Submitted to the Annals of Applied Statistics

GRADATING ARROW'S AXIOMS

BY SUVADIP SANA[1,a], PETER ROCK[2,c] MARTIN T. WELLS[1,b]
AND
MOON DUCHIN[3,d]

[1]*Department of Statistics and Data Science, Cornell University* , [a]*ss2776@cornell.edu*; [b]*mtw1@cornell.edu*

[2]*Cornell Jeb E. Brooks School of Public Policy,* [c]*peter.rock@cornell.edu*

[3]*Data Science Institute, University of Chicago,* [d]*mduchin@uchicago.edu*

Social choice theory is built on the formal/axiomatic study of voting rules, thought of as functions that convert the preferences of a group of voters into a single outcome. Classically, the axioms (often styled as *fairness criteria*) are assessed in a binary manner, so that a voting rule fails the axiom if it fails in even one corner case. Arrow's impossibility theorem—along with numerous variants in the Arrow tradition—tells us that rules that satisfy even a small number of intuitive axioms must amount to dictatorship of a single voter. Many authors have nuanced this by considering *how often* (i.e., on what subsets of preference profiles) a voting rule satisfies some axiom. We set out in a different direction by measuring *how badly* voting rules fail the axioms.

A pass/fail evaluation of axioms masks variation across preference profiles and does not support statistical comparison, uncertainty quantification, or robustness analysis. Treating preference profiles as realizations from an underlying data-generating process, we define continuously gradated measurements of the axiomatic deviations at the profile level. This statistical formulation enables principled empirical comparison of voting rules using standard inferential tools.

Our empirical work analyzes a dataset of over 1000 ranked-choice Scottish local government elections, the 2024 city council elections from Portland, Oregon and the 2025 New York City mayoral primary. We also employ a probabilistic (Bradley-Terry) model designed to produce profiles with varying degrees of polarization. Synthetic experiments let us confirm a finding suggested by the observed elections: the rules that perform most strongly in the Arrow statistical framework are prone to failure for other important democratic norms, like proportional representation.

1. Introduction. The principled aggregation of individual preferences into a collective ranking or decision stands as a foundational challenge across both the classical domain of social choice theory and the modern frontiers of AI and machine learning. From a history centered on systems of election for political representation, the applications have expanded to recommender systems, multi-agent deliberation, and LLM-driven decision pipelines, where the need to reconcile conflicting preferences while preserving structural axioms has become increasingly consequential. Central to this discourse is Arrow's famous impossibility theorem (Arrow, 1950), which establishes that no ranking-based voting rule can simultaneously satisfy three intuitive axioms—Independence of Irrelevant Alternatives (IIA), Unanimity (U), and Non-Dictatorship—when three or more alternatives (candidates) are present.

While the theoretical landscape around Arrow's Theorem has been extensively explored, practical implications have remained less clear. Most studies of voting rules stick with binary criteria—either a rule satisfies IIA as a guarantee over all possible preference profiles, or it does not.[1] This binary framing obscures meaningful variation across real-world instances

Keywords and phrases: Voting theory, Rankings.

[1] A *preference profile*, or simply *profile*, is a record of the ranking of alternatives given by each agent/voter.

and fails to offer a quantitative grasp of how badly or mildly a rule may violate an axiom. This paper offers a new direction: we propose real-valued metrics that quantify violations of the axioms on specific preference profiles, enabling empirics that allow us to give uncertainty quantification, robustness assessment, and to compare voting rules for their degree of compliance.

For example, both score-based voting rules and simple plurality voting violate the IIA axiom in some cases, but the severity and structure of the IIA violations (and the kinds of profile on which they occur) can differ substantially. By shifting from pass/fail fairness conditions to continuous, interpretable metrics that capture the magnitude and profile-specific structure of axiom noncompliance, we can hope to select voting rules that are best adapted to particular uses. We will discuss LLM alignment as an application, but will focus on election of political representatives as the main use case.

1.1. *Contributions.* We introduce a small collection of gradated metrics that score a voting rule f in the range $[0, 1]$ for its level of axiom adherence on a particular preference profile P. Higher scores indicate stronger adherence to the IIA principle (independence of irrelevant alternatives) and a majoritarian principle (pairwise outcomes should be approved by more voters, where possible). We will consider both voting rules f that output a ranking and those that output a winner set—we use the Greek letters ρ for ranking and σ for set.

- $\rho_{\text{IIA}}(f, P)$ for **rankings** and $\sigma_{\text{IIA}}(f, P)$ for **winner sets**: Metrics that capture how much the output for voting rule f on profile P changes when candidates are removed (or "disqualified") from the election.
- $\rho_{\text{UM}}(f, P)$ for **rankings** and $\sigma_{\text{UM}}(f, P)$ for **winner sets**: Metrics that quantify the degree to which the voting outcomes contradict the preference of *most* or even *all* of the voters. A score of $\rho = 0$ or $\sigma = 0$ means that there is some pairwise outcome that is opposed by all of the voters (unanimity failure). A very low value of ρ or σ means that there is some pairwise outcome that is opposed by nearly all of the voters.

Gradated axioms. We define and motivate these gradated IIA and UM metrics for voting rules, then derive their basic properties. They are easily seen to allow a restatement of Arrow's theorem: any voting rule with $\rho_{\text{IIA}} = 1$ and $\rho_{\text{UM}} > 0$ for all profiles must be Dictatorship. If any rule has a lower bound on $\rho_{\text{IIA}}(f, \cdot)$, this provides a *relaxation* of IIA, limiting the amount that rankings can depend on more than pairwise comparisons. On the other hand, a lower bound on $\rho_{\text{UM}}(f, \cdot)$ is a *relaxation* of majoritarianism and at the same time a *strengthening* of unanimity, limiting the share of voters who can disagree with any pair's relative ranking.

Empirical evaluation. We apply our framework to four datasets: (1) a real-world corpus of 1,070 ranked-choice elections for Scottish local government; (2) elections in four districts in Portland's 2024 city council contest; (3) the 2025 New York City mayoral Democratic primary contest; and (4) an inexhaustible supply of synthetic profiles generated under a Bradley–Terry-based model. We find that the voting rule known as *Ranked Pairs*—a strongly Condorcet-consistent method that provably maximizes the UM scores—gives very strong performance for the IIA scores as well. The Borda positional voting rule also has strong performance on these metrics, while plurality (single non-transferable vote) and STV (multi-member ranked choice) receive consistently much lower scores.

Tradeoffs. Finally, we examine the tradeoffs suggested by the empirics. We argue that tolerating lower IIA and UM scores allows voting rules to achieve stronger performance on other normative criteria, especially proportional representation. To explore this tradeoff, we present develop natural techniques:

- A Markov chain to sort candidate lists into slates that are optimized for having the strongest shared support from the most disjoint voter blocs. The scoring function is based on measurements of network modularity.
- A Bradley-Terry parametric model for generating profiles. The parameters let us interpolate from more polarized to less polarized elections, while varying other behavioral assumptions about the voters.

With these tools, we are able to validate the ideas suggested by the real-world elections of Scotland, New York, and Portland: the rules with the strongest performance on the Arrow-style scores have the weakest tendency to deliver proportional representation.

1.2. *Related Work.* Quantifying axiom violations—rather than merely detecting their presence—has emerged as a research direction in multiple domains, including large language models and algorithmic fairness. In this literature, researchers increasingly use real-valued metrics to evaluate departures from normative principles such as stability, internal consistency, or group- and individual-level notions of fairness.[2] Related issues arise in LLM alignment and reinforcement learning with human feedback (RLHF), where preference aggregation plays a central role, but Arrow-like results point to unavoidable tradeoffs. As AI systems and agents are increasingly deployed for deliberative and decision-oriented tasks, the need for principled, interpretable metrics to evaluate axiom adherence has been flagged by leading researchers such as Conitzer et al. (2024).

Classical social choice theory treats preference profiles as deterministic combinatorial inputs. In contrast, the statistical analysis of rankings (Diaconis and Graham, 1977; Diaconis, 1988), models individual rankings as random permutations drawn from a probability distribution on the symmetric group. From this perspective, a preference profile is an empirical sample from an unknown population-level distribution over rankings. In the voting-focused setting of computational social choice theory, the direction has been different. Many papers offer quantified Arrow-style theorems, but they tend to focus on the *probability* of rule violations under stylized types of random voting. A lineage that starts with Kalai and was developed by Mossel and collaborators uses this setup to prove structure theorems, showing that if the probability of a rule violation is near zero, then the voting rule is close to dictatorship (Kalai, 2002a,b; Mossel, O'Donnell and Servedio, 2005; Mossel, 2012). Procaccia and others also use probabilistic framing, particularly by considering average-case manipulability (Procaccia and Rosenschein, 2007). In those and subsequent papers, the random voting models are often approximately uniformly distributed over the permutations (*impartial culture* plus possible noise) or approximately constant (*Mallows models*) (Diaconis, 1988). Social choice researchers have also considered the frequency of rule violations in real elections for axioms like monotonicity (McCune and Graham-Squire, 2024). In economics, researchers such as Dougherty and Heckelman (2020) analyze the frequency of IIA violations under simulated or real-world profiles, contrasting relatively stable Borda and relatively unstable Plurality, for instance. In all of these papers, the axioms remain binary and quantification is external to the axiom.

