,P)\in BL^{-1}(\Lambda))\widetilde{W}_i(\Lambda) - Pr_\phi(f(A_k,P)=A_i)\right| \lesssim c_I N^{-(20/37-\epsilon)}$$

by Claim I again. The triangle inequality then gives

$$|Pr_\phi(f(A_j,P)=A_i) - Pr_\phi(f(A_k,P)=A_i)| \lesssim (2c_I + c_{II})N^{-(20/37-\epsilon)}.$$

The theorem follows, with (say) $\kappa = 20/37 - 2\epsilon$.

□

# Bibliography


[1] Milton Abramowitz and Irene A. Stegun (1972), *Handbook of Mathematical Functions* (Washington: U.S. Government Printing Office), tenth printing.

[2] Fuad Aleskerov and Eldeniz Kurbanov (1999), "Degree of Manipulability of Social Choice Procedures," in Ahmet Alkan, Charalambos D. Aliprantis, and Nicholas C. Yannelis, eds., *Current Trends in Economics: Theory and Applications* (Berlin, Heidelberg: Springer-Verlag), 13-27.

[3] Nabil I. Al-Najjar and Rann Smorodinsky (2000), "Pivotal Players and the Characterization of Influence," *Journal of Economic Theory* 92 (2), 318-342.

[4] Itai Ashlagi, Mark Braverman, and Avinatan Hassidim (2011), "Matching with Couples Revisited," extended abstract in *Proceedings of the 12th ACM Conference on Electronic Commerce* (EC-11), 335.

[5] Eduardo Azevedo and Eric Budish, "Strategyproofness in the Large as a Desideratum for Market Design," unpublished paper, Harvard University.

[6] Salvador Barberá (2001), "An Introduction to Strategy-Proof Social Choice Functions," *Social Choice and Welfare* 18 (4), 619-653.

[7] John J. Bartholdi III and James B. Orlin (1991), "Single Transferable Vote Resists Strategic Voting," *Social Choice and Welfare* 8 (4), 341-354.

[8] J. J. Bartholdi III, C. A. Tovey, and M. A. Trick (1989), "The Computational Difficulty of Manipulating an Election," *Social Choice and Welfare* 6 (3), 227-241.

[9] Dirk Bergemann and Stephen Morris (2005), "Robust Mechanism Design," *Econometrica* 73 (6), 1771-1813.

[10] Eleanor Birrell and Rafael Pass (2011), "Approximately Strategy-Proof Voting," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence* (IJCAI-11), 67-72.

[11] Steven J. Brams and Peter C. Fishburn (1978), "Approval Voting," *American Political Science Review* 72 (3), 831-847.

[12] Eric Budish (2011), "The Combinatorial Assignment Problem: Approximate Competitive Equilibrium from Equal Incomes," *Journal of Politicial Economy* 119 (6), 1061-1103.

[13] Donald E. Campbell and Jerry S. Kelly (2009), "Gains from Manipulating Social Choice Rules," *Economic Theory* 40 (3), 349-371.

[14] Kim-Sau Chung and J. C. Ely (2007), "Foundations of Dominant-Strategy Mechanisms," *Review of Economic Studies* 74 (2), 447-476.

[15] Claude d'Aspremont and Louis-André Gérard-Varet (1979), "Incentives and Incomplete Information," *Journal of Public Economics* 11 (1), 25-45.

[16] Robert Day and Paul Milgrom (2008), "Core-Selecting Package Auctions," *International Journal of Game Theory* 36 (3-4), 393-407.

[17] John Duggan (1996), "A Geometric Proof of Gibbard's Random Dictatorship Theorem," *Economic Theory* 7 (2), 365-369.

[18] Lars Ehlers, Hans Peters, and Ton Storcken (2004), "Threshold Strategy-Proofness: On Manipulability in Large Voting Problems," *Games and Economic Behavior* 49 (1), 103-116.

[19] Aytek Erdil and Paul Klemperer (2010), "A New Payment Rule for Core-Selecting Package Auctions," *Journal of the European Economic Association* 8 (2-3), 537-547.

[20] Pierre Favardin, Dominique Lepelley, and Jérôme Serais (2002), "Borda Rule, Copeland Method and Strategic Manipulation," *Review of Economic Design* 7 (2), 213-228.

[21] Peter C. Fishburn (1978), "A Strategic Analysis of Nonranked Voting Systems," *SIAM Journal of Applied Mathematics* 35 (3), 488-495.

[22] Allan Gibbard (1973), "Manipulation of Voting Schemes: A General Result," *Econometrica* 41 (4), 587-601.

[23] Allan Gibbard (1977), "Manipulation of Schemes that Mix Voting with Chance," *Econometrica* 45 (3), 665-681.

[24] Wassily Hoeffding (1963), "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association* 58 (301), 13-30.

[25] Nicole Immorlica and Mohammed Mahdian (2005), "Marriage, Honesty, and Stability," in *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (SODA 2005), 53-62.

[26] Marcus Isaksson, Guy Kindler, and Elchanan Mossel (2010), "The Geometry of Manipulation: A Quantitative Proof of the Gibbard Satterthwaite Theorem," in *Proceedings of the IEEE 51st Annual Symposium on Foundations of Computer Science* (FOCS-10), 319-328.

[27] Fuhito Kojima and Mihai Manea (2010), "Incentives in the Probabilistic Serial Mechanism," *Journal of Economic Theory* 145 (1), 106-123.

[28] Fuhito Kojima and Parag A. Pathak (2009), "Incentives and Stability in Large Two-Sided Matching Markets," *American Economic Review* 99 (3), 608-627.

[29] Anshul Kothari, David C. Parkes, and Subhash Suri (2005), "Approximately-Strategyproof and Tractable Multiunit Auctions," *Decision Support Systems* 39 (1), 105-121.

[30] Dominique Lepelley and Boniface Mbih (1987), "The Proportion of Coalitionally Unstable Situations under the Plurality Rule," *Economics Letters* 24 (4), 311-315.

[31] Dominique Lepelley and Boniface Mbih (1994), "The Vulnerability of Four Social Choice Functions to Coalitional Manipulation of Preferences," *Social Choice and Welfare* 11 (3), 253-265.

[32] Hitoshi Matsushima (2008), "Behavioral Aspects of Implementation Theory," *Economics Letters* 100 (1), 161-164.

[33] Hitoshi Matsushima (2008), "Role of Honesty in Full Implementation," *Journal of Economic Theory* 139 (1), 353-359.

[34] Stefan Maus, Hans Peters, and Ton Storcken (2007), "Anonymous Voting and Minimal Manipulability," *Journal of Economic Theory* 135 (1), 533-544.

[35] Stefan Maus, Hans Peters, and Ton Storcken (2007), "Minimal Manipulability: Unanimity and Nondictatorship," *Journal of Mathematical Economics* 43 (6), 675-691.

[36] Stefan Maus, Hans Peters, and Ton Storcken (2007), "Minimally Manipulable Anonymous Social Choice Functions," *Mathematical Social Sciences* 53 (3), 239-254.

[37] Stefan Maus, Hans Peters, and Ton Storcken (2007), "Minimal Manipulability: Anonymity and Unanimity," *Social Choice and Welfare* 29 (2), 247-269.

[38] Andrew McLennan (forthcoming), "Manipulation in Elections with Uncertain Preferences," *Journal of Mathematical Economics.*

[39] Frank McSherry and Kunal Talwar (2007), "Mechanism Design via Differential Privacy," in *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science* (FOCS-07), 94-103.

[40] Elchanan Mossel and Miklos Z. Racz (2011), "A Quantitative Gibbard-Satterthwaite Theorem Without Neutrality," arXiv preprint, `arXiv:1110.5888`.

[41] Hervé Moulin (1988), *Axioms of Cooperative Decision Making* (Cambridge: Cambridge University Press).

[42] Roger B. Myerson (1998), "Population Uncertainty and Poisson Games," *International Journal of Game Theory* 27 (3), 375-392.

[43] Roger B. Myerson (2000), "Large Poisson Games," *Journal of Economic Theory* 94 (1), 7-45.

[44] Richard G. Niemi (1984), "The Problem of Strategic Behavior under Approval Voting," *American Political Science Review* 78 (4), 952-958.

[45] Shmuel Nitzan (1985), "The Vulnerability of Point-Voting Schemes to Preference Variation and Strategic Manipulation," *Public Choice* 47 (2), 349-370.

[46] Parag A. Pathak and Tayfun Sönmez (2011), "School Admissions Reform in Chicago and England: Comparing Mechanisms by their Vulnerability to Manipulation," *American Economic Review*, forthcoming.

[47] Bezalel Peleg (1979), "A Note on Manipulability of Large Voting Schemes," *Theory and Decision* 11 (4), 401-412.

[48] Geoffrey Pritchard and Arkadii Slinko (2006), "On the Average Minimum Size of a Manipulating Coalition," *Social Choice and Welfare* 27 (2), 263-277.

[49] Reyhaneh Reyhani, Geoffrey Pritchard, and Mark C. Wilson, "A New Measure of the Difficulty of Manipulation of Voting Rules," unpublished working paper, University of Auckland.

[50] Donald John Roberts and Andrew Postlewaite (1976), "The Incentives for Price-Taking Behavior in Large Exchange Economies," *Econometrica* 44 (1), 115-127.

[51] Donald G. Saari (1990), "Susceptibility to Manipulation," *Public Choice* 64 (1), 21-41.

[52] Mark A. Satterthwaite (1975), "Strategy-proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions," *Journal of Economic Theory* 10 (2), 187-217.

[53] James Schummer (2004), "Almost-Dominant Strategy Implementation: Exchange Economies," *Games and Economic Behavior* 48 (1), 154-170.

[54] Arkadii Slinko (2002), "On Asymptotic Strategy-Proofness of the Plurality and the Run-Off Rules," *Social Choice and Welfare* 19 (2), 313-324.

[55] Arkadii Slinko (2002), "On Asymptotic Strategy-Proofness of Classical Social Choice Rules," *Theory and Decision* 52 (4), 389-398.

[56] David A. Smith (1999), "Manipulability Measures of Common Social Choice Functions," *Social Choice and Welfare* 16 (4), 639-661.

[57] Tayfun Sönmez and M. Utku Ünver (2011), "Matching, Allocation, and Exchange of Discrete Resources," in *Handbook of Social Economics*, vol. 1A, eds. Jess Benhabib, Matthew O. Jackson, and Alberto Bisin (Amsterdam: North-Holland).

[58] Yves Sprumont (1995), "Strategyproof Collective Choice in Economic and Political Environments," *Canadian Journal of Economics* 28 (1), 68-107.

[59] Yasuhito Tanaka (2003), "An Alternative Direct Proof of Gibbard's Random Dictatorship Theorem," *Review of Economic Design* 8 (3), 319-328.

[60] Robert Wilson (1987), "Game-Theoretic Approaches to Trading Processes," in Truman F. Bewley, ed., *Advances in Economic Theory: Fifth World Congress* (Cambridge: Cambridge University Press), 33-77.

[61] Takuro Yamashita (2011), "A Necessary Condition on Robust Implementation: Theory and Applications," unpublished working paper, Toulouse School of Economics.

# Chapter 3

# The Efficiency-Incentive Tradeoff in Double Auction Environments


## Abstract

We consider the tradeoff between efficiency and incentives in large double auction environments with weak budget balance. No mechanism simultaneously gives agents perfect incentives to be truthful and ensures first-best efficiency, but a planner designing a mechanism may be willing to compromise on either of these dimensions for improvements along the other. She would then naturally wish to find where the possibility frontier lies with respect to incentives and efficiency. We make inroads on this question: our main result locates the frontier to within a factor that is logarithmic in the size of the market.



Thanks to (in random order) Alessandro Bonatti, Xiao Yu Wang, Ruitian Lang, Glenn Ellison, Parag Pathak, Rakesh Vohra, and Juuso Toikka, as well as an anonymous referee, for helpful comments and advice.


## 1 Introduction

### 1.1 Overview

Economists have known since Akerlof [2] that private information can prevent markets from reaching efficient outcomes. Moreover, the results of Myerson and Satterthwaite [24], among many others, show that this inefficiency is not specific to competitive

markets but rather is unavoidable under any possible mechanism for allocating goods. However, some mechanisms lead to more severe inefficiency than others, and so the natural next question is what second-best mechanism achieves outcomes that are as efficient as possible. A large literature addresses this question in many different settings.

Customarily, the mechanism design literature assumes that agents optimize perfectly. In particular, applying the revelation principle, it is standard to take as a given constraint that each agent's best possible strategy should be to truthfully reveal his private information, and then describe the optimal mechanism subject to this constraint.

However, in practice, human decision-makers are not perfectly strategic, or at least do not perfectly optimize the material payoffs that are usually modeled. Accordingly, a planner could offer a mechanism asking agents to report their preferences, in which reporting truthfully is not exactly optimal, but the incentives to behave strategically instead are small. The planner might then expect that agents will report truthfully, rather than go to the trouble of figuring out how to strategically manipulate the mechanism. This notion leads to a tradeoff between incentives and efficiency, and motivates a quantitative examination of the tradeoff.

The present paper makes initial inroads into quantitatively studying this tradeoff, in the specific context of large double auction environments with quasilinear preferences and weak budget balance. This is one of the most widely studied economic environments for mechanism design, and can be viewed as an analytically convenient, stylized model of an exchange economy.

By studying the incentive-efficiency tradeoff, we bridge two branches of theoretical research on mechanisms for large markets. On one hand is the literature, going back to Roberts and Postlewaite [26], showing that in large exchange economies, under the competitive equilibrium mechanism, the incentives for strategic misreporting of preferences (assuming other agents are truthful) go to zero. On the other hand is a recent literature studying exact equilibria of large markets and showing that the inefficiency goes to zero [15, 16, 27]. In particular, part of that latter literature [12, 28] takes a

mechanism design approach and identifies the optimal rate at which *any* mechanism can converge to full efficiency as the market becomes large. However, no previous work has explored the space in between these branches, looking for compromises between perfect efficiency and perfect incentives. If it turned out that large gains in efficiency could be achieved at the cost of a very small relaxation of incentives, that would cast a new light on the existing convergence-rate bounds. Conversely, if this were not possible, the existing impossibility results would be strengthened.

