A Quantitative Approach to Incentives: Application to Voting Rules

Gabriel Carroll, Microsoft Research and Stanford University

gdc@alum.mit.edu

May 20, 2013

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

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

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 ϕ 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 ϕ).

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, 51] 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 ϕ: 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 [33, 36]. 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 companion paper [13] 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 paper. The details of the omitted proofs are in Appendices C through H, which are collected in a separate file available online.

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 [49] and including recent work on matching markets [3, 24, 26, 27]. 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 [9] 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 [4, 28, 31, 32, 38, 52], 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 [46] and has been followed by many authors since [1, 20, 25, 33, 34, 35, 36, 39, 44, 54, 55]. 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 [45] for this approach applied to matching mechanisms). Some of the literature also considers manipulation by coalitions rather than individual voters [29, 30, 47, 48, 50]. The measure used by Campbell and Kelly [12], 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 [6, 7].

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 ϵ 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 ϵ, 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 [3, 11, 26, 27], 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 [8, 14, 60] 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 companion paper on double auctions [13]), 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 ϵ 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 σ 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, 33]).

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 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

[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.

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 $\succ$, $f(N+1 \quad \succ)$ 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 [40] 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* [40] — that, if a Condorcet winner exists (see Subsection 3.2), she should be elected — implies simplicity on every pair of candidates.

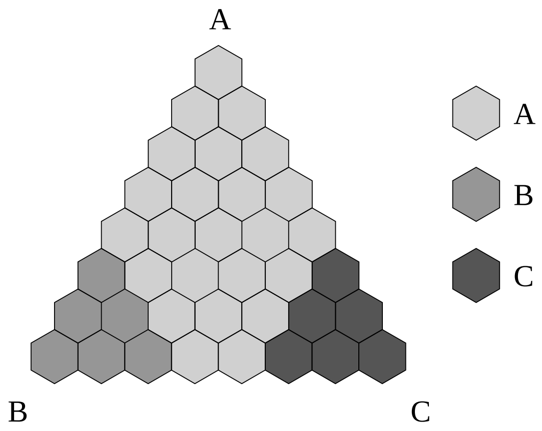

Figure 2.1: A tops-only voting rule

- *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.

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

Following [46], 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]

[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.

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 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_Φ indicates expectation with respect to the distribution Φ for P.

We focus attention on a particular class of beliefs Φ, those for which the other voters' preferences are IID. As argued by McLennan [37], 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 ϕ.

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, σ 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 ϕ.

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 ϕ in the population; but conditional on the realization of ϕ, the opponent-profile P is drawn $IID(\phi)$. Then, for any such belief, the manipulator

[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.

still cannot gain more than σ 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 ϕ, the expected gain from manipulating is at most σ (by definition). So, by the law of iterated expectations, the *unconditional* expected utility gain from manipulating is again at most σ.

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 $\mathrm{cl}(\mathcal{Z})$, and also take the difference inside the expectation:

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

Here the maximum is over the set

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

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

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

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 ϕ) 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 ϕ 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 ϕ may be represented by writing each preference, preceded by its probability. More generally, we 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

$$\mathbf{P}\left(\begin{array}{cc|cc} K_1 & ABC & & \alpha_1\ ABC \\ K_2 & ACB & N; & \alpha_2\ ACB \\ N - K_1 - K_2 & BCA & & \alpha_3\ BCA \end{array}\right).$$

[4] Expression (2.3) is also reminiscent of the notion of *influence* developed by Al-Najjar and Smorodinsky [2]. However, there are some important differences. Influence in [2] is defined with respect to a specific belief ϕ, whereas we take the max over beliefs. The analysis in [2] imposes a noise assumption on ϕ — 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 α 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 ϕ for this same vector thought of as a probability distribution on $\mathcal{L}$ or $\mathcal{C}$.

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 $(K\alpha_1, K\alpha_2, \ldots, K\alpha_r)$ 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 α_i. In particular, the smaller α_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 α_i is attained at $\alpha_i = x_i/K$.*

Lemma 2.4 *The expression*

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

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*

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

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 ϕ_C is the probability (under ϕ) 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 [33, 35, 36], 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 σ_N^{plur} denote the susceptibility of plurality rule.*

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

(b) σ_N^{plur} *satisfies*

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

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 ϕ_C is close to ϕ_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 σ_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 ϕ 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 [10, 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 [1], 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 [1] 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 [1] 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 $A_i \to A_j$. Choose the candidate with the highest score as the winner; break ties alphabetically.