Passing to relaxed axioms also has a long pedigree, such as with relaxations of the Pareto property, monotonicity, transitivity, and especially the use of restricted voting domains. In one very recent example, Eric Maskin introduced an IIA modification called MIIA (Maskin,

[2]In particular, we found two specific loss functions in the AI literature that function as relaxations of IIA Zhao, Wang and Peng (2024); Hornischer and Terzopoulou (2025). The similarity score in Zhao et al. uses edit distance in place of swap distance, but is otherwise similar to our σ_{IIA} score. Hornischer–Terzopoulou define an *independence loss function* that measures the change in output rankings when the input rankings are randomly perturbed while leaving a given pair of candidates fixed.

2025). Where IIA requires that whether $A \succ B$ in the output ranking depends only on the A vs. B preference of individual voters, MIIA also allows for dependence on the number of candidates ranked *between* them, thought of as an indicator of preference strength. However, as with the other relaxed axioms, satisfying MIIA remains a binary yes/no condition.

Social choice does encompass a major theme of measuring suboptimality, going under the heading of *distortion*, but in the main it is not tied to axioms. The main thrust of the distortion literature is to measure how much a social choice output falls short of an ideal outcome, whether in terms of utility or of distance in a latent metric space that tracks voter preferences. The reliance on latent positions or utilities means that distortion cannot be computed directly from ranked preference profiles (which is the classical setting for Arrow's Theorem). Within this literature, the work of Delemazure, Lang and Pierczyński (2024) takes on the IIA axiom specifically, within the spatial and utility setting.

Here, we provide a voting-specific study centered on the magnitude of axiom violations. By considering the degree of violation when specific voting rules are applied to specific profiles, we can keep our view focused on the relative merits of the voting rules—and this in turn provides a better understanding of the axioms.

The remainder of the paper is organized as follows. Section 2 reviews the social choice background and introduces the needed notation. Section 3 defines the gradated axioms, analyzes their properties, and articulates the connection to Arrow's Theorem. Sections 4-5 present empirical results on observed and synthetic data, respectively. Finally, Section 6 concludes with discussion, limitations, and directions for future work.

2. Background.

2.1. *Notation for elections, preferences, and preference profiles.* First, we give some notation for the elections we will consider. We let $\mathcal{V}$ be the set of n voters and $\mathcal{C}$ the set of m candidates. Each voter will offer a complete or partial ranking of the candidates as their *ballot*, and a collection of ballots is a *preference profile* P. We will write $S(\mathcal{C})$ for the set of permutations of the candidates, which is a copy of the symmetric group S_m, and we will write $\hat{S}(\mathcal{C})$ for the extended set of partial rankings. For instance, if $\mathcal{C} = \{A, B, C\}$, then (A, B, C) is an example of a complete ranking and (A) is a partial ranking, both valid ballots.[3] We will adopt the notation $A \succ_{i,P} B$ to mean that voter i ranks A above B in profile P, or simply $A \succ_i B$ when P is understood. Then we write $A \asymp_i B$ to mean that the voter leaves both unranked. Then $A \succeq_i B$ means $A \succ_i B$ or $A \asymp_i B$. Similarly, $A \succ_{f(P)} B$ indicates that when the voting rule is applied to the profile P, the outcome ranking has A over B. In order to measure the difference between rankings, a natural choice of metric is the Kendall tau (Diaconis, 1988, Ch. 6) or swap distance, which we will denote by d_{swap}. This measures the number of neighbor-swaps necessary to transform one ranking to another, and extends naturally to partial rankings.

The set of preference profiles (or simply *profiles*) for elections on $\mathcal{V}, \mathcal{C}$ is given by $\mathcal{P} := \hat{S}(\mathcal{C})^{\mathcal{V}}$. With this notation, a *ranking voting rule* (or just ranking rule) is a function $f : \mathcal{P} \to S(\mathcal{C})$. That is, even though the ballots are allowed to be partial rankings, for us the voting rule must output a complete ranking in a deterministic way, such as by using a tie-breaking procedure. A *winner-set voting rule* is a function $f : \mathcal{P} \to 2^{\mathcal{C}}$. (Here $2^{\mathcal{C}}$ is the power set: the set of all subsets of $\mathcal{C}$.) A *k-winner voting rule* is a function $f^{(k)} : \mathcal{P} \to \binom{\mathcal{C}}{k}$ that outputs a winner set of a specified size k.

[3] We do not allow ties, except implicitly in the sense that partial ballots may be read as giving equal rankings to the non-mentioned candidates. This is in keeping with real-world ranked-choice voting.

Below, we will slightly abuse notation by using a common name to refer to voting rules described in a general enough way to apply to any number of voters and candidates. For instance, the *plurality* rule ranks all candidates in order of their number of first-place votes, or outputs the top k in that list, depending on context.

DEFINITION 2.1 (Misalignment). For a given voting rule and preference profile, the *misalignment score* considers all pairwise comparisons of candidates and reports the lowest share of voters agreeing with a pairwise outcome.

For a ranking rule f, we set

$$M(f,P) := \min_{A \succ_{f(P)} B} \frac{\#\{i \in \mathcal{V} : A \succ_i B\}}{n}. \tag{1}$$

Similarly, for a winner-set rule f, we define

$$M(f,P) := \min_{\substack{A \in f(P) \\ B \notin f(P)}} \frac{\#\{i \in \mathcal{V} : A \succ_i B\}}{n}. \tag{2}$$

This M finds the worst alignment of the voters with the outcome. In the case of complete rankings by voters, this flags an anti-majoritarian outcome if $M < \frac{1}{2}$.

DEFINITION 2.2 (Candidate removal). Given a profile $P \in \mathcal{P}$ or a ballot $\beta \in \hat{S}(\mathcal{C})$ and a candidate $C \in \mathcal{C}$, we write P^C or β^C to represent the condensed profile or ballot on $\mathcal{C}' = \mathcal{C} \setminus \{C\}$, i.e., the candidate set without C.

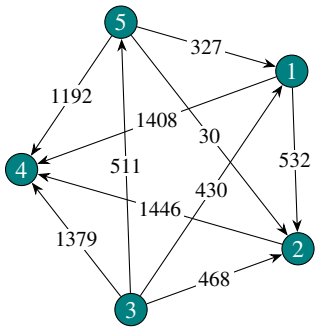
Finally, we record the margins of victory in a complete graph on the candidates and we employ a definition from graph theory that picks out rankings of candidates that respect majority preferences.

DEFINITION 2.3 (Pairwise comparison graph and topological sort). For any profile P, the *pairwise comparison graph* (abbreviated PWCG, also known as the tournament graph) is a directed weighted graph $G(P)$ with vertex set $V = \mathcal{C}$ defined by putting a directed edge from A to B, weighted by the margin by which voters prefer A to B. (That is, the arrow is directed from the more popular to the less popular candidate, with non-negative weight $\#\{i : A \succ_i B\} - \#\{i : B \succ_i A\}$. We take the convention that a tie is represented by a zero-weight edge in each direction.)

We say that a complete ranking of vertices, $\pi \in S(V)$, is a *topological sort* of a directed graph $G = (V, E)$ if $(u, v) \in E \implies u \succ v$ in π. We will similarly call a ranking of candidates a topological sort of a profile if that condition applies to the associated graph. Thus a graph has a topological sort if it has no directed cycles (which by our definitions means there are no exact ties between candidates), and if a topological sort exists, it is unique. Note that this is stronger than simply having a Hamiltonian path, because it puts requirements on the other arrows.

Recall that a *Condorcet candidate* in a profile P is a candidate who beats all others head-to-head, so that all the directed arrows in the PWCG point outwards from the associated node. Not all profiles have such a candidate, but it is often claimed that real-world preferences are overwhelmingly likely to have one, as in Figure 1. We will also refer to a topological sort as a *Condorcet ladder*, since it has the property that the first candidate A in the order is preferred head-to-head to all others (and is thus a Condorcet candidate in P); the second candidate is preferred head-to-head to all but the first (and is thus a Condorcet candidate in P^A), and

so on, until the last candidate in the order loses every head-to-head comparison. A profile with a Condorcet ladder will be called a *ladder profile*. We will call a ranking rule *strongly Condorcet consistent* if it returns the (unique) topological sort for any ladder profile.[4] Figure 1 shows the PWCG for an election in Renfrewshire, Scotland.[5]



1: Edward Grady (Labour)
2: Kate Hughes (Labour)
3: Cathy McEwan (SNP)
4: Dale Nelson (Conservative)
5: Jim Paterson (SNP)

FIG 1. *The pairwise comparison graph for the Scottish local government election in Renfrewshire Ward 2, 2022, where 3761 voters ranked five candidates in order to select three winners. In this case, Cathy McEwan is the Condorcet candidate, preferred head-to-head to all others, and there is a topological sort (total Condorcet order)* $3 \succ 5 \succ 1 \succ 2 \succ 4$. *In the actual STV election, McEwan (3) crossed threshold first, followed by Grady (1) and then Paterson (5). Below, Table 2 shows the results when various ranking rules are applied.*

2.2. *Voting rules.* In this paper, we focus on a small menu of common ranking-based voting rules. The first several rules are in the class called **scoring rules**: a score vector $\mathbf{s} = (s_1, s_2, \ldots, s_m)$ is applied, so that first-place votes are worth s_1 points, second-place votes are worth s_2, and so on. This can be used to output a ranking of the candidates in order of their total score, or to output a k-winner set by taking the top k scores. For simplicity, we adopt the convention that candidates who are unranked in a given ballot receive no points from that ballot. Table 1 summarizes the scoring rules compared below.[6]

Voting rule	Score vector $\mathbf{s}$	Description
Borda	$(m-1, m-2, \ldots, 0)$	Linear scoring.
3-Approval	$(1, 1, 1, 0, \ldots, 0)$	Top three get a point.
2-Approval	$(1, 1, 0, 0, \ldots, 0)$	Top two get a point.
Plurality	$(1, 0, 0, 0, \ldots, 0)$	Only first choice gets a point.