Our modeling framework is fundamentally non-equilibrium-based, intended to study design of market institutions for agents who are not perfectly familiar with their environment. Indeed, our basic motivating assumption — that agents do not effortlessly know how to manipulate to their advantage — would be difficult to justify in an equilibrium model. On the other hand this assumption is reasonable for describing plenty of exchange in real-world markets. The typical shopper at the grocery store is unlikely to think about the demand curve of other shoppers for a pint of strawberries, or to know how he might profitably deviate from pure price-taking behavior so as to influence the prices he faces — or to even want to bother thinking about how he might go about strategically deviating.

To explore quantitatively the tradeoff between incentives and efficiency, we need ways to measure both. As in the second chapter of this dissertation, we work in a direct revelation framework, where a mechanism asks each agent his value for the good being exchanged, and determines trades accordingly; and we take a worst-case approach to the definition of incentives. The *susceptibility to manipulation* of a particular market mechanism is the largest amount of expected utility any agent could possibly gain by reporting his value strategically instead of truthfully; the maximum is taken over all possible beliefs about the distributions from which other agents' behavior is drawn. Likewise, we also use a worst-case measure for inefficiency: it is the largest value, over all possible distributions of agents' valuations, of the expected shortfall in surplus realized by the mechanism compared to the first-best (assuming that agents report truthfully).[1]

[1]We measure inefficiency using the allocation of goods, not the sum of the agents' utilities. These

Our worst-case methodology is appropriate for a planner choosing a trading institution to be used in the future, when she does not have clear priors over agents' valuations or their strategic behavior, and wants to be sure that her mechanism will perform well. (The second chapter of this dissertation fleshes out in detail a positive model of such a planner's choice of mechanism, showing how our measurement methodology fits in.) In addition, when defining susceptibility, note that we take the worst case over *beliefs*: there is no presumption that agents know the true distribution of others' behavior. This is in keeping with our non-equilibrium framework, in which agents may not accurately know the details of their environment.

Two mechanisms in the existing literature represent polar cases with respect to the efficiency-incentive tradeoff. On one end is the *k-double auction*, a version of the competitive mechanism; where the goods are given to the traders whose (reported) values are highest, and trades take place at a market-clearing price. This mechanism achieves first-best efficiency if traders are truthful, but does not provide perfect incentives for truthfulness. On the other end is McAfee's [22] dominant-strategy double auction, which provides perfect incentives, but may fail to realize (at most) one profitable trade.

Our results, presented in Section 3, describe the asymptotic behavior of susceptibility or inefficiency as the number of agents becomes large. We consider environments in which buyers' valuations are independently drawn from one distribution, sellers' valuations are independently drawn from another distribution, and these two distributions are not too dissimilar. More precisely, the distributions have densities that differ everywhere by at most some fixed ratio. Then the $k$-double auction has susceptibility on the order of $1/\sqrt{N}$, and McAfee's double auction has inefficiency on the order of $1/\sqrt{N}$, where $N$ represents the size of the market. Our main result (Theorem 3.3) shows that both mechanisms are close to the possibility frontier: There is a constant $c$ such that any mechanism has either susceptibility or inefficiency at least $c/(\sqrt{N}\log N)$.

---

measures are different if the mechanism runs a surplus. Our measure implicitly assumes that the surplus can be paid to someone outside the mechanism.

The assumption of similar distributions is necessary. If we allow the buyers' and sellers' valuations to come from arbitrarily different distributions, then the lower bound on susceptibility or inefficiency does not go to zero as the market grows (Proposition 3.4).

In Section 4, we address a possible "consequentialist" critique of our methodology: Perhaps a planner designing mechanisms should not be concerned with incentives for strategic manipulation per se, since agents might manipulate in a way that does not adversely affect the outcome of the mechanism. Instead, she should be concerned with the inefficiency that will result from manipulation. It turns out that our results withstand this critique, as long as we make reasonably conservative assumptions about how agents might try to manipulate. Specifically, we allow that agents may attempt any manipulation that gives them a sufficiently large gain in expected utility (they will not necessarily find the optimal manipulation), and we consider the inefficiency that may result. In this formulation, instead of a tradeoff between efficiency under truthfulness and incentives for truthfulness, we have a tradeoff between efficiency under manipulation and the planner's confidence about agents' cost of strategic behavior. Only a little extra work is needed to reformulate our main results in these terms.

In addition to the results themselves, the method of proof for the lower bounds merits attention. We use a straightforward variation on a standard proof of the impossibility of attaining both first-best efficiency and perfect incentives. That proof uses the usual integral formula derived from the envelope theorem to compute the utility that each type of each agent would need to receive, and verifies that the total surplus in the market is not enough to provide that utility to each agent. We introduce error terms into the proof, representing inefficiency and susceptibility to manipulation. By continuity, the same contradiction is still reached if the error terms are sufficiently small; we simply track them explicitly to find out how large they need to be to avoid a contradiction. Some care is needed in working the error terms into the integral formula: it turns into a discrete approximation, and one needs to choose the approximation points appropriately. However, the fact that we can readily adapt a

standard argument to obtain our results is encouraging, since it suggests that similar methods can be applied to study tradeoffs involving incentives in other mechanism design domains.

## 1.2 Literature review

The question of incentives for truthfulness in large markets can be traced back to Roberts and Postlewaite [26], who showed that the benefits from misreporting one's demand function in an exchange economy (under the Walrasian mechanism) go to zero as the economy is replicated. More recent work in the market design literature gives similar results for matching mechanisms [3, 14, 17, 18], argues that this property makes the mechanisms suitable for use in practice. A variety of other literature has also considered mechanisms with small incentives to manipulate [7, 11, 19, 20, 21, 23, 29], but without looking at the possibility frontier between these incentives and other properties of the mechanism, as we emphasize here.

In contrast to this approach, much of the recent work on double auctions has assumed that agents perfectly optimize — thus imposing Bayesian Nash equilibrium, with given valuation distributions — and examined either behavior in specific mechanisms or the design problem of finding the optimal mechanism. Several relevant papers studied rates of convergence to perfect efficiency. In the model of Rustichini, Satterthwaite, and Williams [27], equilibrium behavior in the $k$-double auction leads to inefficiency tending to zero as the market grows, at rate $1/N$. McAfee's dominant-strategy double auction [22] also attains rate $1/N$. Satterthwaite and Williams [28] showed (for the uniform distribution) that any mechanism has inefficiency of order at least $1/N$, so that the the two mechanisms just described are asymptotically optimal, to within a constant factor. (These results appear to contradict our Theorem 3.3 below, which implies worse rates of convergence. The discrepancy arises because we allow for a broader class of value distributions.) There is also recent work on large double auctions with interdependent values, e.g. [25]. However, our focus here is on environments with private values.

# 2 Model

## 2.1 Elements

We consider double auction settings with unit capacity, private values, and quasilinear utility. Thus, there are $N$ sellers who each have a good to sell, and $N$ buyers who each would like to buy a good. Write $b_i$ for the value of the good to buyer $i$, and $s_i$ for the value to seller $i$. These values are normalized to lie in $[0,1]$. We write $P^b, P^s$ for profiles of buyers' and sellers' valuations, $(b_i)_{i=1,\ldots,N}$ and $(s_i)_{i=1,\ldots,N}$, and $P = (P^b, P^s)$ for the profile of all $2N$ agents' valuations. Then $P^b_{-i}$ denotes the profile of valuations of all buyers except the $i$th, and $P^s_{-i}$ similarly.

We focus attention on direct mechanisms. (This, and other assumptions, will be discussed in Subsection 2.2.) Thus, a *mechanism* elicits each agent's valuation, and determines an allocation of the goods (possibly probabilistic) and expected transfer payments as a function of the reported valuations. Formally, a mechanism is a collection of $4N$ functions,

$$M = (p^b_i, p^s_i, t^b_i, t^s_i)_{i=1,\ldots,N}$$

where

$$p^b_i, p^s_i : [0,1]^{2N} \to [0,1]$$

denote each agent's probability of exchange (i.e. $p^b_i$ is buyer $i$'s probability of receiving a good, and $p^s_i$ is seller $i$'s probability of giving up a good); and

$$t^b_i, t^s_i : [0,1]^{2N} \to \mathbb{R}$$

denote the net payment made by each agent. We require the functions $p^b_i, p^s_i, t^b_i, t^s_i$ to be measurable. We also impose the feasibility conditions

$$\sum_i p^b_i(P) = \sum_i p^s_i(P)$$

for every profile of valuations $P \in [0,1]^{2N}$.

We do not allow the mechanism to run a deficit, but we do allow a surplus; thus we impose weak budget balance:

$$\sum_i t_i^b(P) + t_i^s(P) \geq 0$$

for all $P$. (With deficits allowed, the Vickrey-Clarke-Groves mechanism [9, 13] would achieve full efficiency in dominant strategies, so the tradeoff between efficiency and incentives would be uninteresting.)

If the profile of reported valuations is $P$, then the utilities of buyer $i$ and seller $i$, respectively (relative to not participating in the mechanism), are

$$U_i^b(P) = b_i p_i^b(P) - t_i^b(P), \qquad U_i^s(P) = -s_i p_i^s(P) - t_i^s(P).$$

In addition to feasibility and weak budget balance, we also require mechanisms to satisfy ex post individual rationality:

$$U_i^b(P), U_i^s(P) \geq 0$$

for all profiles $P$ and all $i$. Note that individual rationality and weak budget balance imply that the transfers $t_i^b(P), t_i^s(P)$ are bounded.

In the operation of a mechanism, we assume that the buyers' valuations are drawn independently from a distribution $F^b$ on $[0,1]$, and the sellers' valuations are drawn independently from a distribution $F^s$. We will in general not presume these distributions are known, either to the planner or to the agents, but rather allow a set $\mathcal{F}$ of possible pairs $(F^b, F^s)$. We will assume that for all possible pairs, $F^b$, $F^s$ are representable by bounded density functions on $[0,1]$. Our results would be unchanged (and indeed simpler to prove) if we allowed for atoms in the distributions, but by requiring continuity we make clear that atoms are not driving the results. We will sometimes write $f^b, f^s$ for the respective density functions.

The utility achieved by buyer $i$ when the *reported* profile is $\widehat{P}$ but his true valuation

is $b_i$ is

$$U_i^b(\widehat{P}|b_i) = b_i p_i^b(\widehat{P}) - t_i^b(\widehat{P}).$$

Similarly define

$$U_i^s(\widehat{P}|s_i) = -s_i p_i^s(\widehat{P}) - t_i^s(\widehat{P}).$$

We define the *susceptibility to manipulation* of a mechanism $M$ as the strongest possible incentive faced by any agent to misreport his valuation. Formally, for a given set $\mathcal{F}$ of distribution pairs, the *buyer-susceptibility* is

$$\sigma^b = \sup_{i,b_i,\widehat{b}_i,(F^b,F^s)} \Big( E_{(F^b,F^s)}[U_i^b(\widehat{b}_i, P_{-i}^b, P^s|b_i)] - E_{(F^b,F^s)}[U_i^b(P|b_i)] \Big)$$

where the supremum is over buyers $i$, true valuations $b_i \in [0,1]$, possible reports $\widehat{b}_i \in [0,1]$, and distribution pairs $(F^b, F^s) \in \mathcal{F}$. The expectations are with respect to other agents' reported types, where we assume other buyers' reports are drawn from $F^b$ and sellers' from $F^s$ (all independently). Similarly the *seller-susceptibility* is

$$\sigma^s = \sup_{i,s_i,\widehat{s}_i,(F^b,F^s)} \big( E_{(F^b,F^s)}[U_i^s(P^b, \widehat{s}_i, P_{-i}^s|s_i)] - E_{(F^b,F^s)}[U_i^s(P|s_i)] \big).$$

The *susceptibility* is then

$$\sigma = \max\{\sigma^b, \sigma^s\}.$$

The motivating story behind this definition is simple: Suppose a planner knows that agents face a psychological or computational cost of at least $\epsilon$ to behaving strategically. If the planner chooses a mechanism whose susceptibility is known to be less than $\epsilon$, then agents will not bother to behave strategically and instead will simply report their true valuations. This is discussed in more detail in the preceding chapter of this dissertation, which also shows how the above definition of susceptibility is equivalent to one in which players are allowed to be uncertain about the distribution pair $(F^b, F^s)$.

We define the *inefficiency* of a mechanism using an analogous worst-case formulation. For any profile $P$ of valuations, define the first-best welfare $W^{FB}(P)$ to be the

sum of the $N$ highest valuations, and the welfare $W^M(P)$ achieved by the mechanism as $\sum_i b_i p_i^b(P) + \sum_i s_i(1 - p_i^s(P))$. Note that

$$W^M(P) = \left[\sum_i U_i^b(P) + \sum_i U_i^s(P)\right] + \left[\sum_i t_i^b(P) + \sum_i t_i^s(P)\right] + \left[\sum_i s_i\right].$$

The second bracketed expression is the surplus accrued by the mechanism; we implicitly assume when computing welfare that this surplus can be paid to an outside agent. The third expression is independent of the choice of mechanism, so it does not affect the shortfall relative to first best, $W^{FB}(P) - W^M(P)$.

The inefficiency of $M$ relative to $\mathcal{F}$ is then defined as

$$\sup_{(F^b,F^s)} \left(E_{(F^b,F^s)}[W^{FB}(P) - W^M(P)]\right),$$

where the supremum is over $(F^b, F^s) \in \mathcal{F}$, and the expectation is with respect to valuation profiles where each $b_i$ is drawn from $F^b$ and each $s_i$ is drawn from $F^s$ (independently). In particular, this definition of inefficiency assumes truthful reporting; we will address this issue in Section 4. Also, the definition is absolute (not normalized by the size of the market), though our results could just as well be formulated in terms of relative inefficiency.

We will be mainly concerned with a set of distribution pairs $\mathcal{F}$ in which the buyers' and sellers' value distributions are not too different. Specifically, let $\lambda \geq 1$ be an exogenously given constant; then define $\mathcal{F}_\lambda$ to be the family of distribution pairs $(F^b, F^s)$ whose densities satisfy $f^b(x)/\lambda \leq f^s(x) \leq \lambda f^b(x)$ for all $x \in [0, 1]$. (As a special case, $\lambda = 1$ means that the buyers' and sellers' values are drawn from the same distribution.) Our main results apply to $\mathcal{F}_\lambda$. However, we will also consider the set $\mathcal{F}_\infty$, of all possible pairs $(F^b, F^s)$ of distributions representable by bounded density functions on $[0, 1]$.