- *Fishburn's* system (also known as the *uncovered set* system [55]): Say that a candidate A_i *covers* another candidate A_j if, for all k such that $A_k \to A_i$, we also have $A_k \to A_j$. (In particular, this requires $A_i \to A_j$.) 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, the statement is that there is some absolute constant c such that $\sigma \gtrsim c/\sqrt[4]{N}$ for all N and M.)

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 [1] 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.

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 ϕ 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 ϕ 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 α. Then, from Lemma 2.2(b), the probability of being pivotal is $\sim (1/2)\sqrt{1/\pi\alpha N}$. In particular, for small α, the probability of being pivotal is high. For these two voting systems, we can construct beliefs with the relevant α 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 ϵ 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

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

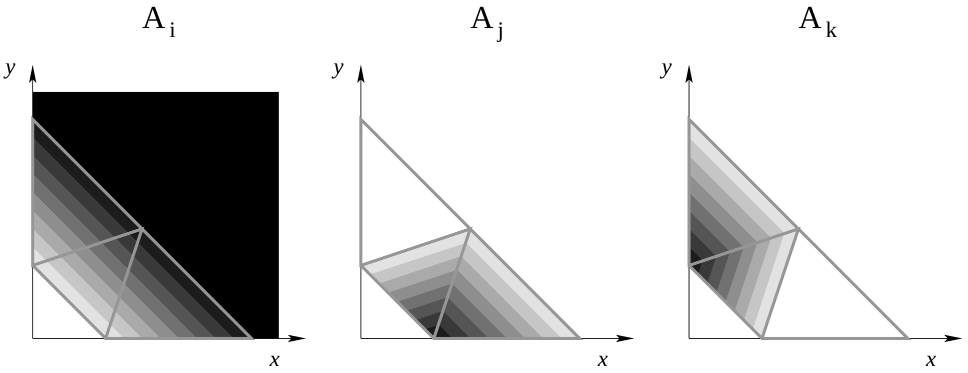

Figure 3.1: Scoring system in case (c) of the pair-or-plurality voting rule. The level plots show what fraction of the $x+y$ points are allocated to each candidate, as a function of x and y. Darker regions represent more points for the candidate indicated. For reference, the gray lines connect the points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

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

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

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.

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 [25] and Mossel and Racz [39] which used a profile-counting measure. (A version of the Gibbard-Satterthwaite theorem for our IID setting was first proved by McLennan [37].)

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.)

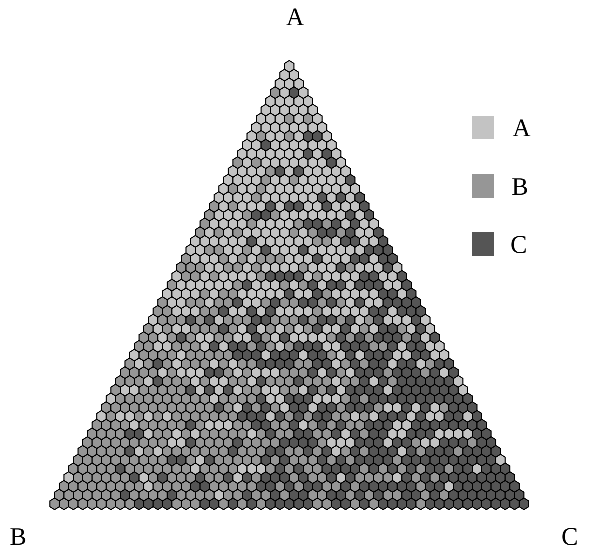

Figure 4.1: The approximate random dictatorship voting rule

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 [33, 36], 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]

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 unanimous, tops-only voting rule that is simple over some pair of candidates has susceptibility $\sigma \geq cN^{-1/2}$.*

In proving all of these lower bounds, we focus on profiles and beliefs ϕ 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

[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.