TABLE 1
Scoring rules.

Next, we compare these with another voting rule that cannot be described with a score vector, but rather operates iteratively in a round-by-round fashion. Single transferable vote (or **STV**) is a multi-winner voting rule for a fixed number k of seats, designed to output a set of k winners. A *threshold of election* is also fixed—often, this is the so-called *Droop quota* $n/(k+1)$—as the amount of support needed to be elected. If any candidate has strictly more than that level of first-place support in the profile P, they are elected. If no candidate

[4]Such rules are called *pairwise-majority consistent* by Caragiannis, Procaccia and Shah (2013), but at least some other authors use the strong Condorcet terminology in the way that we do here (Lamboray, 2006).

[5]The raw preference profile can be found at github.com/mggg/scot-elex.

[6]The Plurality rule discussed here is also called SNTV or single non-transferable vote, not to be confused with Plurality Block voting, which allows voters to cast as many votes as there are seats to be filled.

does, then proceeding round by round, we eliminate the candidate with the least first-place support and transfer their support (with full weight) to the next choice of their voters. When a candidate is elected by surpassing threshold, their *surplus* support is transferred to the next choice of their voters; for instance, if they received 120% of the threshold, then their votes are transferred with weight $20/120 = 1/6$. Though STV is designed as a multi-winner system, we can convert the output to a ranking by filling in winners from the top in order of election, and filling in eliminated candidates from the bottom in order of elimination. Any candidates still left once the seats have been filled are placed in between, in order of first-place votes when the process terminates.

Finally, we consider a strongly Condorcet consistent rule that also takes margins of victory into account. The **Ranked Pairs** voting rule uses the PWCG to output a ranking as follows. Start with all the directed edges turned "off" (inactive). Then sequentially activate edges in order of their margin of victory, rejecting only those that form a directed cycle with other already activated edges. This necessarily produces a cycle-free directed forest. If G had no cycles, then the process activates the path witnessing the topological sort; return that ranking. Otherwise, return the lexicographically first ranking consistent with the arrows.[7]

None of the other rules discussed here (Borda, 3-Approval, 2-Approval, Plurality/SNTV, and STV) is strongly Condorcet consistent; indeed, none of them is even guaranteed to top-rank a Condorcet candidate (one preferred head-to-head to all others) when one exists.

2.3. *Observed and synthetic elections.* We apply our framework to a real-world dataset of Scottish ranked-choice elections, comprising over 1000 local elections.[8] For administrative purposes, Scotland is divided into several hundred wards, each conducting a ranked choice election by STV for its local government every five years since 2012. An example is shown in Figure 1 with the candidates listed as they appeared on the ballots (alphabetically by last name) and the PWCG displayed at right.

3. Gradated axioms and their properties. We propose real-valued metrics that are structured to make it possible to express the classical axioms. These metrics can be evaluated when voting rules are applied to particular preference profiles.

3.1. *Independence of Irrelevant Alternatives (IIA).* We first extend the axiom of Independence of Irrelevant Alternatives (IIA). Informally, a voting rule satisfies IIA if the relative position of any two candidates in the social ranking only depends on their relative ranking by individual voters.

A ranking voting rule f satisfies IIA if for any two profiles $P, P' \in \mathcal{P}$ and any $A, B \in \mathcal{C}$,

$$\{i \in \mathcal{V} : A \succ_i B \text{ in } P\} = \{i \in \mathcal{V} : A \succ_i B \text{ in } P'\} \implies \left(A \succ_{f(P)} B \iff A \succ_{f(P')} B\right).$$

Equivalently, f satisfies IIA iff $f(P^C) = f(P)^C$ for all $P \in \mathcal{P}, C \in \mathcal{C}$.

DEFINITION 3.1 (Gradated IIA for rankings). We define $\rho_{\text{IIA}}(f, P)$ to measure the failure of IIA:

$$\rho_{\text{IIA}}(f, P) := 1 - \frac{\sum_{C \in \mathcal{C}} d_{\text{swap}}(f(P^C), f(P)^C)}{m\binom{m-1}{2}}.$$

[7]For further details, see the documentation for the VoteKit Python package, which was used in the empirical sections of this paper (Data and Democracy Lab, 2024).

[8]This dataset was collected by David McCune and published in a format compatible with VoteKit by the Data and Democracy Lab. It can be found at https://github.com/mggg/scot-elex.

Note that the largest possible swap distance between vectors of length m comes from total reversal, which gives $d_{\text{swap}} = \binom{m-1}{2}$. From this characterization, we obtain some immediate consequences.

REMARK. The IIA property is equivalent to $\rho_{\text{IIA}}(f, P) \equiv 1$ over $P \in \mathcal{P}$. (See Proposition A.1 for a short proof.) The meaning of $\rho_{\text{IIA}}(f, P) = 1 - \alpha$ is that, on average, disqualifying a candidate alters the ranking of the remaining candidates by a fraction α of a complete reversal. In particular, $\rho_{\text{IIA}}(f, P) = 0$ iff $f(P^C)$ is the complete reversal of $f(P)^C$ for every candidate $C \in \mathcal{C}$.

DEFINITION 3.2 (Gradated IIA for winner sets). Let f be a voting rule which outputs a winner set $W \subset \mathcal{C}$. We will sometimes use the notation $f^{(k)}$ to emphasize that the winner set is specified to have size k. We define the winner-set variation of the IIA score by

$$\sigma_{\text{IIA}}(f^{(k)}, P) := \frac{1}{m}\left[\sum_{C \in f^{(k)}(P)} \frac{|f^{(k-1)}(P^C) \cap f^{(k)}(P)^C|}{k-1} + \sum_{C \notin f^{(k)}(P)} \frac{|f^{(k)}(P^C) \cap f^{(k)}(P)^C|}{k}\right].$$

If $k = 1$, we just average over the $m - 1$ non-winner disqualification terms.

This is designed to give a full score ($\sigma = 1$) when the removal of each candidate C is minimally disruptive of the winner set. Removing a winner C should lead to the other winners still succeeding when $k - 1$ are selected from a pool without C. Removing a non-winner C should lead to no change to the winner set.

An easy observation shows that any strongly Condorcet consistent rule must satisfy IIA on the restricted domain of ladder profiles. This is because disqualifying candidates removes their nodes from the PWCG but the remaining arrows and margins stay the same, so both $f(P^C)$ and $f(P)^C$ prioritize candidates in order of the topological sort.

3.2. *Unanimity (U) and majority.* The other axiom we extend is the unanimity criterion (U), and we do so by recognizing it as one pole of a spectrum with respect to misalignment. Informally, a voting rule f satisfies U if, in the case that P contains a unanimous preference for some candidate over another one, the output $f(P)$ respects that preference. Note that this says nothing about profiles in which there are no unanimous preferences—which means this axiom is silent in most real-world elections. This motivates us to create a metric that also gives nontrivial information in cases where majority preferences are not unanimous.

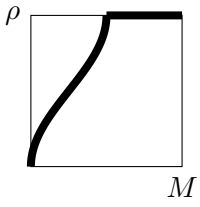
DEFINITION 3.3 (Gradated unanimity and majoritarianism). A voting rule f satisfies U if for any $P \in \mathcal{P}$ and $A, B \in \mathcal{C}$, if $A \succ_i B$ for all voters $i \in \mathcal{V}$, then $A \succ_{f(P)} B$.

Equivalently,

$$f \text{ satisfies U} \iff M(f, P) > 0 \quad \forall P.$$

Then we can define a metric of majority alignment that extends the unanimity criterion.

$$\rho_{\text{UM}}(f, P) := \begin{cases} \frac{2}{\pi} \arcsin \sqrt{2M}, & M(f, P) < 1/2 \\ 1, & M(f, P) \geq 1/2, \end{cases}$$



where M is as defined in (1). The definition for σ_{UM} is identical but with M as in (2).

Any monotonic function $[0, 1/2] \to [0, 1]$ could be used to relate ρ and σ to M, and we selected the arcsin for two reasons. First, an arcsin transformation is frequently used in statistics applications for variance stabilization for binomial data (Anscombe, 1948). By controlling heteroskedasticity of the data, it produces statistics that are suitable for simple averaging, without an additional dispersion weighting step. Secondly, the vertical tangent (infinite slope) of ρ and σ as $M \searrow 0$ and $M \nearrow 1/2$ also creates sharp distinctions in the scores near those critical values.

This is set up so that $\rho_{\text{UM}}(f, P) = 0$ when P is a witness to the failure of the unanimity criterion. In other words, $\rho_{\text{UM}}(f, P) = 0$ exactly means that P contains a unanimous preference between some two candidates and $f(P)$ ranks them in reverse. By contrast, the definition makes it easy to see that $\rho_{\text{UM}}(f, P) = 1$ precisely means that every majority preference in P is respected. (And clearly this implies that every unanimous preference is respected.) Similar statements hold for $\sigma_{\text{UM}}(f, P)$.