Note that we have not required mechanisms to be *anonymous* — that is, to treat all buyers and all sellers identically. Formally, a mechanism $M$ is anonymous if, for

all profiles $(b_i, s_i)$ and all permutations $\pi^b, \pi^s$ of $\{1, \ldots, N\}$, we have

$$p_i^b(b_{\pi^b(1)}, \ldots, b_{\pi^b(N)}, s_{\pi^s(1)}, \ldots, s_{\pi^s(N)}) = p_{\pi^b(i)}^b(b_1, \ldots, b_N, s_1, \ldots, s_N)$$

for each $i$, and similarly for the functions $p_i^s, t_i^b, t_i^s$. However, to study the inefficiency-susceptibility frontier, it is enough to consider anonymous mechanisms. Indeed, if $M$ is any mechanism with susceptibility $\sigma$ and inefficiency $\eta$, we can define an anonymous mechanism $\widetilde{M}$ by randomly permuting the buyers and the sellers and then applying $M$: that is,we define

$$\widetilde{p}_i^b(b_1, \ldots, b_N, s_1, \ldots, s_N) = \frac{1}{(N!)^2} \sum_{\pi^b, \pi^s} p_{(\pi^b)^{-1}(i)}^b(b_{\pi^b(1)}, \ldots, b_{\pi^b(N)}, s_{\pi^s(1)}, \ldots, s_{\pi^s(N)})$$

and define $\widetilde{p}_i^s, \widetilde{t}_i^b, \widetilde{t}_i^s$ likewise; these comprise the mechanism $\widetilde{M}$. Then $\widetilde{M}$ is an average of $(N!)^2$ mechanisms, all of which (by symmetry) have gains at most $\sigma$ to any agent from manipulating and all of which have an expected welfare loss at most $\eta$ relative to the first-best, so the same is true of $\widetilde{M}$. Thus we have an anonymous mechanism whose susceptibility and inefficiency are at most those of $M$.

Given this, we will henceforth restrict attention to anonymous mechanisms without further comment.

## 2.2 Discussion

There are a couple of assumptions implicit in the above modeling framework which call for elaboration. Our restriction to direct mechanisms really entails two assumptions: first, that each agent's strategy depends only on his valuation (and no other information); second, that the strategy space can be taken to be the space of valuations, with honest reporting as the default behavior of agents who do not strategize.

The second assumption is actually not a serious restriction. We view double auction environments as a stylized model of competitive markets, and truthfulness as a metaphor for price-taking. This seems a natural assumption about default behavior (especially for inexperienced participants). But more generally, we could take an

indirect-mechanism approach, allowing a mechanism $M$ to specify any strategy space for each player, together with probabilities of trade and transfers as functions of the strategy profile, *and* a specification of a default strategy for each player (possibly mixed) that depends on that player's valuation. By a straightforward variation of the usual revelation principle, $M$ could be converted into a direct mechanism $M'$, where default behavior consists of honest reporting, and where $M'$ has the same inefficiency as $M$ and susceptibility no higher than $M$ (it may have strictly lower susceptibility, due to the elimination of strategies in $M$ that were not default strategies for any type). Since we are concerned only with the inefficiency-susceptibility frontier, it suffices to focus on direct mechanisms as we have done above.

The assumption that players' behavior depends only on their valuations is more serious. This assumption invites the critique of Bergemann and Morris [6] that a planner could potentially do better by designing a mechanism in which agents also condition their strategies on their beliefs about other agents' behavior. If we were to formulate the mechanism design problem in full generality taking this into account, a direct mechanism would have agents report their full types, where a type consists not only of a valuation but an entire belief hierarchy (including beliefs about any parameters relevant to agents' manipulative behavior — see the discussion in Section 4 below).

However, recall that we have chosen to make no assumptions about the correctness of agents' beliefs about others' behavior. The appropriate worst-case measure of inefficiency in this framework would specify that for a mechanism to have inefficiency at most $\eta$, the expected welfare loss relative to first-best should be at most $\eta$ for every possible distribution of buyer and seller *types*, regardless of whether or not their beliefs reflected the true distribution. With such a definition, it turns out that our results would remain valid even in this more fully-specified setting. This is because the proofs of our lower bounds rely only on a single "true" distribution pair $(F^b, F^s)$ when analyzing the incentive to misreport, and so these lower bounds actually hold for the subset of the type space on which it is common knowledge among the agents that values are drawn from this $(F^b, F^s)$. On this subset, two types of a

given agent differ only in their valuation, so assuming that agents report only their valuations is without loss of generality. (However, the proofs do analyze *inefficiency* using distributions other than the fixed $(F^b, F^s)$, so it is crucial that the definition of inefficiency allows the "true" distribution to be different from the one that agents believe to be correct.) Thus the critique of [6] does not bind here. The formal details of this argument would be notationally involved and not relevant to the main point of this paper, so we omit them.

We should also comment here on the interpretation of the individual rationality constraints, which we have written in an ex post form. These can be thought of as normative constraints on acceptable mechanisms. They can also be viewed in positive terms, if agents have the opportunity to renege after the mechanism has operated. However, this latter interpretation is less tidy: as pointed out by Compte and Jehiel [10], the proper formulation of such a constraint is as a *veto constraint*, which not only requires ex post individual rationality but also imposes stronger incentive constraints — agents should not be able to benefit by misreporting their valuation and then potentially vetoing the outcome depending on the realizations of other agents' types. This distinction turns out to be immaterial for our results, however: our negative results under individual rationality still hold a fortiori under the stronger veto constraint, and it can be checked that our positive results also hold, since the relevant mechanisms (the McAfee and $k$-double auctions) satisfy the veto constraint.

Alternatively, using a richer type space as outlined above, in which strategies reflect an agent's full type, would allow us to instead use an interim version of the individual rationality constraints — each agent has nonnegative expected utility from participation — in which case the positive interpretation would be straightforward. Our lower bounds would still hold with these weaker constraints rather than the ex post constraints, again for the reason that the proofs invoke the constraints only for agents whose beliefs coincide with the true distributions $(F^b, F^s)$. Again, we omit the details.

## 2.3 Polar mechanisms

We now describe in precise terms our two polar mechanisms. We will content ourselves with verbal descriptions, rather than tediously write out all the algebraic expressions.

For any $k \in [0,1]$, the *k-double auction* (described e.g. in [27]) is as follows. For any profile $P$ of $2N$ reported valuations, sort them as $v_{(1)} \geq v_{(2)} \geq \cdots \geq v_{(2N)}$, and define the price $p^* = kv_{(N)} + (1-k)v_{(N+1)}$. Allocate the goods to the agents with the $N$ highest valuations. (If there is a tie at $v_{(N)}$, ration uniformly at random; ties are not really important since they occur with probability zero in our model.) Every buyer who receives a good pays $p^*$, and every seller who sells a good receives $p^*$. It is clear that this mechanism satisfies feasibility, budget balance, and individual rationality, and that it achieves inefficiency of 0.

*McAfee's double auction*, from [22], is a bit more complex. The rules are as follows. Sort the buyers' reported valuations in decreasing order, $b_{(1)} \geq \cdots \geq b_{(N)}$, and the sellers' in increasing order, $s_{(1)} \leq \cdots \leq s_{(N)}$. Also define $b_{(0)} = s_{(N+1)} = 1$ and $b_{(N+1)} = s_{(0)} = 0$ for convenience. Let $k$ be the highest value satisfying $b_{(k)} \geq s_{(k)}$; this is the efficient number of trades. We have $0 \leq k \leq N$. Define the price $p^* = (b_{(k+1)} + s_{(k+1)})/2$.

If $p^* \in [s_{(k)}, b_{(k)}]$, then have the $k$ highest-value buyers buy the good from the $k$ lowest-value sellers at price $p^*$. (Again, break ties uniformly at random.) Otherwise, note that $k > 0$, and have the $k-1$ highest-value buyers each receive a good and pay $b_{(k)}$, while the $k-1$ lowest-value sellers each sell their good for price $s_{(k)}$. The mechanism thus carries out $k-1$ trades and earns a budget surplus $(k-1)(b_{(k)} - s_{(k)}) \geq 0$.

This mechanism is again feasible, weakly budget-balanced, and individually rational. It has been established that reporting truthfully is a dominant strategy for all agents in this mechanism [22, Theorem 1]. Therefore, it has a susceptibility of 0.

# 3 The efficiency-incentive tradeoff

We can now properly introduce our results on the efficiency-incentive tradeoff.

The results are illustrated in Figure 3.1, where the gray region represents the (inefficiency, susceptibility) pairs $(\eta, \sigma)$ attained by some mechanism. The frontier must be convex, as shown in the figure: If mechanism $M$ has inefficiency $\eta$ and susceptibility $\sigma$, and mechanism $M'$ has inefficiency $\eta'$ and susceptibility $\sigma'$, then for any $\alpha \in [0, 1]$ we can take the convex combination $(1-\alpha)M + \alpha M'$ (defined by taking corresponding convex combinations of the $p_i^b, p_i^s, t_i^b, t_i^s$ functions), and this mechanism has inefficiency at most $(1-\alpha)\eta + \alpha\eta'$ and susceptibility at most $(1-\alpha)\sigma + \alpha\sigma'$.

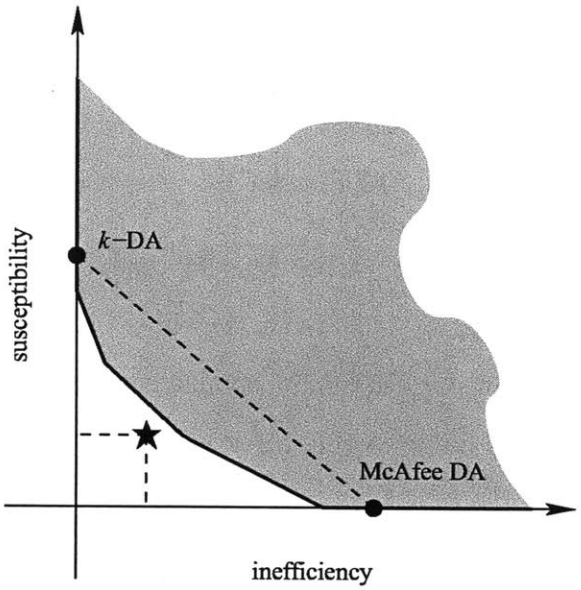


Figure 3.1: The possibility frontier

For the main results, we consider the class of distribution pairs $\mathcal{F}_\lambda$, in which some similarity is imposed between the buyers' and sellers' value distributions. We give the approximate locations of the two polar mechanisms, which lie on the two axes of the possibility set, at a distance of order $1/\sqrt{N}$ from the origin. On the other hand, we identify a point lying below the possibility set (indicated by the star in the figure), whose coordinates are of order $1/(\sqrt{N}\log N)$. Thus, these results together pin down the location of the possibility frontier to within a factor that is logarithmic in the size of the market.

If we look at the class of distribution pairs $\mathcal{F}_\infty$, where the distribution of buyers' values can be arbitrarily different from the distribution of sellers' values, then a

similar picture applies but on a different scale: the lower bound on inefficiency or susceptibility (the star point) does not go to zero as the market becomes large. This will be shown in Subsection 3.2.

## 3.1 Main results

We first bound the inefficiency attained by the McAfee double auction, over $\mathcal{F}_\lambda$. As the number of agents grows, the inefficiency shrinks on the order of $N^{-1/2}$. More specifically:

**Proposition 3.1** *There is a constant $c$ such that the McAfee double auction has inefficiency at most $c/\sqrt{N}$ on $\mathcal{F}_\lambda$. (The value of $c$ depends on $\lambda$.)*

The calculation is routine, but rather lengthy, so we leave it for Appendix A. For a quick overview: Inefficiency is at most the value of the least valuable trade; a change-of-variables argument implies that this value is no greater than the probability that the least valuable trade involves of a buyer with value above $x^*$ and a seller with value below $x^*$, for a suitable (fixed) $x^*$. For this to happen, in turn, it must be that either (a) the number of agents with values above $x^*$ is close to $N$, which happens with probability on the order of $N^{-1/2}$ by a law-of-large-numbers argument; or (b) when all $2N$ agents are arranged from highest value to lowest, there is a long run of consecutive buyers or consecutive sellers, which happens with probability decreasing exponentially in the length of the run.

We can also bound the susceptibility of the $k$-double auction; it is also on the order of $N^{-1/2}$. This is because the probability that any given misreport is pivotal — that is, that it advantageously changes the market price — is of order at most $N^{-1/2}$, by a central-limit-theorem argument.

**Proposition 3.2** *There is a positive constant $c$ such that the $k$-double auction has susceptibility at most $cN^{-1/2}$. (Again, $c$ may depend on $\lambda$.)*

**Proof:** Consider a buyer with value $b$, reporting a false value $\widehat{b}$. We may assume $\widehat{b} < b$, since reporting $\widehat{b} > b$ can never be profitable: holding fixed the realizations of

other agents' reports, such a misreport cannot decrease the price, nor can it change the buyer's outcome from not receiving a good to receiving one, unless the trade occurs at a price higher than $b$.

Moreover, again holding fixed the other agents' reports, the misreport can only be beneficial if is pivotal — more specifically, if exactly $N-1$ other agents report values higher than $\widehat{b}$. Indeed, if more than $N-1$ other agents report higher values, then the misreporting buyer gets no good and hence utility zero; if fewer than $N-1$ other agents report higher values, then the misreport has no effect on the price at which he trades.

Since the buyer's realized utility is always between 0 and 1, his expected gain from misreporting is at most the probability that exactly $N-1$ other agents report a value greater than $\widehat{b}$. Letting $J$ be the number of other buyers whose values are less than $\widehat{b}$, we can express this probability as a sum over possible values of $J$:

$$\sum_{J=0}^{N-1} \binom{N-1}{J} \binom{N}{J} F^b(\widehat{b})^J F^s(\widehat{b})^{N-J} (1-F^b(\widehat{b}))^{N-1-J} (1-F^s(\widehat{b}))^J. \qquad (3.1)$$

We finish by invoking Lemma A.2 in Appendix A. If $F^b(\widehat{b}) \leq 1/2$, then using $\binom{N-1}{J} \leq \binom{N}{J}$, the expression in (3.1) is

$$\leq 2 \sum_{J=0}^{N} \binom{N}{J}^2 F^b(\widehat{b})^J F^s(\widehat{b})^{N-J} (1-F^b(\widehat{b}))^{N-J} (1-F^s(\widehat{b}))^J$$

which, according to the lemma (with $\kappa = 1/2$, say, and $K = 0$), is at most $c\sqrt{\lambda/N}$ for some *absolute* constant $c$. This certainly implies the desired bound on the buyer's probability of being pivotal.