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	$\sigma \geq \sigma^*_N \quad (\sim cN^{-1/2})$	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 σ. 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.

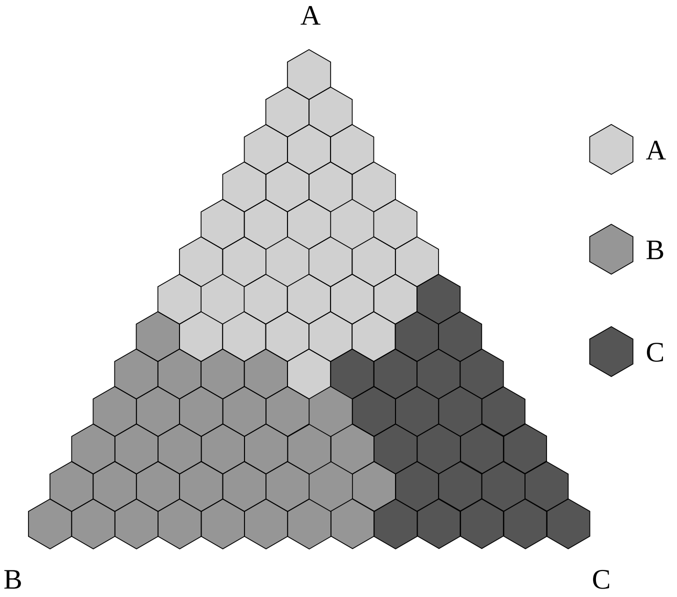

Figure 4.2: A unanimous, monotone, tops-only voting rule

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 ϕ is then chosen to maximize the probability of this pivotal profile, and the bound follows via Lemmas 2.3 and 2.4.

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 ϕ. 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 ϕ. 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*

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

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 σ. There exists a constant c, independent of N (or f or σ), 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 $\mathcal{C}^+$ consisting of the L highest-ranked candidates under $\succ$, for some L, we have

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

(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 $\mathcal{C}'$ of candidates not containing

A_i or A_j,

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

(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,

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

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 expectation on the right side is at most σ. So $\frac{dg}{dx} \leq \alpha\widetilde{N}\sigma$ for all x, hence

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

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 $\mathcal{C}^+ = \mathcal{C} \backslash \mathcal{C}'$. 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 κ, for which the following holds.*

Suppose there are $\widetilde{N}$ voters. 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

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

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

Let A_i, A_j be two different candidates. Suppose that f is a voting rule with susceptibility σ, and the following conditions are satisfied:

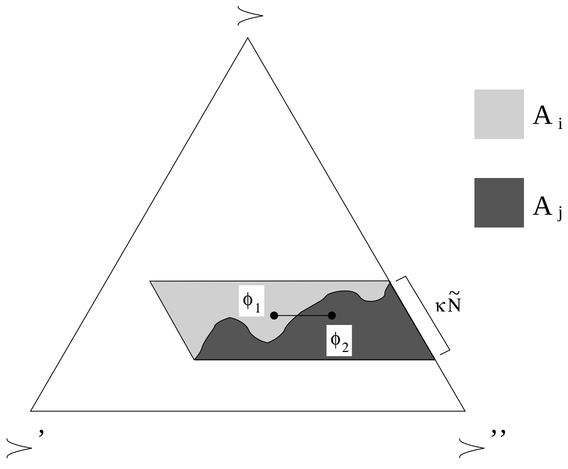

Figure 4.3: Illustration of Lemma 4.10

(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 for N *large enough,*

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

The lemma is proven by identifying two distributions ϕ_1, ϕ_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 the online appendix.)

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.