This description also gives a natural interpretation to scores less than one. If $\rho_{\text{UM}} < 1$, then $M < 1/2$, and there is some pair of candidates where only M share of voters agrees with the output ranking, and M is monotonically related to ρ by $M = \frac{1}{2}\sin^2(\pi\rho/2)$.

REMARK. Recall that Arrow's impossibility theorem (Arrow, 1950) asserts that no ranking rule can satisfy IIA, U, and non-dictatorship simultaneously. This gives us the following statement, visualized in Figure 2.

Visual Arrow's Theorem:

A ranking voting rule f is a Dictatorship if and only if $\rho_{\text{IIA}}(f, P) = 1$ and $\rho_{\text{UM}}(f, P) > 0$ for all profiles P.

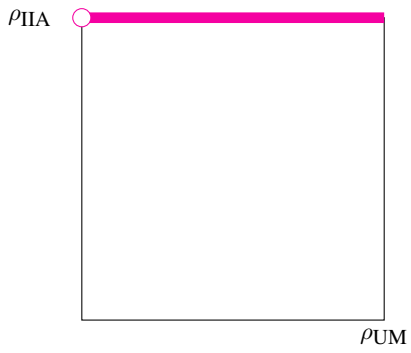


FIG 2. *In Arrow's framework, the half-open line ($\rho_{\text{UM}} > 0, \rho_{\text{IIA}} \equiv 1$) is desirable, but forces Dictatorship.*

Importantly, there are rules that always have $\rho_{\text{UM}} > 0$ (like Borda) but there are no rules that have $\rho_{\text{UM}} \equiv 1$. In any profile with a tie-free Condorcet cycle, some comparison of a pair of candidates must displease a majority, so $\rho_{\text{UM}}(f, P) < 1$ on such a profile for any f. Ranked Pairs is explicitly constructed to displease the smallest possible majorities. Thus $(\rho_{\text{UM}}, \rho_{\text{IIA}}) = (1, 1)$ can be a design goal for voting rules, but no rule, not even Dictatorship, can always attain it.

Appendix A goes into greater detail on some of the choices made in the construction of the ρ, σ metrics.

4. Empirical results.

4.1. *Scottish elections.* In Table 2, we return to the 2022 Scottish election of the Renfrewshire Ward 2 first shown in Figure 1.

	Borda	**3-App**	**2-App**	**Plurality**	**STV/IRV**	**Ranked Pairs**
1st	3	3	3	3	3	3
2nd	5	2	5	1	1	5
3rd	1	5	1	5	5	1
4th	2	1	2	4	4	2
5th	4	4	4	2	2	4
ρ_{IIA}	0.93	0.90	0.93	0.80	0.73	1
ρ_{UM}	1	0.75	1	0.44	0.44	1

TABLE 2

Using the preferences from Renfrewshire Ward 2, 2022, we compute the output rankings and the ρ metrics under the different systems of election. Borda, 2-Approval, and Ranked Pairs all return the topological sort, respecting all pairwise majority preferences. Plurality and STV give the same output ranking, which ensures that their ρ_{UM} values are equal, but they perform differently when candidates are removed, leading to different ρ_{IIA}.

Next, we sweep our scores over the full Scottish dataset and show the results in Figures 3–5. Unsurprisingly, the strongly Condorcet consistent rule (Ranked Pairs) dominates the field of voting rules in terms of these scores. This is in large part because the great majority of Scottish preference profiles are ladder profiles, with a total Condorcet order on candidates. This occurs outright in 1025 out of 1070 elections. The other 45 contests have a Condorcet cycle, but the cycles depends on an exact tie in 5 cases. (See also Supplementary Figure 10.) Putting these together, a voting rule with strong Condorcet consistency and consistent tie-breakers guarantees $(\rho_{\text{UM}}, \rho_{\text{IIA}}) = (1, 1)$ in 1030/1070 cases, about 96.3% of the time. Though this does suggest that Ranked Pairs is an unequivocally strong choice of voting rule, we will see below that this dominant performance on Arrow scores comes at a cost with respect to other norms.

The figures compare the six indicated voting rules, either used in ranking form to compute ρ scores or in winner-set form to compute σ scores. In the winner-set case, the number of winners is set equal to its real-world value from which the data was drawn. A seventh column shows the results of a random voting rule. This is executed by choosing a uniformly random permutation of the candidates as the output ranking for ρ scoring, and the top k names from that ranking as the winner set for σ scoring.

In Figure 3, the data is separated out by the number of candidates, in order to examine how the normalization handles the size of the candidate pool. The six- to nine-candidate contests make up 812 out of the 1070 available elections. There is a visible decay of most of the scores as the candidate pool grows, but the effect is quite slight.

Figures 4 and 5 show the scatterplot of the IIA score against the UM score over all 1070 elections, first for rankings output and then for winner-set versions.

Appendix C reports bootstrap confidence intervals quantifying the uncertainty of the scores under each ranking rule. We are able to confirm that, up to overlapping intervals, the four scores $\rho_{\text{IIA}}, \rho_{\text{UM}}, \sigma_{\text{IIA}}, \sigma_{\text{UM}}$ sort the voting rules in the same order on the Scottish data.

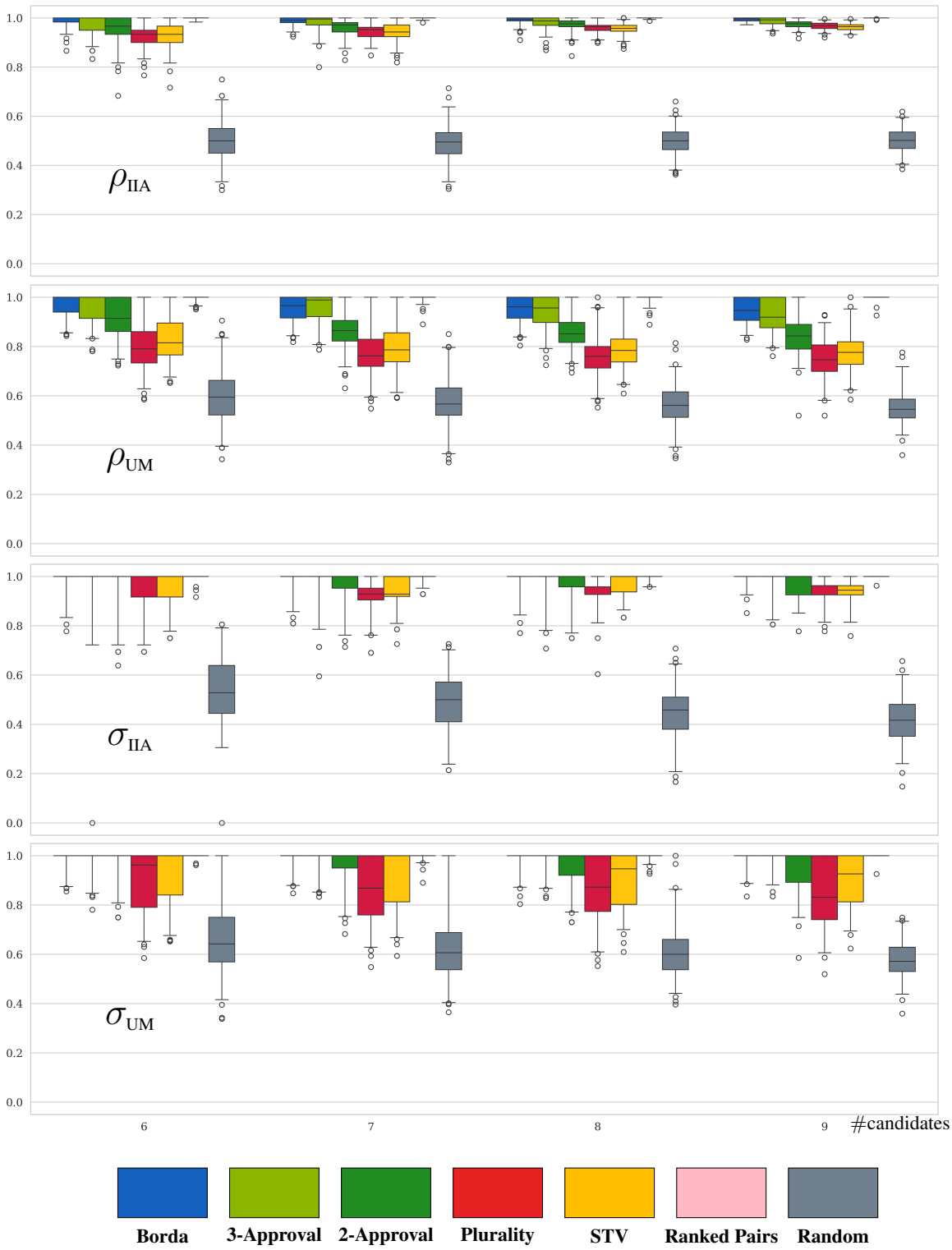


FIG 3. *The UM and IIA scores are run on the Scottish ranked choice elections with 6 candidates (205 elections), 7 candidates (282 elections), 8 candidates (207 elections), and 9 candidates (118 elections). The boxes show the 25th-75th percentile values of the scores, and the whiskers run from the 1st-99th percentiles. The Ranked Pairs method scores so high on these metrics that the boxes are essentially invisible. Random rankings are shown for contrast.*