If $F^b(\widehat{b}) > 1/2$, then using $\binom{N-1}{J} \leq \binom{N}{J+1}$, the expression in (3.1) is

$$\leq 2 \sum_{J=0}^{N-1} \binom{N}{J+1} \binom{N}{J} F^b(\widehat{b})^{J+1} F^s(\widehat{b})^{N-J} (1-F^b(\widehat{b}))^{N-1-J} (1-F^s(\widehat{b}))^J,$$

and by a change of variable, this is

$$= 2\sum_{J=1}^{N} \binom{N}{J-1}\binom{N}{J}(1-F^s(\widehat{b}))^{N-J}(1-F^b(\widehat{b}))^{J-1}F^s(\widehat{b})^J F^b(\widehat{b})^{N-J+1}$$

which, again according to the lemma (with $K = 1$), is at most $c\sqrt{\lambda/N}$ for an absolute constant $c$.

This shows in both cases that the buyer-susceptibility of the $k$-double auction satisfies the bound. The argument for seller-susceptibility is identical.

□

Having established these estimates for the two polar mechanisms, we can proceed to our main result: a lower bound showing that the two polar mechanisms are close to the optimal rate of convergence of inefficiency or susceptibility as the number of agents becomes large.

**Theorem 3.3** *There exists a positive constant c such that, on $\mathcal{F}_1$, every mechanism has either inefficiency at least $c/(\sqrt{N}\log N)$ or susceptibility at least $c/(\sqrt{N}\log N)$.*

Of course, the same bounds a fortiori hold for any $\mathcal{F}_\lambda$, $\lambda \geq 1$.

The idea behind the proof is as follows: Consider the incentives facing a given agent — say, a buyer — when he believes the other agents' values are drawn from a distribution with mass concentrated near 0 and 1. Let $\overline{p}^b(b)$ be the probability that the buyer gets a good when his value is $b$ (and he reports truthfully). Let $\overline{U}^b(b)$ be the expected utility he attains if his value is $b$. Similarly define $\overline{p}^s(s)$ and $\overline{U}^s(s)$.

Suppose the mechanism were to have inefficiency and susceptibility zero. Then the first-best allocation would determine $\overline{p}^b$ and $\overline{p}^s$ completely. In turn, these determine the functions $\overline{U}^b$ and $\overline{U}^s$ via the familiar integral formula coming from the envelope theorem (up to a constant, which is bounded below by individual rationality). These expected utility functions are not consistent with weak budget balance — there is not enough expected surplus in the market to give all agent types the needed utility levels.

Let $\overline{p}^{b*}, \overline{p}^{s*}, \overline{U}^{b*}, \overline{U}^{s*}$ be the functions obtained in the above calculations assuming zero inefficiency and susceptibility. With a small amount of wiggle room, we know only that $\overline{p}^b$ and $\overline{p}^s$ have to be close to $\overline{p}^{b*}$ and $\overline{p}^{s*}$, and in turn that $\overline{U}^b, \overline{U}^s$ have to be close to $\overline{U}^{b*}, \overline{U}^{s*}$. Requiring the agents' expected utility levels to be far enough from $\overline{U}^{b*}, \overline{U}^{s*}$ to avoid exceeding the total surplus in the market then leads to a lower bound on either inefficiency or susceptibility.

**Proof of Theorem 3.3:** It is enough to prove the result for $N$ sufficiently large; we can then adjust the constant $c$ to ensure the result holds for small $N$ as well.[2]

Suppose that the number $c$ is such that some mechanism $M$ has susceptibility $\sigma$ and inefficiency $\eta$ both less than $c/(\sqrt{N}\log N)$. Our goal is to show that $c$ must be larger than some absolute constant. Specifically, we will show that $c \geq 1/7000$. (This is far from best possible, but we are not concerned here with fine-tuning constants.) Thus, suppose that $c < 1/7000$, and seek a contradiction.

Let $\gamma$ be a sufficiently small positive number. At several points in the course of the proof, we will use the fact that $\gamma$ is smaller than various functions of $N$, $\eta$, and $c$. Rather than writing out explicit bounds here, we will simply assume without further comment that all needed bounds are satisfied (there will be only finitely many of them, so this assumption is safe).

Define the density function $f$ by

$$f(x) = \begin{cases} 1/2\gamma, & 0 \leq x \leq \gamma; \\ 0, & \gamma < x < 1-\gamma; \\ 1/2\gamma, & 1-\gamma \leq x \leq 1. \end{cases}$$

Let $F$ be the corresponding cumulative distribution function. We focus on the incentives facing a given agent when all other agents' reports are independently drawn from $F$.

[2]To be precise, this requires knowing that for each small $N$, either inefficiency or susceptibility must be bounded away from 0. By continuity arguments, it is enough to show that there is no mechanism with inefficiency and susceptibility both 0. This can be proven e.g. by using revenue equivalence to show that any such mechanism would have to be equivalent to a VCG mechanism, which always runs a deficit; see [31].

**Step 1 (buyers).** As in the sketch above, let $\overline{p}^b(b)$ be the probability that buyer $i$ receives a good, when his value is $b$. (By anonymity, this is independent of $i$.) In this first step, we use efficiency to show that $\overline{p}^b$ is close to its first-best value.

However, because we are working with continuous distributions, efficiency in expectation imposes no restrictions on $\overline{p}^b(b)$ itself for any single value of $b$. Instead, we need to talk about averages. Accordingly, for $\frac{\gamma}{2} \leq b \leq 1 - \frac{\gamma}{2}$, define

$$\overline{p}^b_\gamma(b) = \frac{1}{\gamma} \int_{b-\frac{\gamma}{2}}^{b+\frac{\gamma}{2}} \overline{p}^b(b')\, db'.$$

We will show that $\overline{p}^b_\gamma(b)$ is approximately bounded below b times $1/2$ minus a constant times $\eta/b$. Specifically, for any $b > 3\gamma/2$,

$$\overline{p}^b_\gamma(b) \geq \frac{1}{2} - \frac{16\eta}{b - 3\gamma/2}. \tag{3.2}$$

To show this, suppose otherwise, so that

$$\frac{1}{16}\left(b - \frac{3\gamma}{2}\right)\left(\frac{1}{2} - \overline{p}^b_\gamma(b)\right) > \eta. \tag{3.3}$$

Define a density function $g$ by

$$g(x) = \begin{cases} \frac{1}{\gamma}, & b - \frac{\gamma}{2} < x < b + \frac{\gamma}{2}; \\ 0 & \text{otherwise.} \end{cases}$$

Define the density $h(x) = \left(1 - \frac{1}{N}\right) f(x) + \left(\frac{1}{N}\right) g(x)$. Let $G, H$ be the distributions associated with $g, h$.

Suppose that we draw all $2N$ agents' values independently from $H$. This is equivalent to generating values as follows: we mark each agent as an $F$-type or $G$-type agent, randomly with probability $1 - \frac{1}{N}$ or $\frac{1}{N}$ respectively, and then draw the valuations from $F$ or $G$ accordingly. Let $E$ denote the event that there is exactly one

$G$-type buyer and no $G$-type seller. We have

$$Pr(E) = \left[N \cdot \left(1 - \frac{1}{N}\right)^{N-1} \cdot \left(\frac{1}{N}\right)\right] \cdot \left(1 - \frac{1}{N}\right)^{N} \geq \frac{1}{16}. \tag{3.4}$$

Conditional on $E$, the $G$-type buyer receives a good with probability $\overline{p}^b_\gamma(b)$.

On the other hand, conditional on $E$, all $F$-type agents have values distributed uniformly on the set $[0, \gamma] \cup [1-\gamma, 1]$. In this case, the probability that at least half the $F$-type agents have values in $[0, \gamma]$ is $1/2$, by symmetry. Thus, conditional on $E$, we have probability at least $1/2$ that the $G$-type buyer is among the top $N$ values, and the next lower value is at most $\gamma$. In particular, conditional on $E$, there is probability at least $1/2 - \overline{p}^b_\gamma(b)$ that the $G$-type buyer is among the top $N$ values but does not receive a good, and the next highest value is at most $\gamma$. When this occurs, there is an efficiency loss (relative to first-best) of at least $b - \frac{3\gamma}{2}$.

Therefore, conditional on $E$, we have an expected efficiency loss (relative to first-best) of at least $\left(\frac{1}{2} - \overline{p}^b_\gamma(b)\right)\left(b - \frac{3\gamma}{2}\right)$. Since $Pr(E) \geq 1/16$, we have an unconditional inefficiency of at least $\frac{1}{16}\left(\frac{1}{2} - \overline{p}^b_\gamma(b)\right)\left(b - \frac{3\gamma}{2}\right)$. But this amount is greater than $\eta$ by (3.3). We have a contradiction. Therefore, (3.2) must hold.

In fact, we have the simpler bound

$$\overline{p}^b_\gamma(b) \geq \frac{1}{2} - \frac{32\eta}{b} \tag{3.5}$$

(as long as $\gamma < 64\eta/3$). Indeed, (3.2) implies (3.5) for $b \geq 3\gamma$, and for $b < 3\gamma$ the right side of (3.5) is negative, so the inequality holds trivially.

Henceforth we will only need this latter bound.

**Step 2 (buyers).** We next construct a discrete approximation for the standard integral formula, leading to a lower bound on the utilities of buyer types with high values. Specifically, we will show that buyers with values in the interval $[1 - \gamma, 1]$ must, on average, achieve utility at least $1/2 - 1/20\sqrt{N}$.

To this end, let $\overline{t}^b(b)$ be the expected payment by buyer $i$, when his value is $b$, and other values are drawn independently from $F$. Again, this is independent of $i$. Let

$\overline{U}^b(b) = b\overline{p}^b(b) - \overline{t}^b(b)$ be the expected utility achieved by a buyer with value $b$.

Take $K = \lfloor \log N \rfloor$. Define buyer values $b_0, b_1, \ldots, b_K$ by

$$b_j = \left(1 - \frac{\gamma}{2}\right)^{1-\frac{j}{K}} \left(\frac{1}{20\sqrt{N}}\right)^{\frac{j}{K}}.$$

(The subscripts here simply index the values; they do not denote different buyer identities.) These buyer values will essentially serve as the interval endpoints in our approximation to the integral formula. However, instead of using these values exactly, we will need to average over small perturbations of the values. This is because our available bounds on probabilities of trade apply to the averages $\overline{p}^b_\gamma$, not to $\overline{p}^b$ for any single type.

Define $\rho$ to be the ratio of successive $b_j$'s:

$$\rho = \frac{b_j}{b_{j+1}} = \left(\frac{1 - \frac{\gamma}{2}}{1/20\sqrt{N}}\right)^{1/K},$$

and note that

$$\rho \leq (20\sqrt{N})^{1/K} \leq 20^{2/\log N}(\sqrt{N})^{2/\log N} = 20^{2/\log N} e < 3 \tag{3.6}$$

(as long as $N$ is large enough, as usual).

Now, by definition of $\sigma$, for any $r \in [-\gamma/2, \gamma/2]$, a buyer of type $b_j + r$ (for any $j$) cannot benefit by more than $\sigma$ from misreporting as type $b_{j+1} + r$.

Consider any such $r$ . We have

$$\begin{aligned}
\overline{U}^b(b_j + r) &= (b_j + r)\overline{p}^b(b_j + r) - \overline{t}^b(b_j + r) \\
&\geq (b_j + r)\overline{p}^b(b_{j+1} + r) - \overline{t}^b(b_{j+1} + r) - \sigma \\
&= \overline{U}^b(b_{j+1} + r) + (b_j - b_{j+1})\overline{p}^b(b_{j+1} + r) - \sigma
\end{aligned}$$

for each $j$. By combining these inequalities for each $j$ we get

$$\begin{aligned}\overline{U}^b(b_0+r) &\geq \overline{U}^b(b_K+r)+\sum_{j=0}^{K-1}(b_j-b_{j+1})\overline{p}^b(b_{j+1}+r)-K\sigma \\ &\geq \sum_{j=0}^{K-1}(b_j-b_{j+1})\overline{p}^b(b_{j+1}+r)-K\sigma \qquad (3.7)\end{aligned}$$

where the last step is by individual rationality.

Now average over $[-\gamma/2,\gamma/2]$. For each $j$, we know from (3.5) that

$$\frac{1}{\gamma}\int_{-\frac{\gamma}{2}}^{\frac{\gamma}{2}}\overline{p}^b(b_{j+1}+r)\,dr \;\geq\; \frac{1}{2}-\frac{32\eta}{b_{j+1}}.$$

Therefore,

$$\begin{aligned}\frac{1}{\gamma}\int_{-\frac{\gamma}{2}}^{\frac{\gamma}{2}}\overline{U}^b(b_0+r)\,dr &\geq \sum_{j=0}^{K-1}(b_j-b_{j+1})\left[\frac{1}{\gamma}\int_{-\frac{\gamma}{2}}^{\frac{\gamma}{2}}\overline{p}^b(b_{j+1}+r)\,dr\right]-K\sigma \\ &\geq \sum_{j=0}^{K-1}(b_j-b_{j+1})\left(\frac{1}{2}-\frac{32\eta}{b_{j+1}}\right)-\frac{c}{\sqrt{N}} \\ &= (b_0-b_K)\left(\frac{1}{2}\right)-K(\rho-1)\,(32\eta)-\frac{c}{\sqrt{N}} \\ &\geq \left(1-\frac{\gamma}{2}-\frac{1}{20\sqrt{N}}\right)\left(\frac{1}{2}\right)-(\log N)\left(64\frac{c}{\sqrt{N}\log N}\right)-\frac{c}{\sqrt{N}} \\ &> \frac{1}{2}-\frac{1}{40\sqrt{N}}-\frac{70c}{\sqrt{N}} \\ &> \frac{1}{2}-\frac{1}{20\sqrt{N}}. \qquad (3.8)\end{aligned}$$

(The fourth line uses (3.6), and the sixth uses the assumption $c < 1/7000 < 1/2800$.)