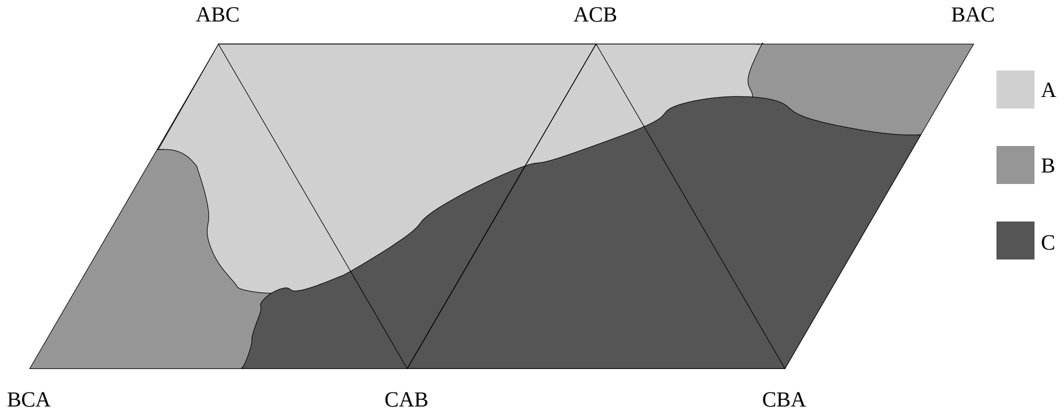

Figure 4.4: Proof of Theorem 4.4

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 σ 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 ϕ_1 and ϕ_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 ϕ_1 and ϕ_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

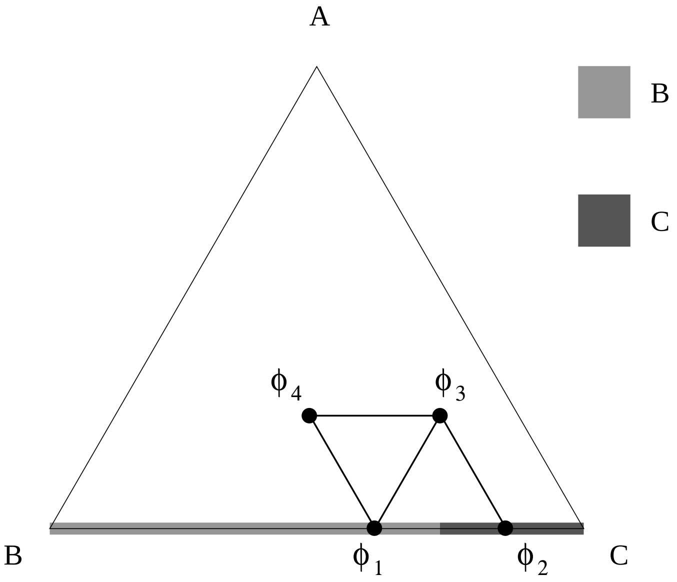

Figure 4.5: Proof of Theorem 4.7

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.

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 γ_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 $\overline{f}(\phi_2) = C$).

Write $\Delta = \alpha_1 - \alpha_2 \leq 2\sqrt{2/\widetilde{N}}$. Then we have $\phi_1 - \phi_2 = \Delta(B - C)$, as an equality of vectors in $\mathbb{R}^M$ (where we have identified each candidate with the corresponding unit vector).

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 ϕ_1 and ϕ_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 ϕ_2 and ϕ_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 ϕ_1 places probability at least Δ on C. If $\widetilde{N}$ is large enough then

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

so this requirement is satisfied.

Applying Lemma 4.9(c) to ϕ_1 and ϕ_4 with $\mathcal{C}' = \{B\}$ gives that $\overline{f}(\phi_4)$ places weight $> 3/4$ on B. Applying Lemma 4.9(c) again to ϕ_3 and ϕ_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 ϕ 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: Viewing each profile P as an integer point in $\mathbb{R}^{M!}$ as usual, we can define a probabilistic voting rule by $\widehat{f}(P) = \overline{f}(P/\widetilde{N})$; the local average lemma implies that if f has low susceptibility then $\widehat{f}$ 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 a linear function of P. 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

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

give the needed linearity.

Although Gibbard's original proof was quite involved, our assumptions of anonymity and weak unanimity make the argument less difficult. (See also [17] and [58] 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:*

- *$\succ_1, \succ_2$ 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 σ, we have the bound

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

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 σ.

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

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

This is easy to show using beliefs along 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*

$$\left|\left(\overline{f}\begin{pmatrix} x & ABC\ldots \\ y+z & BAC\ldots \end{pmatrix} - \overline{f}\begin{pmatrix} x+y & ABC\ldots \\ z & BAC\ldots \end{pmatrix}\right) - \right.$$

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

where c_0 is the constant from Lemma 4.9.