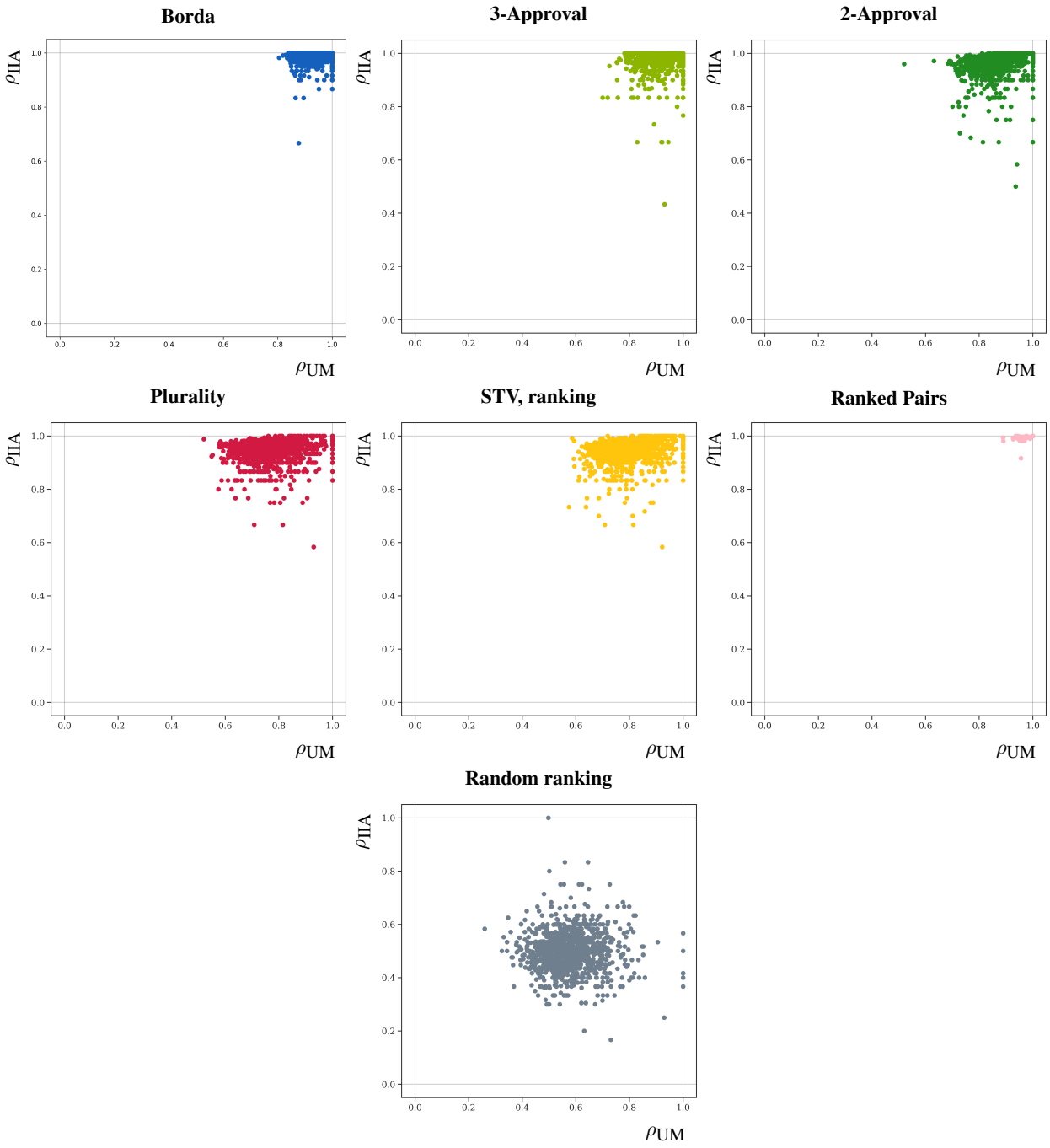


FIG 4. *Scatterplots show the results for* ρ_{UM} *vs* ρ_{IIA} *when various ranking rules are applied to preference profiles from Scottish local government elections.*

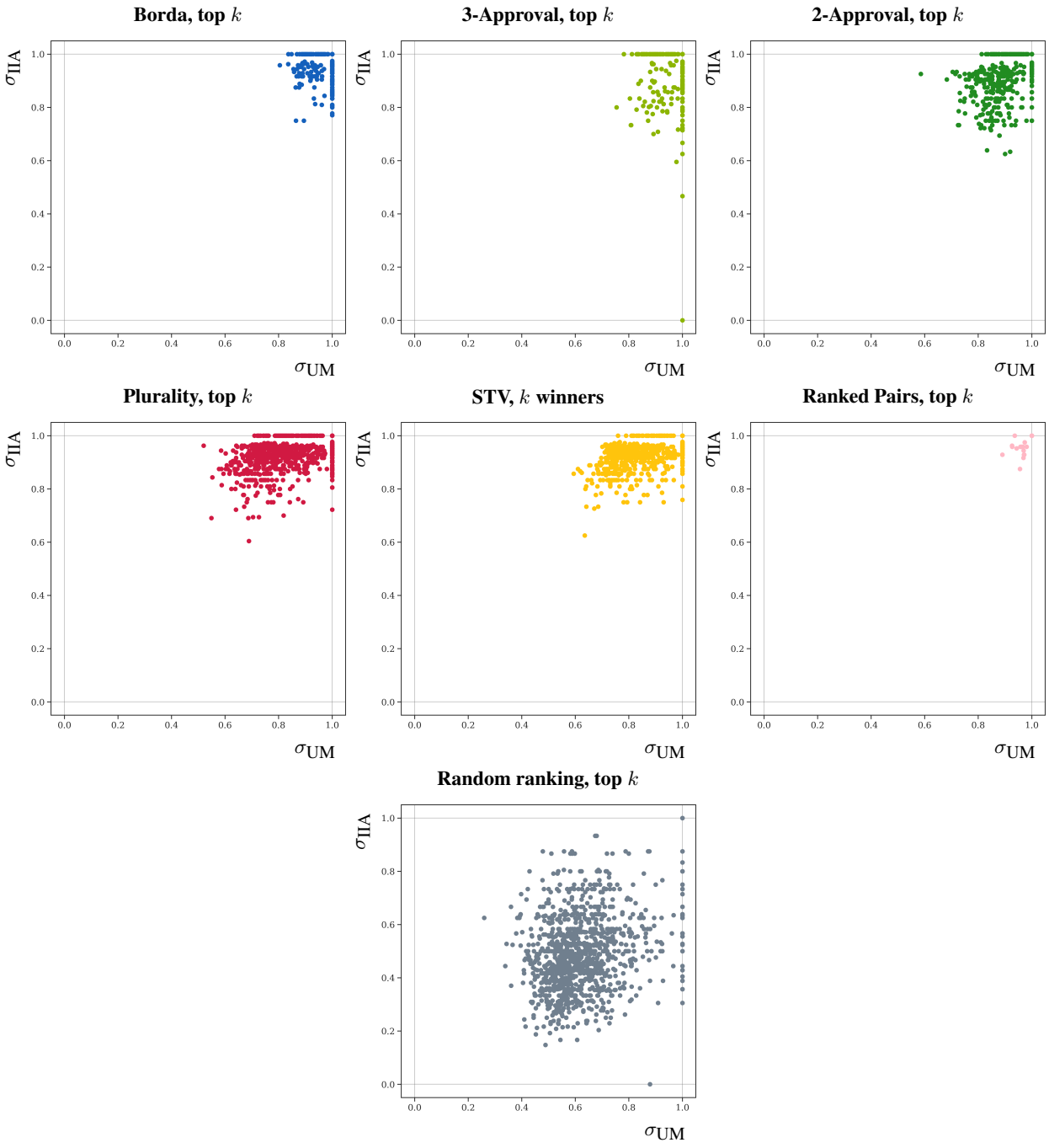


FIG 5. *The same picture as before, but this time the voting rules are configured to take rankings as input and give a set of winners as output. The number of winners* $1 \leq k \leq 5$ *used to make each data point is the one actually used in the corresponding Scottish election, with* $k = 3, 4$ *by far the most common.*

4.2. *Portland and New York City.* In this section, we turn to two quite interesting ranked choice elections in the United States—both recent, at the time of writing. In June 2025, IRV was used to choose the Democratic nominee for mayor of New York City. And earlier, in November 2024, the city council of Portland was expanded to 12 members and elected via 3-winner STV in each of four districts.