Now, to wrap up this step of the proof, consider the expected utility accruing to buyer $i$, when all agents' values are drawn independently from $F$ (and all agents report truthfully). With probability $1/2$, buyer $i$ has a value in the interval $[1-\gamma,1]$; and conditional on being in this interval, buyer $i$'s value is uniformly distributed on

the interval. Therefore, buyer $i$'s unconditional expected utility is at least

$$\frac{1}{2}\left(\frac{1}{2} - \frac{1}{20\sqrt{N}}\right) = \frac{1}{4} - \frac{1}{40\sqrt{N}}.$$

**Steps 1, 2 (sellers).** The analysis up to this point has focused on incentives for buyers. However, exactly the same calculations can be performed with incentives for sellers. We briefly outline the arguments. Let $\overline{p}^s(s)$ be the probability that a seller with value $s$ sells his good, when all other agents' values are independently drawn from $F$. Define

$$\overline{p}^s_\gamma(s) = \frac{1}{\gamma}\int_{s-\frac{\gamma}{2}}^{s+\frac{\gamma}{2}} \overline{p}^s(s')\,ds'.$$

We can use the same efficiency arguments as before to obtain a counterpart to (3.5):

$$\overline{p}^s_\gamma(s) \geq \frac{1}{2} - \frac{32\eta}{1-s}. \tag{3.9}$$

Now define $\overline{t}^s(s)$ be the expected net transfer paid by a seller with value $s$, and let $\overline{U}^s(s) = -s\overline{p}^s(s) - \overline{t}^s(s)$ be the expected utility such a seller attains. Define $K$ as before, and define the seller values $s_0, \ldots, s_K$ by

$$s_j = 1 - \left(1 - \frac{\gamma}{2}\right)^{1-\frac{j}{K}}\left(\frac{1}{20\sqrt{N}}\right)^{\frac{j}{K}}.$$

As before, for any $r \in [-\gamma/2, \gamma/2]$,

$$\overline{U}^s(s_j + r) \geq \overline{U}^s(s_{j+1} + r) + (s_{j+1} - s_j)\overline{p}^s(s_{j+1} + r) - \sigma$$

for each $j$. Summing over $j$, averaging over $r \in \left[-\frac{\gamma}{2}, \frac{\gamma}{2}\right]$, and applying (3.9), we obtain

$$\frac{1}{\gamma}\int_{-\frac{\gamma}{2}}^{\frac{\gamma}{2}} \overline{U}^s(s_0 + r)\,dr > \frac{1}{2} - \frac{1}{40\sqrt{N}} - \frac{70c}{\sqrt{N}} > \frac{1}{2} - \frac{1}{20\sqrt{N}}.$$

Finally, as with the buyers, we conclude that when all agents' values are independently drawn from $F$, each seller's expected utility is at least $1/4 - 1/40\sqrt{N}$.

**Step 3.** To complete the proof, we use the lower bound on each agent's utility from Step 2, compare to the total expected surplus available, and obtain a contradiction.

The lower bound of $1/4 - 1/40\sqrt{N}$ obtained at the end of Step 2 holds for each of the $2N$ agents, and therefore the expected surplus generated by the mechanism — that is, the expected sum of the agents' utilities — is bounded below as

$$E[\sum_i U_i^b(P) + \sum_i U_i^s(P)] \geq \frac{N}{2} - \frac{\sqrt{N}}{20}. \tag{3.10}$$

On the other hand, due to weak budget balance, the surplus at any profile $P$ satisfies

$$\sum_i U_i^b(P) + \sum_i U_i^s(P) \leq \sum_i b_i p_i^b(P) - \sum_i s_i p_i^s(P) \leq W^{FB}(P) - \sum_i s_i. \tag{3.11}$$

Let's bound the expectation of the first-best welfare $W^{FB}$. Each agent's value is either in $[0, \gamma]$ or in $[1-\gamma, 1]$, independently with probability $1/2$. Letting $K$ be the number of agents with high values, we can bound the first-best by summing over possible values of $K$:

$$\begin{aligned} E[W^{FB}(P)] &\leq \sum_{K=0}^{2N} \binom{2N}{K} \left(\frac{1}{2}\right)^{2N} [\min\{N,K\} \cdot 1 + (N - \min\{N,K\}) \cdot \gamma] \\ &\leq N\gamma + \sum_{K=0}^{2N} \binom{2N}{K} \left(\frac{1}{2}\right)^{2N} \min\{N,K\}. \end{aligned}$$

Break the sum into terms with $K \leq N - \lfloor\sqrt{N}/4\rfloor$ and $K > N - \lfloor\sqrt{N}/4\rfloor$, rearrange, and then use Lemma A.3 from Appendix A (a crude central-limit-theorem

approximation) to bound from below the probability that $K \leq N - \lfloor\sqrt{N}/4\rfloor$:

$$\begin{aligned}
E[W^{FB}(P)] &\leq N\gamma + \sum_{K=0}^{N-\lfloor\sqrt{N}/4\rfloor} \binom{2N}{K}\left(\frac{1}{2}\right)^{2N} \cdot \left(N - \left\lfloor\frac{\sqrt{N}}{4}\right\rfloor\right) \\
&\qquad + \sum_{K=N-\lfloor\sqrt{N}/4\rfloor+1}^{2N} \binom{2N}{K}\left(\frac{1}{2}\right)^{2N} \cdot N \\
&= N\gamma + \sum_{K=0}^{2N} \binom{2N}{K}\left(\frac{1}{2}\right)^{2N} \cdot N \\
&\qquad - \sum_{K=0}^{N-\lfloor\sqrt{N}/4\rfloor} \binom{2N}{K}\left(\frac{1}{2}\right)^{2N} \cdot \left\lfloor\frac{\sqrt{N}}{4}\right\rfloor \\
&\leq N\gamma + N - \frac{1}{4}\left\lfloor\frac{\sqrt{N}}{4}\right\rfloor \\
&< N - \frac{\sqrt{N}}{20}.
\end{aligned}$$

This bounds the expectation of $W^{FB}(P)$.

The expression (3.11) also involves a $\sum_i s_i$ term. But since each seller has expected value $N$, the expectation of this sum is simply $N/2$. Consequently, (3.11) implies that the expected surplus is less than

$$\frac{N}{2} - \frac{\sqrt{N}}{20}.$$

Comparing with (3.10), we have a contradiction, which completes the proof.

□

## 3.2 Unrestricted distributions

We now show how the results change when no restrictions are imposed on the pair of distributions — we use the full class $\mathcal{F}_\infty$, rather than $\mathcal{F}_\lambda$

Trivially, the McAfee double auction has inefficiency at most 1 (since it omits at most one desirable trade), and the $k$-double auction has susceptibility at most 1 (since no agent can ever achieve utility greater than 1). Thus, it is possible to achieve zero

inefficiency or susceptibility and a constant, independent of the market size, along the other dimension. The following result shows that it is not possible to do better:

**Proposition 3.4** *There exists a positive constant c such that, on $\mathcal{F}_\infty$, every mechanism has either inefficiency or susceptibility at least c.*

The argument is somewhat similar to that of Theorem 3.3, but simpler. Since the proof is relatively brief, we will not bother explicitly breaking it into steps.

**Proof:** We will give a proof with $c = 1/128$. So suppose for contradiction that some mechanism $M$ has susceptibility $\sigma$ and inefficiency $\eta$ both less than $1/128$. Let $\gamma$ be a positive number, chosen to be very small; as in the proof of Theorem 3.3, we will not bother being explicit about the bounds needed on $\gamma$.

Let the distributions $F^b$, $F^s$ be given by the densities

$$f^b(x) = \begin{cases} 1/\gamma, & 1-\gamma \le x \le 1; \\ 0, & 0 \le x < 1-\gamma; \end{cases}$$

$$f^s(x) = \begin{cases} 1/\gamma, & 0 \le x \le \gamma; \\ 0, & \gamma < x \le 1. \end{cases}$$

Also let $G^b$ be the distribution with density

$$g^b(x) = \begin{cases} 1/\gamma, & \frac{1}{4}-\gamma \le x \le \frac{1}{4}; \\ 0 & \text{otherwise}, \end{cases}$$

and take $H^b(x) = \left(1-\frac{1}{N}\right) F^b(x) + \left(\frac{1}{N}\right) G^b(x)$. Drawing a buyer's value from $H^b$ is equivalent to designating the buyer as $F^b$-type or $G^b$-type, with probabilities $1-\frac{1}{N}$ or $\frac{1}{N}$ respectively, and then drawing the value from $F^b$ or $G^b$ accordingly.

Suppose all buyers' values are drawn from $H^b$ and all sellers' values are drawn from $F^s$. Let $E$ be the event that there is exactly one $G^b$-type buyer. By calculations similar to (3.4), we have

$$Pr(E) \ge \frac{1}{4}.$$

Whenever $E$ occurs, the first-best allocation assigns all the goods to the buyers, and any failure to assign a good to some buyer entails an efficiency loss of at least $(1/4-\gamma)-(\gamma) \geq 1/8$. In particular, if $\pi$ is the probability that the $G^b$-type buyer ends up with a good (conditional on $E$), we have

$$\eta \geq Pr(E) \cdot (1-\pi) \cdot \frac{1}{8} \geq \frac{1-\pi}{32}$$

from which

$$\pi \geq 1 - 32\eta > \frac{3}{4}.$$

Now let $\overline{p}^b(b)$ denote the probability that a buyer receives a good, when he reports $b$ and all other agents' reports are drawn from $(F^b, F^s)$. The above implies that the average of $\overline{p}^b(b)$ with respect to $G^b$ is at least 3/4:

$$\frac{1}{\gamma}\int_{\frac{1}{4}-\gamma}^{\frac{1}{4}} \overline{p}^b(b)\, db \geq \frac{3}{4}.$$

Let $\overline{U}^b(b)$ be the utility attained by a buyer with value $b$, when other agents' reports are drawn from $(F^b, F^s)$.

For any $r \in [0,\gamma]$, a buyer of value $1-r$ cannot benefit by more than $\sigma$ by misreporting as value $1/4-r$. And so, as in Step 2 of the proof of Theorem 3.3, we have

$$\overline{U}^b(1-r) \geq \overline{U}^b\left(\frac{1}{4}-r\right) + \left(1-\frac{1}{4}\right)\overline{p}^b\left(\frac{1}{4}-r\right) - \sigma.$$

Averaging over $r \in [0,\gamma]$ gives

$$\begin{aligned}
\frac{1}{\gamma}\int_{1-\gamma}^{1} \overline{U}^b(b)\, db &\geq \frac{1}{\gamma}\int_{\frac{1}{4}-\gamma}^{\frac{1}{4}} \overline{U}^b(b)\, db + \frac{3}{4}\left[\frac{1}{\gamma}\int_{\frac{1}{4}-\gamma}^{\frac{1}{4}} \overline{p}^b(b)\, db\right] - \sigma \\
&\geq \frac{3}{4}\left[\frac{1}{\gamma}\int_{\frac{1}{4}-\gamma}^{\frac{1}{4}} \overline{p}^b(b)\, db\right] - \sigma \\
&\geq \frac{9}{16} - \sigma \\
&> \frac{1}{2}.
\end{aligned}$$

Now finally suppose all agents' values are drawn from $(F^b, F^s)$. Each buyer's expected utility from the mechanism is $(1/\gamma)\int_{1-\gamma}^{1} \overline{U}^b(b)\,db$, which is greater than $1/2$, by the above calculation.

By identical arguments, each seller's expected utility is also greater than $1/2$.

But this means that when all agents' values are drawn from $(F^b, F^s)$, the total expected surplus generated by the mechanism must be more than $2N \cdot 1/2 = N$. Since it is never possible to generate a surplus of more than $N$, we have a contradiction.

□

# 4 A consequentialist approach

The exposition so far has focused on incentives for truthful reporting. This follows a substantial literature that treats strategyproofness as a basic normative criterion for evaluating mechanisms (see [5] for a survey, and [4] for a succinct summary of several justifications). However, others [6, 8] have raised the criticism that truthfulness should not be an end in itself; rather, what matters is the outcome that occurs as a result of any manipulations. In particular, in the double auction environment, there is an unambiguous objective available to a planner with such a "consequentialist" philosophy — namely, the efficiency of the realized allocation of goods — and so it is especially natural to frame the design problem in terms of this objective.

Fortunately, it turns out that there is a close connection between our formulation of the efficiency-incentive tradeoff and an alternative formulation that focuses on outcome efficiency. To motivate the latter formulation, imagine a planner who wants to ensure an allocation within $\eta$ of the first-best welfare, and who is uncertain not only about the distributions $(F^b, F^s)$ but also about the agents' strategic behavior. Thus, the planner wants to ensure that no matter what manipulations the agents perform, welfare is always within $\eta$ of the first-best (in expectation over realizations of the agents' types).

To describe the planner's problem, we must specify how she expects agents to manipulate. As sketched in Subsection 2.1, we presume there is a computational

cost of at least $\epsilon$ to behaving strategically, so the planner is confident that agents will not manipulate if they cannot gain more than $\epsilon$ expected utility by doing so. What if they can gain more than $\epsilon$? We could assume that agents will choose the manipulation that is optimal (with respect to their beliefs), but this would stray from our motivating notion of inexperienced, boundedly-rational agents. Instead we will take a more agnostic approach: agents may potentially make any misreport that would gain at least $\epsilon$ expected utility.[3]

We formalize this approach as follows. Given a mechanism $M$, a class $\mathcal{F}$ of distribution pairs, and a minimum manipulation cost $\epsilon$, define the *manipulation set* for each possible buyer's valuation $b_i \in [0,1]$ as

$$W^b(b_i;\epsilon) = \{b_i\} \cup$$
$$\{\widehat{b_i} \mid E_{(F^b,F^s)}[U_i^b(\widehat{b_i}, P_{-i}^b, P^s|b_i)] - E_{(F^b,F^s)}[U_i^b(b_i, P_{-i}^b, P^s)] \geq \epsilon$$
$$\text{for some } (F^b, F^s) \in \mathcal{F}\}.$$

(This is independent of $i$, by anonymity.) This set represents the set of all valuations that the planner believes a buyer might report, given that his true valuation is $b_i$. Note that we always include $b_i$: no matter what the mechanism is, we allow for the possibility that strategizing is so costly that the buyer just tells the truth. Similarly, for each seller's valuation $s_i$ we define

$$W^s(s_i;\epsilon) = \{s_i\} \cup$$
$$\{\widehat{s_i} \mid E_{(F^b,F^s)}[U_i^s(P^b, \widehat{s_i}, P_{-i}^s|s_i)] - E_{(F^b,F^s)}[U_i^s(P^b, s_i, P_{-i}^s)] \geq \epsilon$$
$$\text{for some } (F^b, F^s) \in \mathcal{F}\}.$$

From the planner's point of view, each buyer's true valuation and his report are drawn from a *joint* distribution $H^b$ on $[0,1] \times [0,1]$, independently across buyers. We say that such a joint distribution $H^b$ is *possible* if it places probability 1 on the set of

[3]In the preceding chapter of the dissertation, we gave a positive model that effectively assumes the planner considers any misreport to be possible if she is not certain that the agents cannot gain more than $\epsilon$. The model here is more refined.

pairs $(b, \widehat{b})$ such that $\widehat{b} \in W^b(b; \epsilon)$; and similarly for joint distributions $H^s$ of sellers' valuations and reports.