This key step is proven by repeated applications of Lemma 4.11. The bound (4.9) 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 x, y, z, x', z' with $x + y + z = x' + y + z' = 1$, and*

$$\left|\left(\overline{f}\begin{pmatrix} x & ABC\ldots \\ y+z & BAC\ldots \end{pmatrix} - \overline{f}\begin{pmatrix} x+y & ABC\ldots \\ z & BAC\ldots \end{pmatrix}\right) - \right.$$

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

This simply quantifies how much the fact that f is deterministic 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 (4.9) and (4.10) both hold, from which $\sigma \geq c_2/192c_0\widetilde{N}^{3/2}$. □

If we impose the additional requirement of simplicity over candidates A and B, the bound $c_2/\sqrt{\widetilde{N}}$ on the right side of (4.10) 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 [41, 42], 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 companion paper on double auctions [13], which studies the quantitative tradeoff between incentives to manipulate and efficiency in that setting, 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 ϵ. 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 [5, 57, 56]. 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 [59].

The space in between the dominant-strategy and Bayesian approaches — explored by the recent literature on robust mechanism design [8, 14, 60] — 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 ϵ 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 [3, 11, 26, 27], 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 $\succ^*$: $[0,1]^M \rightarrow \mathcal{L}$ 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 $V : \mathcal{C} \times \mathcal{L}^{N+1} \rightarrow \mathbb{R}$ 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 ϵ. 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

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

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

A mathematical state ω has the following interpretation: in this state, if the voting rule in use is f, a voter expects others' votes to follow distribution ϕ, 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 ω_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 ω, 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 ω_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 Ω^*, 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 ψ 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 ω.[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, in state ω, for a voter with utility function u, if he votes $\succ$ and expects others to vote according to ϕ, 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 ω_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 ω, with utility function u, and strategizing cost ϵ. His utility if he truthfully reports preference $\succ^* (u)$, and other

[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.

voters' votes are IID draws from ϕ, 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 ω.

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 ϵ, 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 ξ 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 ϵ.

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 companion paper on double auctions [13] 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 ϵ 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 [10, 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 [43] 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 [10, 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 ϵ. Suppose the manipulator's belief ϕ 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 [10, 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 (σ 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 [10] 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.

References

[1] 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.

[2] Nabil I. Al-Najjar and Rann Smorodinsky (2000), "Pivotal Players and the Characterization of Influence," *Journal of Economic Theory* 92 (2), 318-342.

[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 Barberá (2001), "An Introduction to Strategy-Proof Social Choice Functions," *Social Choice and Welfare* 18 (4), 619-653.

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

[7] 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.

[8] Dirk Bergemann and Stephen Morris (2005), "Robust Mechanism Design," *Econometrica* 73 (6), 1771-1813.

[9] 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.

[10] Steven J. Brams and Peter C. Fishburn (1978), "Approval Voting," *American Political Science Review* 72 (3), 831-847.

[11] Eric Budish (2011), "The Combinatorial Assignment Problem: Approximate Competitive Equilibrium from Equal Incomes," *Journal of Politicial Economy* 119 (6), 1061-1103.

[12] Donald E. Campbell and Jerry S. Kelly (2009), "Gains from Manipulating Social Choice Rules," *Economic Theory* 40 (3), 349-371.

[13] Gabriel Carroll (2011), "The Efficiency-Incentive Tradeoff in Double Auction Environments," in preparation.

[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] 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.

[25] 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.

[26] Fuhito Kojima and Mihai Manea (2010), "Incentives in the Probabilistic Serial Mechanism," *Journal of Economic Theory* 145 (1), 106-123.

[27] Fuhito Kojima and Parag A. Pathak (2009), "Incentives and Stability in Large Two-Sided Matching Markets," *American Economic Review* 99 (3), 608-627.

[28] Anshul Kothari, David C. Parkes, and Subhash Suri (2005), "Approximately-Strategyproof and Tractable Multiunit Auctions," *Decision Support Systems* 39 (1), 105-121.