	Borda				**3-Approval**				**2-Approval**			
	ρ_{IIA}	ρ_{UM}	σ_{IIA}	σ_{UM}	ρ_{IIA}	ρ_{UM}	σ_{IIA}	σ_{UM}	ρ_{IIA}	ρ_{UM}	σ_{IIA}	σ_{UM}
NYC	1	0.9532	1	1	0.9909	0.8185	1	1	0.9909	0.7410	1	1
D1	0.9968	0.8669	0.9394	0.9523	0.9970	0.8669	0.9242	0.9523	0.9950	0.8669	0.9773	0.9523
D2	0.9993	0.8297	0.9524	1	0.9972	0.7818	1	0.9977	0.9970	0.8297	0.9643	1
D3	0.9985	0.8575	1	1	0.9983	0.7897	1	1	0.9973	0.7596	1	1
D4	0.9987	0.9586	1	1	0.9986	0.9565	1	1	0.9978	0.8348	0.9722	0.8348

	Plurality				**STV / IRV**				**Ranked Pairs**			
	ρ_{IIA}	ρ_{UM}	σ_{IIA}	σ_{UM}	ρ_{IIA}	ρ_{UM}	σ_{IIA}	σ_{UM}	ρ_{IIA}	ρ_{UM}	σ_{IIA}	σ_{UM}
NYC	0.9621	0.7410	1	1	0.9606	0.7410	1	1	1	1	1	1
D1	0.9935	0.6890	1	0.8401	0.9695	0.6890	0.9545	0.8401	1	1	1	1
D2	0.9975	0.7823	1	1	0.9842	0.7823	1	1	1	1	1	1
D3	0.9961	0.7387	1	1	0.9886	0.7387	1	1	0.9999	0.9612	1	1
D4	0.9954	0.7673	0.9769	0.7673	0.9868	0.7948	1	0.8348	1	1	1	1

In New York's primary, there were 1,071,730 valid ballots cast, and voters had the right to rank five, choosing from eleven named candidates and arbitrary write-ins. Outside coverage clearly depicted the leading candidates as Zohran Mamdani (a Democratic Socialist who ultimately won the primary and then the general election) and Andrew Cuomo (the former governor, running as an establishment candidate). Though those were the last two left in the elimination order, it is notable that voter preferences form a ladder profile, and those are not the top two in the Condorcet order. Rather, the order goes

$$\begin{aligned}\text{Zohran Mamdani} \to \text{Brad Lander} \to \text{Adrienne Adams} \to \text{Andrew Cuomo} \to \\ \to \text{Zellnor Myrie} \to \text{Scott Stringer} \to \text{Michael Blake} \to \text{Jessica Ramos} \to \\ \to \text{Whitney Tilson} \to \text{Selma Bartholomew} \to \text{Paperboy Love Prince}\end{aligned}$$

This is because there was a de facto progressive slate in the voting; for example, the left-wing Working Families Party (WFP) issued a ranked list in May 2025, encouraging its voters to "rank the slate" with Mamdani followed by Lander, Adams, Myrie, and Ramos.[9] In fact, Mamdani and Lander were ranked on a nearly-identical number of ballots (639,693 and 631,838, respectively), while Cuomo was ranked on only 489,682. With mentions on fewer than half (about 45.7%) of the ballots, this means that Cuomo was not viable—he could not have reached the threshold of 50% of ballots even if he received every possible ballot transfer in his favor. (See Def 4.1.)

In Portland, similar dynamics can be observed, where the Condorcet order frequently lists several candidates from one de facto slate before switching to the other. For instance, Portland's D4 had 30 candidates listed on the ballot, and voters had the right to rank six. In this case, because three candidates are being elected, the threshold of support needed to win is lower than in New York's IRV election; securing 25% of the votes after any round's transfers

[9] See workingfamilies.org/2025/05/nywfp-ranks-zohran-mamdani-1-for-nyc-mayor. Late in the campaign, Ramos cross-endorsed with Cuomo and was dropped by WFP.

are executed is enough to be elected. In D4, there were eight viable candidates, meaning that they were ranked on enough ballots to make it possible to cross threshold.

The candidates in D4 cleave neatly into two groups: Clark, Zimmerman, Arnold, and Weinstein enjoy mutual support; this is largely disjoint from a second group comprised of Green, Sylkie, Lykins, and Freeman (see Figure 6.) We call those the CZAW and GSLF slates.

DEFINITION 4.1 (Mentions, viable candidates, and boost). In a ranking-based election, let $V(i)$ be the set of ballots that include candidate X in the rankings, so that $|V(i)|$ is the number of ballots listing candidate i, also known as the number of *mentions* of X.

In an STV election with strict threshold T, we say candidate i is *viable* if $|V(i)| > T$. The terminology is justified because to be non-viable means that a candidate cannot cross threshold under any pattern of ballot transfers.

For distinct candidates $i, j \in \mathcal{C}$, let the *boost* to i from j be

$$B_{ij} := \frac{|V(i) \cap V(j)|}{|V(j)|} - \frac{|V(i)|}{n},$$

where n is the number of voters/ballots. (Set all $B_{ii} = 0$.) This is the difference between the rate of mentions of i among those who listed j and the overall rate. If we consider a preference profile as inducing an $n \times m$ binary matrix of mentions, this is just the normalized covariance $B_{ij} = \frac{Cov(i,j)}{|V(j)|/n}$, which shows us that B_{ij} and B_{ji} have the same sign, but not necessarily the same magnitude. If $B_{ij} > 0$, we say the candidates are *mutually boosting*.

In District 4, the partition of the viable candidates into two slates is extremely clear, as shown in Figure 6. If you consider the share of first-place support going to each slate, about 62.7% of voters who ranked a viable candidate first chose one of CZAW and the other 37.3% chose one of GSLF. Under other measures of relative support, like mentions or Borda points, the CZAW slate is preferred at a rate of 55-61% (see Table 3). Intuitively, the proportional outcome would have two winners from CZAW and one from GSLF. However, voter preferences form a Condorcet ladder, and the order is $C > Z > A > G > S > L > F > W$. So with three seats to be filled, any rule with $\sigma_{\text{UM}} = 1$ would have to award all seats to the CZAW slate. Under the STV system that was in place, on the other hand, Clark crossed threshold first, followed by Green, and then Zimmerman—since each slate gets representation, this is a superior outcome from the point of view of proportionality. Given the actual preference profile observed in Portland D4, the only way to get the more proportional outcome is to sacrifice the higher UM score.

District 4 is the only one admitting a strict sorting with $B_{ij} \geq 0$ within slates and $B_{ij} < 0$ between slates. To examine the same phenomenon in the other three Portland districts, we generalize to optimized slate partitions as follows. Form a candidate graph whose nodes are the viable candidates, with a directed edge with weight B_{ij} from candidate i to candidate j. There are edges going both ways between each pair of candidates. We then employ a generalization of the standard network modularity score to weighted, directed graphs: we measure the surplus in positive edge-weight and the deficit of negative edge-weight within slates, against a null model with similarly polarized structure. Finally, we use a Glauber-style Markov chain in a local search to heuristically optimize the slates. See Appendix B for details.

Under this methodology, District 1 cleaves into $4 + 3 = 7$ viable candidates in two de facto slates; District 2 has $4 + 5 = 9$; District 3 has $2 + 3 = 5$; and we have already seen $4 + 4 = 8$ viable candidates in District 4. These slate partitions were the best observed in every one of 1000 independent heuristic optimization runs in each district; they are reflected in Table 3, along with several measures of their relative popularity with voters. The table goes on to compare the *vote* support to the *seats* for each slate under the six main voting rules considered in this paper.

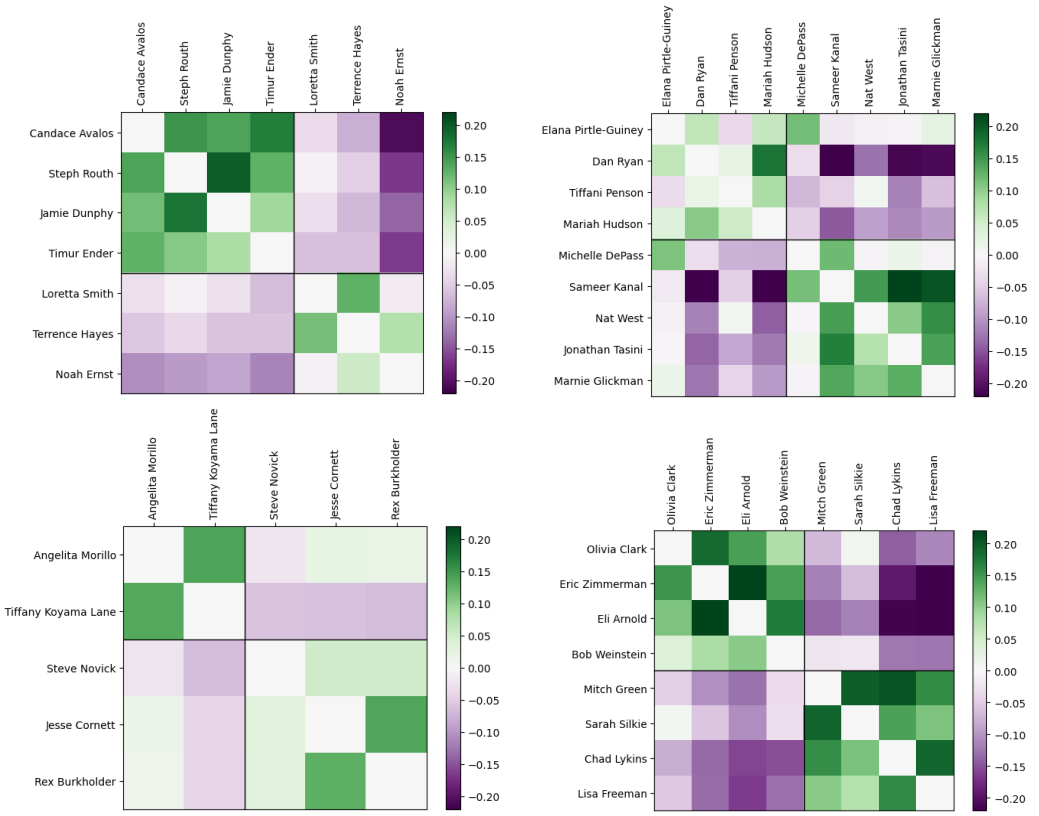


FIG 6. *A modularity-based Markov chain is used to sort the viable candidates into two slates in each of the four Portland districts. Green cells indicate a positive boost: voters ranking one candidate are more likely to rank the other. Purple shading indicates a retarding effect ("anti-boost"). The goal is to choose a partition that has pairwise mutually boosting relationships within slates, but not between slates. This amounts to permuting the rows and columns to obtain green blocks on the diagonal and purple blocks off-diagonal.*