The planner's measure of inefficiency is given by the worst case over all possible *joint* distributions $H^b$ and $H^s$. For any profile $P$ of true valuations and $\widehat{P}$ of reports, define the realized welfare $W^M(\widehat{P}|P)$ as $\sum_i b_i p_i^b(\widehat{P}) + \sum_i s_i(1 - p_i^s(\widehat{P}))$. Then define the *consequentialist inefficiency* of the mechanism $M$ (with minimum manipulation cost $\epsilon$) as

$$\eta^c = \sup_{(H^b, H^s)} \left( E_{(H^b, H^s)}[W^{FB}(P) - W^M(\widehat{P}|P)] \right),$$

where the expectation is over true profiles $P$ and reported profiles $\widehat{P}$ obtained by drawing each $(b_i, \widehat{b_i}) \sim H^b$ and each $(s_i, \widehat{s_i}) \sim H^s$ independently, and the supremum is over pairs such that

- $H^b$ is possible for the buyers,
- $H^s$ is possible for the sellers, and
- the marginals $F^b$ of $H^b$ and $F^s$ of $H^s$ on true valuations satisfy $(F^b, F^s) \in \mathcal{F}$.

The value of $\eta^c$ of course depends on $\epsilon$. The higher $\epsilon$ is, the smaller the manipulation sets are, the smaller the set of $(H^b, H^s)$ over which the sup is taken, and so the smaller is consequentialist inefficiency. Note also that the consequentialist inefficiency $\eta^c$ is always at least as large as the truthful inefficiency $\eta$.

This leads to our main definition: we say that a mechanism $M$ has a $(\sigma, \eta)$ *consequentialist tradeoff* on the class of distribution pairs $\mathcal{F}$ if, for any manipulation cost $\epsilon < \sigma$, the mechanism's consequentialist inefficiency on $\mathcal{F}$ is at least $\eta$. This expresses the tradeoff faced by a planner: she must either be willing to assume that agents have a manipulation cost at least $\sigma$, or accept an allocative inefficiency of at least $\eta$.

With these definitions behind us, we can proceed to convert our results into the consequentialist framework. Our earlier results were of the form

> for a given class of distribution pairs $\mathcal{F}$, every mechanism either has inefficiency greater than *[bound]* or susceptibility greater than *[bound]*.

We would now like to have results of the form

> for a given $\mathcal{F}$, every mechanism has a (*[bound]*,*[bound]*) consequentialist tradeoff.

To make this leap, we focus on misreports that are not too small. The intuition is as follows: Suppose a mechanism makes a buyer of value $b$ willing to misreport as a value $\widehat{b}$, and $\widehat{b}$ is far from $b$ — say $\widehat{b} < b$ for example. Then when all other agents report values in between $b$ and $\widehat{b}$, the mechanism cannot distinguish whether the buyer reporting $\widehat{b}$ actually has value $\widehat{b}$ (in which case efficiency would imply that the buyer should not get the good) or actually has value $b$ (in which case the buyer should get the good). So whatever allocation the mechanism specifies will be bounded away from efficiency in one of the two cases.

Once again, the intuition requires some elaboration because of our restriction to continuous distributions — misreports by just a single type, or by a finite set of types, have zero effect on expected efficiency. The technical apparatus needed to make the argument work is as follows.

We define a *quasi-misreport* for a buyer to be a triple $(b, \widehat{b}, \delta)$, where $\delta \in [0, 1]$ and $b, \widehat{b} \in [\delta, 1 - \delta]$, and $\widehat{b} \neq b$. The interpretation of a quasi-misreport is not just that buyers of type $b$ are willing to misreport as $\widehat{b}$, but rather that a positive measure of types $b'$ within $\delta$ of $b$ are each incentivized to misreport by the amount $\widehat{b} - b$. A quasi-misreport for a seller is analogously a triple $(s, \widehat{s}, \delta)$.

Formally: we say that the mechanism $M$ is *$\sigma$-susceptible* to the quasi-misreport $(b, \widehat{b}, \delta)$ of a buyer under $\mathcal{F}$, if the set

$$\{b'_i \in [b - \delta, b + \delta] \mid b'_i + (\widehat{b} - b) \in W^b(b'_i; \sigma)$$

has positive Lebesgue measure. We define $\sigma$-susceptibility to quasi-misreports of a seller analogously.

It is clear that if a mechanism is $\sigma$-susceptible to any quasi-misreport, then it has susceptibility at least $\sigma$. Thus we can think of susceptibility to a particular set of quasi-misreports as a strengthening of susceptibility. This strengthening ties in with

consequentialist inefficiency via the following lemma:

**Lemma 4.1** *Assume $N \geq 2$. Let $\mathcal{F}$ be a set of distribution pairs with $\mathcal{F}_1 \subseteq \mathcal{F}$.*

*If $M$ is $\sigma$-susceptible to the buyer's quasi-misreport $(b, \widehat{b}, \delta)$, and*

$$\eta < \frac{b - \widehat{b} - 4\delta}{64}, \tag{4.1}$$

*then $M$ has a $(\sigma, \eta)$ consequentialist tradeoff over $\mathcal{F}$. Similarly, if $M$ is $\sigma$-susceptible to the seller's quasi-misreport $(s, \widehat{s}, \delta)$, and*

$$\eta < \frac{\widehat{s} - s - 4\delta}{64}, \tag{4.2}$$

*then $M$ has a $(\sigma, \eta)$ consequentialist tradeoff over $\mathcal{F}$.*

**Proof:** We give the proof for (4.1); the argument for (4.2) is essentially identical. Note (4.1) implies $b - \widehat{b} \geq 4\delta$.

Let $R$ be the set of values $r \in [-\delta, \delta]$ such that a buyer of type $b + r$ can benefit by at least $\sigma$ from misreporting as $\widehat{b} + r$, for some distribution pair in $\mathcal{F}$. Thus, for every $r \in R$, the manipulation set $W^b(b + r; \sigma)$ contains $\widehat{b} + r$, and $R$ has Lebesgue measure $\mu > 0$.

Define density functions $f, g$ as follows:

$$f(x) = \begin{cases} 2/\delta, & \frac{b+\widehat{b}}{2} - \delta < x < \frac{b+\widehat{b}}{2} + \delta; \\ 0 & \text{otherwise;} \end{cases}$$

$$g(x) = \begin{cases} 1/\mu, & x = b + r \text{ for some } r \in R; \\ 0 & \text{otherwise.} \end{cases}$$

Define the density $h(x) = \left(1 - \frac{1}{N}\right) f(x) + \left(\frac{1}{N}\right) g(x)$. Let $F, G, H$ be the associated distributions.

Drawing an agent's value from $H$ is equivalent to designating the agent as "$F$-type" or "$G$-type" with probabilities $1 - 1/N$ or $1/N$, respectively, then drawing a valuation from $F$ or $G$ accordingly.

Certainly $(H, H) \in \mathcal{F}$. Suppose all agents' values are drawn independently from $H$, and that the agents report as follows: any $G$-type buyer misreports by $\widehat{b} - b$ (so if his true value is $b + r$, he reports $\widehat{b} + r$); all other agents report truthfully.

Let $E$ denote the event that there is exactly one $G$-type buyer, and the $2N - 1$ other agents are all $F$-type. As in (3.4), we have

$$Pr(E) \geq \frac{1}{16}. \tag{4.3}$$

Conditional on $E$, let $\pi$ be the probability that the $G$-type buyer ends up with a good under the mechanism. Notice that in event $E$, the first-best always requires this buyer to receive a good; when he does not, the resulting efficiency loss is at least as large as the difference between his value and the next-highest value, which is at least

$$(b - \delta) - \left(\frac{b + \widehat{b}}{2} + \delta\right) = \frac{b - \widehat{b}}{2} - 2\delta.$$

Therefore, the consequentialist inefficiency of the mechanism (with minimum manipulation cost $\sigma$) satisfies the lower bound

$$\eta^c \geq Pr(E) \cdot (1 - \pi) \cdot \left(\frac{b - \widehat{b}}{2} - 2\delta\right) \geq \frac{\frac{b-\widehat{b}}{2} - 2\delta}{16} \cdot (1 - \pi). \tag{4.4}$$

On the other hand, let $\widehat{g}$ be the density defined by

$$\widehat{g}(x) = \begin{cases} 1/\mu, & x = \widehat{b} + r \text{ for some } r \in R; \\ 0 & \text{otherwise.} \end{cases}$$

and $\widehat{h}(x) = \left(1 - \frac{1}{N}\right) f(x) + \left(\frac{1}{N}\right) \widehat{g}(x)$. Define $\widehat{G}, \widehat{H}$ the distributions associated with $\widehat{g}, \widehat{h}$. Note that $\widehat{G}$ represents the distribution of *reports* by a $G$-type buyer in the previous scenario who misreports his value.

Suppose now that all agents' values are drawn from $\widehat{H}$, instead of $H$, and that all agents report truthfully. We can label agents as $F$-type or $\widehat{G}$-type, as before. Let $\widehat{E}$ denote the event that there is one $\widehat{G}$-type buyer and all other agents are

$F$-type. We then have $Pr(\widehat{E}) \geq 1/16$ once again. Moreover, the distribution of profiles conditional on $\widehat{E}$, when values are drawn from $\widehat{H}$, is exactly the same as the distribution of *reported* profiles conditional on $E$, when values were drawn from $H$ and when $G$-type buyers were misreporting. Consequently, conditional on $\widehat{E}$, the probability that the $\widehat{G}$-type buyer receives a good is again $\pi$.

However, conditional on $\widehat{E}$, the first-best requires that the $\widehat{G}$-type buyer never receive a good; and when he does receive one, the efficiency loss is at least the difference between his value and the next higher value, which is at least

$$\left(\frac{b+\widehat{b}}{2} - \delta\right) - \left(\widehat{b} + \delta\right) = \frac{b-\widehat{b}}{2} - 2\delta.$$

So we have

$$\eta \geq Pr(\widehat{E}) \cdot \pi \cdot \left(\frac{b-\widehat{b}}{2} - 2\delta\right) \geq \frac{\frac{b-\widehat{b}}{2} - 2\delta}{16} \cdot \pi. \tag{4.5}$$

Adding (4.4) and (4.5), and dividing by 2 gives $\eta^c \geq (b - \widehat{b} - 4\delta)/64$. Combining with (4.1), we see that the mechanism has a $(\sigma, \eta)$ consequentialist tradeoff, as claimed.

□

We can now use Lemma 4.1 to restate our main results from Section 3 in terms of consequentialist tradeoffs. The following theorem extends Theorem 3.3:

**Theorem 4.2** *There exists a positive constant $c$ such that every possible mechanism has a $(c/(\sqrt{N}\log N), c/(\sqrt{N}\log N))$ consequentialist tradeoff on $\mathcal{F}_1$.*

**Proof:** Suppose not: some mechanism $M$ has consequentialist inefficiency $\eta^c$ less than $c/(\sqrt{N}\log N)$ for manipulation cost $\sigma < c/(\sqrt{N}\log N)$. We repeat exactly the steps of the proof of Theorem 3.3. Since $\eta < \eta^c$, the only assumption from that theorem that is no longer present was the assumption that each agent can gain at most $\sigma$ by misreporting. That assumption was used only once in the original proof — in Step 2, in the line "a buyer of type $b_j + r$ (for any $j$) cannot benefit by more than $\sigma$ from misreporting as type $b_{j+1} + r$" (and the analogous argument for sellers). This line now requires elaboration. In particular, it must be reformulated in terms of quasi-misreports.

We claim that for each $j$, the mechanism $M$ cannot be $\sigma$-susceptible to the quasi-misreport $\left(b_j, b_{j+1}, \frac{\gamma}{2}\right)$. For suppose otherwise. Then by Lemma 4.1, we have

$$\eta \geq \frac{b_j - b_{j+1} - 2\gamma}{64} > \frac{b_j - b_{j+1}}{70}.$$

Now, the ratio $\rho$ satisfies the lower bound

$$\rho \geq \frac{3}{4}(20\sqrt{N})^{1/K} \geq \frac{3}{4}(\sqrt{N})^{1/\log N} = \frac{3}{4}e^{1/2} > \frac{6}{5}, \tag{4.6}$$

which gives us

$$b_j - b_{j+1} = (\rho - 1)b_{j+1} \geq \frac{b_{j+1}}{5} \geq \frac{1}{100\sqrt{N}}$$

and therefore

$$\eta \geq \frac{1/100\sqrt{N}}{70} = \frac{1}{7000\sqrt{N}}.$$

Since $\eta < c/(\sqrt{N}\log N)$, we obtain $c \geq 1/7000$ (as long as $N \geq 3$), contradicting our assumption at the beginning of the proof.

Thus, $M$ is not susceptible to the quasi-misreport $\left(b_j, b_{j+1}, \frac{\gamma}{2}\right)$. So it remains true that for *almost* all $r \in [-\gamma/2, \gamma/2]$, a buyer of type $b_j + r$ (for any given $j$) cannot benefit by more than $\sigma$ from misreporting as type $b_{j+1} + r$. Hence, for almost all $r$, this holds for all $j$ simultaneously.