[29] Dominique Lepelley and Boniface Mbih (1987), "The Proportion of Coalitionally Unstable Situations under the Plurality Rule," *Economics Letters* 24 (4), 311-315.

[30] 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.

[31] Hitoshi Matsushima (2008), "Behavioral Aspects of Implementation Theory," *Economics Letters* 100 (1), 161-164.

[32] Hitoshi Matsushima (2008), "Role of Honesty in Full Implementation," *Journal of Economic Theory* 139 (1), 353-359.

[33] Stefan Maus, Hans Peters, and Ton Storcken (2007), "Anonymous Voting and Minimal Manipulability," *Journal of Economic Theory* 135 (1), 533-544.

[34] Stefan Maus, Hans Peters, and Ton Storcken (2007), "Minimal Manipulability: Unanimity and Nondictatorship," *Journal of Mathematical Economics* 43 (6), 675-691.

[35] Stefan Maus, Hans Peters, and Ton Storcken (2007), "Minimally Manipulable Anonymous Social Choice Functions," *Mathematical Social Sciences* 53 (3), 239-254.

[36] Stefan Maus, Hans Peters, and Ton Storcken (2007), "Minimal Manipulability: Anonymity and Unanimity," *Social Choice and Welfare* 29 (2), 247-269.

[37] Andrew McLennan (forthcoming), "Manipulation in Elections with Uncertain Preferences," *Journal of Mathematical Economics.*

[38] 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.

[39] Elchanan Mossel and Miklos Z. Racz (2011), "A Quantitative Gibbard-Satterthwaite Theorem Without Neutrality," arXiv preprint, `arXiv:1110.5888` .

[40] Hervé Moulin (1988), *Axioms of Cooperative Decision Making* (Cambridge: Cambridge University Press).

[41] Roger B. Myerson (1998), "Population Uncertainty and Poisson Games," *International Journal of Game Theory* 27 (3), 375-392.

[42] Roger B. Myerson (2000), "Large Poisson Games," *Journal of Economic Theory* 94 (1), 7-45.

[43] Richard G. Niemi (1984), "The Problem of Strategic Behavior under Approval Voting," *American Political Science Review* 78 (4), 952-958.

[44] Shmuel Nitzan (1985), "The Vulnerability of Point-Voting Schemes to Preference Variation and Strategic Manipulation," *Public Choice* 47 (2), 349-370.

[45] 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.

[46] Bezalel Peleg (1979), "A Note on Manipulability of Large Voting Schemes," *Theory and Decision* 11 (4), 401-412.

[47] Geoffrey Pritchard and Arkadii Slinko (2006), "On the Average Minimum Size of a Manipulating Coalition," *Social Choice and Welfare* 27 (2), 263-277.

[48] 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.

[49] Donald John Roberts and Andrew Postlewaite (1976), "The Incentives for Price-Taking Behavior in Large Exchange Economies," *Econometrica* 44 (1), 115-127.

[50] Donald G. Saari (1990), "Susceptibility to Manipulation," *Public Choice* 64 (1), 21-41.

[51] 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.

[52] James Schummer (2004), "Almost-Dominant Strategy Implementation: Exchange Economies," *Games and Economic Behavior* 48 (1), 154-170.

[53] Arkadii Slinko (2002), "On Asymptotic Strategy-Proofness of the Plurality and the Run-Off Rules," *Social Choice and Welfare* 19 (2), 313-324.

[54] Arkadii Slinko (2002), "On Asymptotic Strategy-Proofness of Classical Social Choice Rules," *Theory and Decision* 52 (4), 389-398.

[55] David A. Smith (1999), "Manipulability Measures of Common Social Choice Functions," *Social Choice and Welfare* 16 (4), 639-661.

[56] 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).

[57] Yves Sprumont (1995), "Strategyproof Collective Choice in Economic and Political Environments," *Canadian Journal of Economics* 28 (1), 68-107.

[58] Yasuhito Tanaka (2003), "An Alternative Direct Proof of Gibbard's Random Dictatorship Theorem," *Review of Economic Design* 8 (3), 319-328.

[59] 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.

[60] Takuro Yamashita (2011), "A Necessary Condition on Robust Implementation: Theory and Applications," unpublished working paper, Toulouse School of Economics.