Portland	D1 ARDE vs. SHE	D2 PRPH vs. DKWTG	D3 MK vs. NCB	D4 CZAW vs. GSLF
First-place share	.604	.534	.537	.627
Mentions share	.631	.463	.455	.551
Borda share	.632	.475	.488	.575
Borda top-3 share	.627	.498	.523	.606
Borda	ARS- 2/3	RPK- 2/3	NKM- 2/3	CZA- 3/3
3-Approval	ARS- 2/3	RPD- 2/3	NKM- 2/3	CZA- 3/3
2-Approval	ARS- 2/3	RPK- 2/3	NMK- 2/3	CZG- 2/3
Plurality	ASD- 2/3	PRK- 2/3	NMK- 2/3	CGA- 2/3
STV	ASD- 2/3	KPR- 2/3	NMK- 2/3	CGZ- 2/3
Ranked Pairs	ARD- 3/3	PRK- 2/3	NMK- 2/3	CZA- 3/3

TABLE 3

The top half of the table shows quantifies voter support for the first slate; this is measured by first-place votes (1-0-0-0-0-0), mentions (1-1-1-1-1-1), standard Borda points (6-5-4-3-2-1), or top-three Borda points (3-2-1-0-0-0), as a share of the viable-candidate total. The bottom half shows the seats actually awarded to the first slate in the November 2024 elections, with the names listed in order. (By score in the first four voting rules, by order of election in STV, and by rank in Ranked Pairs.) All voter support shares are between 1/3 and 2/3 for each slate, but several voting rules award a seat sweep to the first slate.

These findings heighten the worry that high scores on IIA and UM might be coming at a cost to proportionality. The voting rules with the highest IIA and UM performance in real-world (Scotland, NYC, and Portland) elections have been Ranked Pairs, then Borda, then 3-Approval; these are exactly the voting rules that sometimes give a 3-seat sweep to a Portland candidate slate even when proportionality would counsel 1-2 seats for each side.

Four Portland districts is a very small sample, so this motivates a turn to synthetic data.

5. Synthetic data: Bradley-Terry profiles. Since IIA and UM axiom adherence may conflict with other norms like proportionality, we turn to parametric voter models to generate a large number of preference profiles with flexible control over profile features. We design methods that build on reasonable descriptions of the voter behavior assumptions, and that highlight preference structures that can lead to lower scores. This section reports the outcomes of trials run on synthetic preference profiles generated using a model derived from a Bradley–Terry (BT) process: this is a classical statistical model for generating ranked preferences based on latent candidate strengths (Bradley and Terry, 1952).[10]

5.1. *The generative model.* The profiles are constructed as follows. First, the cohesion parameter π is used to put a relative weight on each ballot type. We use A for the slate with more support and B for the slate with less support. Suppose we are in the $(r, s) = (2, 6)$ case, meaning that voters are faced with two A candidates and six B candidates. Consider for example the ballot types $AABBBBBB$ and $BABBABBB$. Each of these ballots contains 12 pairwise comparisons of an A with a B candidate. In the first ballot, all 12 rank $A > B$, while the second has 8 comparisons with $A > B$ and four with $B > A$. Suppose that the A voting bloc has cohesion parameter π, thought of as their probabilistic tendency to prefer an A to a B candidate head-to-head. Then the relative weight of the first ballot is $w_1 = \pi^{12}$ and the second is $w_2 = \pi^8(1-\pi)^4$. The probability of drawing the first ballot is then $w_1/\sum w_i$, where the sum is over all $\binom{8}{2}$ ballot types for this candidate pool. We run this in four scenarios by letting each slate have cohesion $\pi = 0.7, 0.9$ independently.

Each generated ballot type is then used to construct a full ballot by ordering the r candidates in the A slate and the s candidates in the B slate. To do this, we use the symmetric Dirichlet parameter α to generate utilities for the candidates within a slate. The setting $\alpha = 1$ means that the utilities are selected uniformly (by Lebesgue measure) in the simplex on the r (respectively, s) extreme points corresponding to the candidates. For instance, if $r = 2$, then a draw from $\alpha = 1$ might give .41 utility to candidate A_1 and .59 to candidate A_2. The voter who drew these utility values is then 41% likely to use the order $A_1 > A_2$. We use seven different Dirichlet parameter settings to see a variety of ranking behavior.

The net effect of all of these variations is to create a diverse dataset of preference profiles, interpolating from modestly to highly polarized and from equal-sized blocs to a heavy predominance of one type of voter.

Figures 7 and 8 show scatterplots similar to the ones presented for the Scottish election data, but instead generated from 84,000 synthetic profiles. Each of these profiles was generated with 15,000 voters in an election to fill three seats; we vary over 840 combinations of other parameters and use each combination 100 times. The other parameters include the number of candidates in each slate ($(r, s) = (2, 6), (4, 4), (6, 2), (2, 8), (5, 5), (8, 2)$); the proportions of voter belonging to the larger bloc $(0.5, 0.6, 0.7, 0.8, 0.9)$; the cohesion parameters controlling the tendency to favor one's own slate ($\pi = .7, .9$ for each bloc); and the Dirichlet parameters governing candidate strength within slates (with seven variations).[11]

[10] This sometimes goes by the name Bradley–Terry–Luce model in the statistics literature.

[11] Letting α_{ij} denote the strength parameter when bloc i voters rank slate j candidates, we consider $(\alpha_{AA}, \alpha_{AB}, \alpha_{BA}, \alpha_{BB})$ ranging over $(1, 1, 1, 1)$, $(1/2, 1/2, 1/2, 1/2)$, $(2, 2, 2, 2)$, $(1, 1/2, 1, 1/2)$, $(1, 2, 1, 2)$, $(1/2, 1, 1/2, 1)$, $(2, 1, 2, 1)$.

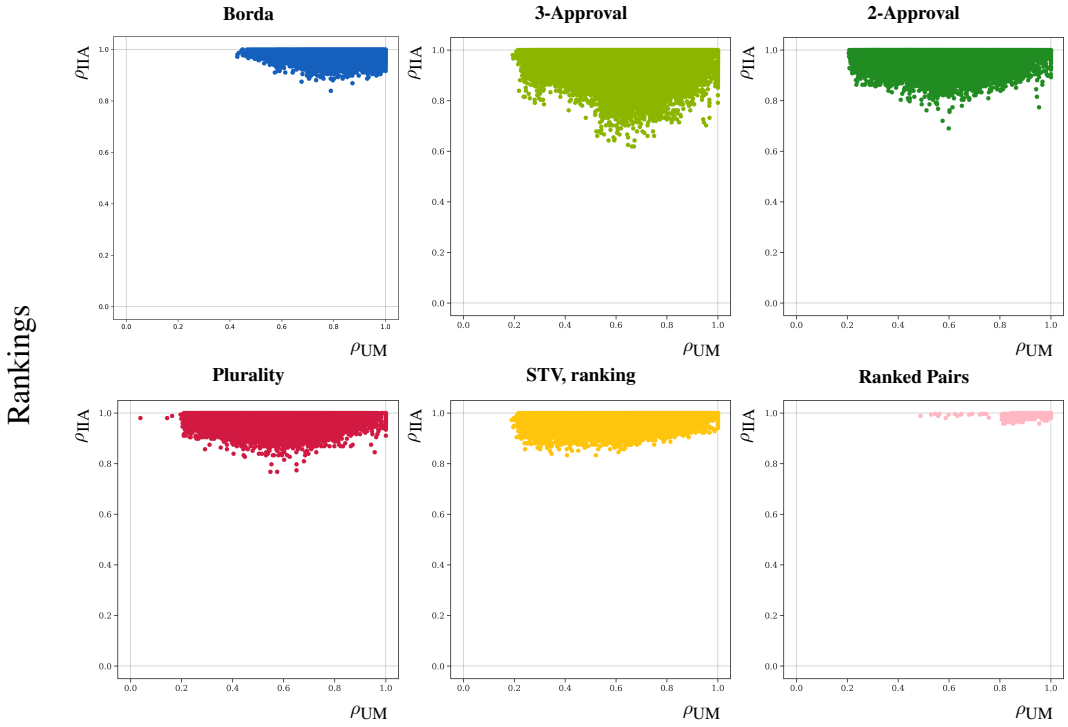


FIG 7. *These scatterplots show scores for the Bradley-Terry profiles using ranking rules. Each datapoint represents the score when the indicated voting rule is applied a single profile.*