For each such $r$, the argument leading to (3.7) remains valid. Then (3.8) continues to hold as well, since that inequality is derived by integrating over $r \in [-\gamma/2, \gamma/2]$ (and the integrand is bounded). Thus, the conclusion of Step 2 on the average utility of each buyer still applies. An entirely analogous argument shows that we also still have the same lower bound on average utility for the sellers.

From there, the rest of the argument for Theorem 3.3 leads to the same contradiction as before. □

Similarly, the following result extends Proposition 3.4 to the consequentialist framework:

**Proposition 4.3** *There exists a positive constant $c$ such that every mechanism has*

*a* $(c, c)$ *consequentialist tradeoff on* $\mathcal{F}_\infty$.

**Proof:** We prove the proposition with $c = 1/128$. Thus suppose for contradiction that some mechanism has consequentialist inefficiency $\eta^c < 1/128$ with $\sigma < 1/128$. Again, we repeat line-for-line the proof of Proposition 3.4, making a change analogous to the one we applied to prove Theorem 4.2.

Specifically, the only line in the proof of Proposition 3.4 that needs to be changed is the assertion "for any $r \in [0, \gamma]$, a buyer of value $1 - r$ cannot benefit by more than $\sigma$ by misreporting as value $1/4 - r$." Instead, this line now only holds for almost all $r \in [0, \gamma]$; that is, the mechanism is not $\sigma$-susceptible to the quasi-misreport $\left(1 - \frac{\gamma}{2}, \frac{1}{4} - \frac{\gamma}{2}, \frac{\gamma}{2}\right)$. Proof: if it were $\sigma$-susceptible, then by Lemma 4.1, we would have

$$\eta \geq \frac{3/4 - 2\gamma}{64} > \frac{1}{128},$$

contrary to assumption. (An analogous change would be made in the argument for sellers.)

Again, the fact that misreports are prevented for almost all $r \in [0, \gamma]$ rather than all $r$ is immaterial, since the proof of Proposition 3.4 then proceeds by integrating over $r$. The rest of that proof then carries through, and we reach the same contradiction.

□

To summarize this section: although our main results were originally expressed in terms of the tradeoff between incentives for strategic manipulation and efficiency under truth-telling, they can be easily rephrased in terms of the tradeoff between costs of strategic behavior and efficiency under manipulation. The proofs carry over with only minor enhancements needed.

Before closing, we should mention that all of the above discussion has used only the allocation of goods as the relevant welfare criterion. In fact, with our assumption that agents face a cost to behaving strategically, it would arguably be appropriate to count this cost as a welfare loss whenever it is incurred. Of course, doing so would only strengthen our lower bounds on consequentialist inefficiency.

# 5 Onwards

In this paper, we have looked at the tradeoff between efficiency and incentives for strategic manipulation in large double auction mechanisms. In so doing, we have begun to fill a gap between two earlier literatures on large double auctions — one looking only at incentives for manipulation, and one looking at inefficiency in perfectly incentive-compatible mechanisms. By looking at the tradeoff, we have addressed the question of whether it would be possible to achieve much-improved convergence to full efficiency by making a small sacrifice in terms of incentives for truthful behavior. Our main result, Theorem 3.3, gives a negative answer to this question, by providing a near-optimal bound for the rate at which either the inefficiency or the susceptibility to manipulation of any mechanism can converge to zero as the size of the market becomes large. We have also reinterpreted the bound in terms of the severity of inefficiency that may result when agents actually do manipulate (Theorem 4.2).

There are several clear technical directions in which to extend this paper. One direction would be to strengthen Theorem 3.3 to give a sharp bound on the inefficiency-susceptibility tradeoff. Ideally, it should be possible to parameterize the curve to which the inefficiency-susceptibility frontier (depicted in Figure 3.1) converges as $N$ becomes large, analogously to the deficit-inefficiency frontier parameterized by Tatur [30].

One might also wish to give analogous bounds for other classes of distribution pairs $\mathcal{F}$, besides those we have looked at. For example, one might consider the family of all pairs $(F^b, F^s)$ given by continuous densities taking values in some interval $[\underline{p}, \overline{p}]$, where $0 < \underline{p} < \overline{p}$ are fixed; this would be more comparable with previous literature [12, 22, 27, 30].

The present paper fits into the program advanced in the previous chapter of this dissertation, which argues that it can be useful to quantify incentives for strategic behavior in mechanisms, and that a natural approach to doing so — defining a mechanism's susceptibility to manipulation as the maximum expected utility an agent could gain by manipulating — is analytically tractable. By looking at incentives in

this way, rather than treating incentive constraints as rigid, we open up a new quantitative dimension to mechanism design. Understanding this dimension may be useful in designing and evaluating mechanisms for practical use.

# A Omitted proofs

We begin by introducing some asymptotic notation used in the proofs. We follow the conventions of the previous chapter and keep explicit track of constant factors. Specifically, for functions $F(N), G(N)$, we write $F(N) \sim G(N)$ to mean that $F(N)/G(N) \to 1$ as $N \to \infty$, and $F(N) \lesssim G(N)$ to mean $\limsup_{N\to\infty} F(N)/G(N) \leq 1$.

Now, we prove a technical result, Lemma A.2, that is used in the proofs of Propositions 3.1 and 3.2. It provides a central-limit-theorem-style approximation on the probability of a given split between the number of high-value and the number of low-value agents.

We first need the following preliminary calculation:

**Lemma A.1** *Fix $0 < \kappa < 1$. Then*

$$\max_{0\leq J\leq N} \binom{N}{J} \kappa^J (1-\kappa)^{N-J} \lesssim \frac{1}{\sqrt{2\pi\kappa(1-\kappa)N}}.$$

**Proof:** The maximum of the left-hand side over $J$ is attained at $J = \lfloor (N+1)\kappa \rfloor$ (this can be proven by computing the ratio of its values over successive $J$). Now expand explicitly, use Stirling's approximation [1, eq. 6.1.38] for the factorials, and simplify. □

**Lemma A.2** *Let $0 < \kappa < 1$ and $\lambda \geq 1$ be given. There exist a constant $c$ and an integer $N_0$ with the following property: For all $N > N_0$, all $K \leq (1-\kappa)N$, and all $a, b \in [0,1]$ such that*

$$b \leq \lambda a, \qquad 1 - b \leq \lambda(1-a),$$

*we have the inequality*

$$\sum_{J=K}^{N} \binom{N}{J}\binom{N}{J-K} a^{N-J} b^{J-K} (1-a)^J (1-b)^{N-J+K} \leq c\sqrt{\frac{\lambda}{\kappa N}}. \tag{A.1}$$

**Proof:** We consider three cases, depending on the values of $a$ and $b$.

(i) Suppose that $a < \kappa/4\lambda$. Then $b < \kappa/4$. For every $J$ we have either

$$\frac{\kappa}{2\lambda} \leq \frac{N-J}{N} \tag{A.2}$$

or

$$\frac{\kappa}{2} \leq \frac{J-K}{N}, \tag{A.3}$$

since otherwise adding would give $\frac{N-K}{N} < \kappa\frac{1+1/\lambda}{2} \leq \kappa$, a contradiction.

If (A.2) holds then consider

$$\binom{N}{J} a^{N-J}(1-a)^J$$

which is log-concave in $a$, maximized at $a = (N-J)/N$. The constraint $a < \kappa/4\lambda$ then implies

$$\begin{aligned} \binom{N}{J} a^{N-J}(1-a)^J &< \binom{N}{J}\left(\frac{\kappa}{4\lambda}\right)^{N-J}\left(1-\frac{\kappa}{4\lambda}\right)^J \\ &\lesssim \frac{1}{\sqrt{2\pi N\left(\frac{\kappa}{4\lambda}\right)\left(1-\frac{\kappa}{4\lambda}\right)}} \end{aligned}$$

using Lemma A.1.

If (A.3) holds then consider

$$\binom{N}{J-K} b^{J-K}(1-b)^{N-J+K},$$

which is log-concave in $b$, maximized at $b = (J-K)/N$. The constraint $b < \kappa/4$

implies

$$\begin{aligned}\binom{N}{J-K}b^{J-K}(1-b)^{N-J+K} &< \binom{N}{J-K}\left(\tfrac{\kappa}{4}\right)^{J-K}\left(1-\tfrac{\kappa}{4}\right)^{N-J+K} \\ &\lesssim \frac{1}{\sqrt{2\pi N\left(\frac{\kappa}{4}\right)\left(1-\frac{\kappa}{4}\right)}}\end{aligned}$$

again by Lemma A.1.

So there is an absolute constant $c$ such that, for every $J$, one of the two factors

$$\binom{N}{J}a^{N-J}(1-a)^{J}, \qquad \binom{N}{J-K}b^{J-K}(1-b)^{N-J+K}$$

is at most $c\sqrt{\lambda/\kappa N}$ (as long as $N$ is large enough). Then the sum in (A.1) is at most

$$c\sqrt{\frac{\lambda}{\kappa N}}\left[\sum_{J=K}^{N}\binom{N}{J-K}b^{J-K}(1-b)^{N-J+K}+\sum_{J=K}^{N}\binom{N}{J}a^{N-J}(1-a)^{J}\right]$$

$$\leq c\sqrt{\frac{\lambda}{\kappa N}}\left[(b+(1-b))^{N}+(a+(1-a))^{N}\right]=2c\sqrt{\frac{\lambda}{\kappa N}}.$$

(ii) Suppose that $1-a<\kappa/4\lambda$. Then $1-b<\kappa/4$. Here the analysis is quite similar to case (i): For every $J$ we have either

$$\frac{1}{2}\leq\frac{J}{N} \tag{A.4}$$

or

$$\frac{1}{2}\leq\frac{N-J+K}{N}. \tag{A.5}$$

If (A.4) holds then

$$\begin{aligned}\binom{N}{J}a^{N-J}(1-a)^{J} &< \binom{N}{J}\left(1-\frac{\kappa}{4\lambda}\right)^{N-J}\left(\frac{\kappa}{4\lambda}\right)^{J} \\ &\lesssim \frac{1}{\sqrt{2\pi N\left(1-\frac{\kappa}{4\lambda}\right)\left(\frac{\kappa}{4\lambda}\right)}}\end{aligned}$$

and if (A.5) holds then

$$\begin{aligned}\binom{N}{J-K}b^{J-K}(1-b)^{N-J+K} &< \binom{N}{J-K}\left(1-\frac{\kappa}{4}\right)^{J-K}\left(\frac{\kappa}{4}\right)^{N-J+K} \\ &\lesssim \frac{1}{\sqrt{2\pi N\left(1-\frac{\kappa}{4}\right)\left(\frac{\kappa}{4}\right)}}.\end{aligned}$$

This case is completed exactly as in the previous case.

(iii) The remaining possibility is $\kappa/4\lambda \leq a \leq 1-\kappa/4\lambda$. In this case, we hold $a$ fixed and let $N$ and $J$ vary. We use Lemma A.1, which gives

$$\max_J \binom{N}{J}a^{N-J}(1-a)^J \lesssim \frac{1}{\sqrt{2\pi N a(1-a)}} \leq c\sqrt{\frac{\lambda}{\kappa N}}$$

for an appropriate constant $c$. Then the sum in (A.1) is at most

$$c\sqrt{\frac{\lambda}{\kappa N}}\sum_{J=K}^{N}\binom{N}{J-K}b^{J-K}(1-b)^{N-J+K} \leq c\sqrt{\frac{\lambda}{\kappa N}}[(b+(1-b))^N] = c\sqrt{\frac{\lambda}{\kappa N}}.$$

□

**Proof of Proposition 3.1:** We will show the following stronger result: there is an absolute constant $c$ such that the expected value of the least valuable trade, under any distribution pair $(F^b, F^s) \in \mathcal{F}_\lambda$, is at most $c\lambda^{5/2}N^{-1/2}$, as long as $N$ is sufficiently large relative to $\lambda$. Denote this expected value by $\zeta(F^b, F^s)$.

First, fix any $N$ and any $(F^b, F^s) \in \mathcal{F}$. For each $x \in [0,1]$, let $H(x)$ denote the probability that $s_{(k)} < x < b_{(k)}$, where $s_{(k)}, b_{(k)}$ denote the values involved in the lowest-value trade as in Subsection 2.3. Conditional on the realized profile, the value of this lowest-value trade, $b_{(k)} - s_{(k)}$, equals the probability that $s_{(k)} < x < b_{(k)}$ when $x$ is drawn uniformly from $[0,1]$. Hence, the unconditional expected value of $b_{(k)} - s_{(k)}$ is just the expected value of $H(x)$, over $x \sim U[0,1]$. That is,

$$\zeta(F^b, F^s) = E[b_{(k)} - s_{(k)}] = \int_0^1 H(x)\,dx.$$

So it suffices to show that $\max_{x\in[0,1]} H(x)$ is bounded above by $c\lambda^{5/2}N^{-1/2}$.

Thus, fix $x^* \in [0,1]$. Call a valuation *high* if it is in $[x^*,1]$ and *low* if it is in $[0,x^*)$. Notice that $b_{(k)}$ is the lowest buyer's value among the top $N$ values, and $s_{(k)}$ is the highest seller's value among the bottom $N$ values. Therefore, $s_{(k)} < x^* < b_{(k)}$ if and only if all buyers among the top $N$ values are high and all sellers among the bottom $N$ values are low. Call this event $E^*$. Thus, $H(x^*) = Pr(E^*)$.

To bound the probability of $E^*$, we define the following events:

- $E_K$, for each integer $K = -N, -N+1, \ldots, N$, is the event that there are exactly $N+K$ high values.
- $E'_K$, for $K = 0, \ldots, N$, is the event that $E_K$ happens and the $(N+1)$th, $\ldots$, $(N+K)$th highest values are all buyer values.
- $E'_K$, for $K = -N, \ldots, -1$, is the event that $E_K$ happens and the $N$th, $(N-1)$th, $\ldots$, $(N+K+1)$th highest values are all seller values.

Note that $E^*$ is contained in the union of the $E'_K$.

We claim that for $|K| \leq N/2$, $Pr(E_K) \leq c\lambda^{1/2}N^{-1/2}$, where $c$ is an absolute constant (as long as $N$ is large enough). Indeed, if we let $J$ denote the number of high buyer values, we can sum over possible realizations of $J$ to obtain (when $K \geq 0$) the equality

$$Pr(E_K) = \sum_{J=K}^{N} \binom{N}{J}\binom{N}{N+K-J} F^b(\gamma)^{N-J}F^s(\gamma)^{J-K}(1-F^b(\gamma))^J(1-F^s(\gamma))^{N+K-J}. \tag{A.6}$$

A direct application of Lemma A.2, with $\kappa = 1/2$, then implies that $Pr(E_K) \leq c\lambda^{1/2}N^{-1/2}$ as claimed. The argument for the case $K < 0$ is identical.

Next, we claim that

$$Pr(E'_K|E_K) \leq \left(1+\frac{1}{2\lambda^2}\right)^{-|K|}, \qquad \text{for } |K| \leq \frac{N}{2}. \tag{A.7}$$

To show this, we argue in terms of the joint density of the $2N$ values $(b_i, s_i)$. We will again assume $K \geq 0$; the argument for $K < 0$ is identical.

For any weakly decreasing sequence of values $v = (v_{(1)} \geq \cdots \geq v_{(2N)})$ and any sequence of labels $t = (t_{(1)}, \ldots, t_{(2N)})$ with each $t_{(i)} \in \{\mathtt{b}, \mathtt{s}\}$, let

$$Q(v,t) = \prod_{i:\ t_{(i)}=\mathtt{b}} f^b(v_{(i)}) \cdot \prod_{i:\ t_{(i)}=\mathtt{s}} f^s(v_{(i)}).$$

If the buyers' and sellers' values are drawn independently from $F^b$ and $F^s$, then the probability density of a given profile $P$ of values is exactly $Q(v,t)$, where $v$ consists of the values in $P$ sorted in decreasing order, and $t_{(i)} = \mathtt{b}$ if the value $v_{(i)}$ belongs to a buyer and $\mathtt{s}$ if a seller. For any set $T$ of label sequences $t$, let $Q(v,T) = \sum_{t \in T} Q(v,t)$. For $J = 0, \ldots, K$, let $T_J$ be the set of label sequences consisting of $N$ $\mathtt{b}$'s and $N$ $\mathtt{s}$'s, such that exactly $J$ of the labels $t_{(N+1)}, \ldots, t_{(N+K)}$ are equal to $\mathtt{s}$; and let $T_\cup = \cup_{J=0}^{K} T_J$, the set of all label sequences consisting of $N$ $\mathtt{b}$'s and $N$ $\mathtt{s}$'s.

Let $V_K$ be the set of value sequences consisting of $N+K$ high values and $N-K$ low values. Then

$$Pr(E_K) = (N!)^2 \int_{V_K} Q(v, T_\cup)\, dv. \tag{A.8}$$

(The $(N!)^2$ factor comes from the fact that each sequence $v$ of distinct values and label sequence $t$ distinguishing the buyer values from the seller values should be counted multiple times, once for each of the $N!$ possible assignments of buyer identities to buyer values and $N!$ assignments of seller identities to seller values.) Similarly

$$Pr(E'_K) = (N!)^2 \int_{V_K} Q(v, T_0)\, dv. \tag{A.9}$$

On the other hand, for any fixed $v$ and any fixed $J \in \{0, \ldots, K-1\}$, we can relate $Q(v, T_J)$ with $Q(v, T_{J+1})$ as follows. Call an element $t_J \in T_J$ and $t_{J+1} \in T_{J+1}$ *connected* if $t_{J+1}$ is obtained from $t_J$ by switching some $t_{(i)}$ from $\mathtt{b}$ to $\mathtt{s}$, where $i \in \{N+1, \ldots, N+K\}$, and switching some $t_{(j)}$ from $\mathtt{s}$ to $\mathtt{b}$, where $j \notin \{N+1, \ldots, N+K\}$. Each element of $T_J$ is connected to exactly $(K-J)(N-J)$ elements of $T_{J+1}$, and each element of $T_{J+1}$ is connected to exactly $(J+1)(N-K+J+1)$ elements of $T_J$. Moreover, if $t_{J+1}$ is connected to $t_J$, then $Q(v, t_J) \leq \lambda^2 Q(v, t_{J+1})$, since the ratio

between $f^b$ and $f^s$ is always bounded by $\lambda$. Summing over all connected pairs, we have

$$(K-J)(N-J)\sum_{t_J \in T_J} Q(v,t_J) \leq (J+1)(N-K+J+1)\sum_{t_{J+1}\in T_{J+1}} \lambda^2 Q(v,t_{J+1})$$

from which

$$Q(v,T_{J+1}) \geq \frac{(K-J)(N-J)}{(J+1)(N-K+J+1)\lambda^2} Q(v,T_J).$$

Since $N \geq 2K$ and $J \leq K-1$ this gives

$$Q(v,T_{J+1}) \geq \frac{K-J}{J+1}\cdot\frac{1}{2\lambda^2}Q(v,T_J) = \frac{\binom{K}{J+1}}{\binom{K}{J}}\cdot\frac{1}{2\lambda^2}Q(v,T_J).$$

Now by induction we have

$$Q(v,T_J) \geq \binom{K}{J}\cdot\left(\frac{1}{2\lambda^2}\right)^J Q(v,T_0)$$

for all $J$. Summing gives

$$Q(v,T_\cup) = \sum_{J=0}^{K} Q(v,T_J) \geq \left(1+\frac{1}{2\lambda^2}\right)^K Q(v,T_0).$$

Combining with (A.8) and (A.9) gives

$$Pr(E_K) \geq \left(1+\frac{1}{2\lambda^2}\right)^K Pr(E'_K).$$

This is exactly (A.7) for $K \geq 0$. The $K < 0$ case is identical.

In addition, if $K > N/2$, then any draw in $E'_K$ requires the $(N+1)$th, ..., $\lfloor 3N/2 \rfloor$th highest values all to be buyer values; so an identical argument gives

$$Pr(E'_K|E_K) \leq \left(1+\frac{1}{2\lambda^2}\right)^{-\lfloor N/2 \rfloor}$$

for $K > N/2$. And by the same argument, this conclusion also holds when $K < -N/2$.

We conclude that

$$\begin{aligned} Pr(E^*) &\leq \sum_{K=-N}^{N} Pr(E'_K) \\ &\leq \sum_{K=-\lfloor N/2 \rfloor}^{\lfloor N/2 \rfloor} c\lambda^{1/2} N^{-1/2} \cdot \left(1 + \frac{1}{2\lambda^2}\right)^{-|K|} + \sum_{\substack{-N \leq K \leq N \\ |K| > N/2}} \left(1 + \frac{1}{2\lambda^2}\right)^{-\lfloor N/2 \rfloor} \\ &\leq 2\sum_{K=0}^{\infty} \left(1 + \frac{1}{2\lambda^2}\right)^{-K} \cdot c\lambda^{1/2} N^{-1/2} + (N+2)\left(1 + \frac{1}{2\lambda^2}\right)^{-\lfloor N/2 \rfloor} \\ &\leq 7c\lambda^{5/2} N^{-1/2}. \end{aligned}$$

The last inequality holds because $\sum_{K=0}^{\infty}(1+1/2\lambda^2)^{-K} = 2\lambda^2 + 1 \leq 3\lambda^2$, and the final term $(N+2)(1+1/2\lambda^2)^{-\lfloor N/2 \rfloor}$ is exponentially decreasing in $N$, so is certainly at most $c\lambda^{5/2}N^{-1/2}$ when $N$ is large enough.

Thus we have shown that there is an absolute constant $c$ for which $H(x^*) = Pr(E^*) \leq c\lambda^{5/2}N^{-1/2}$ when $N$ is large enough. Moreover, at no step in the proof did we use the specific value of $x^*$ or the distribution $(F^b, F^s) \in \mathcal{F}$; therefore the constant $c$ and the threshold for $N$ are independent of these choices. We conclude that $\sup_{(F^b,F^s)\in\mathcal{F}} \zeta(F^b, F^s) \leq c\lambda^{5/2}N^{-1/2}$, which is what we wanted.

□

Finally, we prove a simple central-limit-theorem approximation used in the proof of Theorem 3.3.

**Lemma A.3** *If $N$ is sufficiently large, then*

$$\sum_{K=0}^{N-\lfloor\sqrt{N}/4\rfloor} \binom{2N}{K} \left(\frac{1}{2}\right)^{2N} \geq \frac{1}{4}.$$

**Proof:** From Stirling's approximation, we have

$$\binom{2N}{K} \left(\frac{1}{2}\right)^{2N} \leq \binom{2N}{N} \left(\frac{1}{2}\right)^{2N} \lesssim \sqrt{\frac{2}{\pi N}}$$

and in particular

$$\binom{2N}{K}\left(\frac{1}{2}\right)^{2N} < \frac{1}{\sqrt{N}}$$

for all $K$, as long as $N$ is large enough. Then, we have

$$\sum_{K=0}^{N-\lfloor\sqrt{N}/4\rfloor}\binom{2N}{K}\left(\frac{1}{2}\right)^{2N} \geq \sum_{K=0}^{N}\binom{2N}{K}\left(\frac{1}{2}\right)^{2N} - \lfloor\sqrt{N}/4\rfloor\frac{1}{\sqrt{N}} \geq \frac{1}{4}.$$

□

# Bibliography


[1] Milton Abramowitz and Irene A. Stegun (1972), *Handbook of Mathematical Functions* (Washington: U.S. Government Printing Office), tenth printing.

[2] George A. Akerlof (1970), "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics* 84 (3), 488-500.

[3] Itai Ashlagi, Mark Braverman, and Avinatan Hassidim (2011), "Matching with Couples Revisited," extended abstract in *Proceedings of the 12th ACM Conference on Electronic Commerce* (EC-11), 335.

[4] Eduardo Azevedo and Eric Budish, "Strategyproofness in the Large as a Desideratum for Market Design," unpublished paper, Harvard University.

[5] Salvador Barberà (2001), "An Introduction to Strategy-Proof Social Choice Functions," *Social Choice and Welfare* 18 (4), 619-653.

[6] Dirk Bergemann and Stephen Morris (2005), "Robust Mechanism Design," *Econometrica* 73 (6), 1771-1813.

[7] Eleanor Birrell and Rafael Pass (2011), "Approximately Strategy-Proof Voting," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 67-72.

[8] Kim-Sau Chung and J. C. Ely (2007), "Foundations of Dominant-Strategy Mechanisms," *Review of Economic Studies* 74 (2), 447-476.

[9] Edward H. Clarke (1971), "Multipart Pricing of Public Goods," *Public Choice* 11 (1), 17-33.

[10] Olivier Compte and Philippe Jehiel (2009), "Veto Constraint in Mechanism Design: Inefficiency with Correlated Types," *American Economic Journal: Microeconomics* 1 (1), 182-206.

[11] Lars Ehlers, Hans Peters, and Ton Storcken (2004), "Threshold Strategy-Proofness: On Manipulability in Large Voting Problems," *Games and Economic Behavior* 49 (1), 103-116.

[12] Thomas A. Gresik and Mark A. Satterthwaite (1989), "The Rate at Which a Simple Market Converges to Efficiency as the Number of Traders Increases: An Asymptotic Result for Optimal Trading Mechanisms," *Journal of Economic Theory* 48 (1), 304-332.

[13] Theodore Groves (1973), "Incentives in Teams," *Econometrica* 41 (4), 617-631.

[14] Nicole Immorlica and Mohammed Mahdian (2005), "Marriage, Honesty, and Stability," in *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (SODA-05), 53-62.

[15] Matthew O. Jackson (1992), "Incentive Compatibility and Competitive Allocations," *Economics Letters* 40 (3), 299-302.

[16] Matthew O. Jackson and Alejandro M. Manelli (1997), "Approximately Competitive Equilibria in Large Finite Economies," *Journal of Economic Theory* 77 (2), 354-376.

[17] Fuhito Kojima and Mihai Manea (2010), "Incentives in the Probabilistic Serial Mechanism," *Journal of Economic Theory* 145 (1), 106-123.

[18] Fuhito Kojima and Parag A. Pathak (2009), "Incentives and Stability in Large Two-Sided Matching Markets," *American Economic Review* 99 (3), 608-627.

[19] Anshul Kothari, David C. Parkes, and Subhash Suri (2005), "Approximately-Strategyproof and Tractable Multiunit Auctions," *Decision Support Systems* 39 (1), 105-121.

[20] Hitoshi Matsushima (2008), "Behavioral Aspects of Implementation Theory," *Economics Letters* 100 (1), 161-164.

[21] Hitoshi Matsushima (2008), "Role of Honesty in Full Implementation," *Journal of Economic Theory* 139 (1), 353-359.

[22] R. Preston McAfee (1992), "A Dominant Strategy Double Auction," *Journal of Economic Theory* 56 (2), 434-450.

[23] Frank McSherry and Kunal Talwar (2007), "Mechanism Design via Differential Privacy," in *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science* (FOCS-07), 94-103.

[24] Roger B. Myerson and Mark Satterthwaite (1983), "Efficient Mechanisms for Bilateral Trading," *Journal of Economic Theory* 29 (2), 265-281.

[25] Philip J. Reny and Motty Perry (2006), "Toward a Strategic Foundation for Rational Expectations Equilibrium," *Econometrica* 74 (5), 1231-1269.

[26] Donald John Roberts and Andrew Postlewaite (1976), "The Incentives for Price-Taking Behavior in Large Exchange Economies," *Econometrica* 44 (1), 115-127.

[27] Aldo Rustichini, Mark A. Satterthwaite, and Steven R. Williams (1994), "Convergence to Efficiency in a Simple Market with Incomplete Information," *Econometrica* 62 (5), 1041-1063.

[28] Mark A. Satterthwaite and Steven R. Williams (2002), "The Optimality of a Simple Market Mechanism," *Econometrica* 70 (5), 1841-1863.

[29] James Schummer (2004), "Almost-Dominant Strategy Implementation: Exchange Economies," *Games and Economic Behavior* 48 (1), 154-170.

[30] Tymon Tatur (2005), "On the Tradeoff between Deficit and Inefficiency and the Double Auction with a Fixed Transaction Fee," *Econometrica* 73 (2), 517-570.

[31] Steven R. Williams (1999), "A Characterization of Efficient, Bayesian Incentive Compatible Mechanisms," *Economic Theory* 14 (1), 155-180.