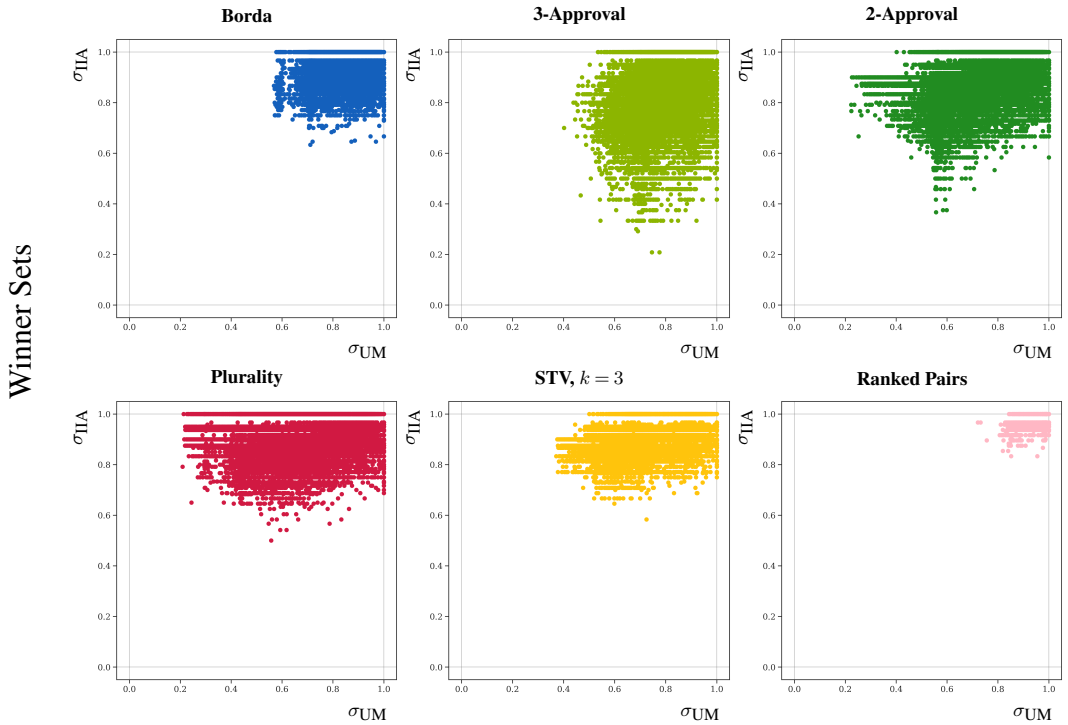


FIG 8. *These scatterplots show scores for the Bradley-Terry profiles when the rules output winner sets. Each datapoint represents the score when the indicated voting rule is applied a single profile.*

5.2. *Proportionality.* Above, we motivated the intuition that our Arrovian scores are in tension with another representative norm: proportionality. This was suggested by the presence of a seat sweep for one slate of candidates under the Ranked Pairs, Borda, and 3-Approval rules, which score highest on the IIA and UM gradated scores. At least for UM, the underlying reason for this may be the one suggested in New York and Portland: if the slate with plurality support has voters who rank them in consistent order, then they may receive the top several positions in the Condorcet order, i.e., several candidates from that slate beat all others head-to-head. If that occurs, then respecting majority preferences mandates a sweep.

The results of testing against Bradley-Terry profiles, shown in Figure 9, strongly bear that out. The STV rule always respects proportionality up to rounding. By contrast, Ranked Pairs and Borda can often give all the seats to one side, even when voter support for each slate is quite close to $1/2$.

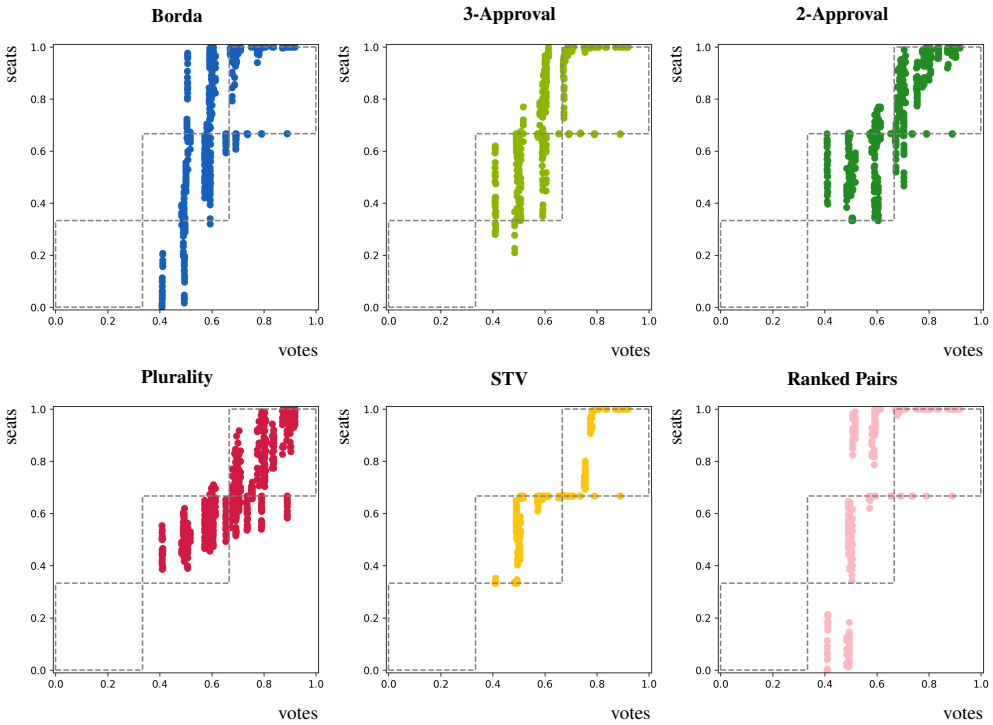


FIG 9. *The same set of 84,000 Bradley-Terry profiles is now used to illustrate the proportionality properties of the various voting rules, each set up to elect three winners. The A slate has 50-90% of the voters, and the other candidate and voter parameters are varied as well. Each plot has one data point for each of the 840 parameter tuples described above. The x position of the dot ("votes") shows the expected first-place vote share for slate A under those parameters. The y position ("seats") shows the average slate A seat share when the voting rule is applied to 100 profiles generated from the same parameters. The dotted lines show direct seats-to-votes proportionality, rounded up or down to whole numbers of seats ($\lfloor V \rfloor \leq S \leq \lceil V \rceil$).*

A closer inspection of the results shows the value of experimenting with parameter interactions. For instance, all voting rules can give just two seats to slate A even when they have as much as 90% of the first-place votes. This is because the candidate strength parameters create the possibility that a single A candidate is overwhelmingly popular. It may at first seem paradoxical that the plurality rule, which simply counts first-place votes, can also give just one seat to the A slate—this can be observed in the relatively large number of points with $V > 2/3$ and $S < 2/3$. But this is exactly what we should expect to occur when there is one popular

A candidate while the B candidates have several candidates of roughly equal strength.[12] It would be very hard to observe this occurring under simpler models of voter preference, like the popular *impartial culture* model that makes all permutations equally likely.[13]

6. Discussion and Significance. In this paper, we introduce quantitative metrics that offer gradations of two foundational social choice axioms: Independence of Irrelevant Alternatives (IIA) and the Unanimity criterion (U). We place unanimity on a scale with majoritarianism to facilitate comparison across profiles. Our analysis demonstrates the value of extending classical binary axioms to gradated metrics, enabling the assessment of voting rules along a continuum of axiom compliance. The Ranked Pairs rule achieves optimal majoritarian performance due to its construction that selects for maximum alignment with voter pairwise preference; in addition, its strong Condorcet consistency guarantees perfect stability in the great majority of elections that have a total Condorcet order. Across both real-world elections from Scottish local government and synthetic datasets generated with a Bradley-Terry statistical model, we find that the Borda voting rule consistently achieves high marks for stability and majoritarianism without being strongly Condorcet consistent. We believe that our statistical approach may be of independent interest for quantifying axiom compliance in welfare economics and in applications to alignment for large language models, where incorporating more axioms in the regularization objectives—rather than focusing solely on utility or main accuracy loss—can lead to more balanced and principled decision-making.

However, readers will note that we have refrained from using the common phrasing in which IIA and U are called axioms of *fairness*. Our analysis has led us to a more critical view of the desirability of the axioms under discussion. By situating Arrow's Impossibility Theorem in a metric setting, gradated scores provide a new lens for identifying trade-offs between objectives.

There is little apparent downside to prioritizing stability to candidate availability (IIA) and emphasizing majority preferences (UM) when designing multi-agent cooperative protocols, but the situation is more nuanced in the setting of political representation. In the electoral setting, these attributes could easily cut against other important normative values, like *responsiveness* (the ease with which outcomes shift to reflect changing preferences) and *proportionality* (which in particular requires that sizable minorities receive meaningful representation).

We leave many directions open for future development. We hope to develop useful gradations for other binary axioms that are prominent in voting theory and machine learning, like monotonicity. A fuller exploration of the Pareto frontier for $(\rho_{\text{IIA}}, \rho_{\text{UM}})$ would be informative. Relatedly, it is unclear what structural features of a voting rule are best suited to maximizing IIA scores in the presence of Condorcet cycles. Finally, it would be extremely valuable to conduct a sustained analysis of the negative relationship between IIA and UM on one hand and proportionality on the other. The Portland analysis in particular (§4.2) gives intriguing hints that undeserved sweeps may occur when the Condorcet order fails to alternate frequently enough between the preferences of distinct blocs of voters. This direction seems richly worthy of future investigation.

[12] Indeed, when the A voter bloc has 80-90% of the voters, their cohesion is $\pi_A = .9$, and the candidate strengths include $\alpha_{AA} = 1/2$ (there is a strongest candidate) and $\alpha_{BB} = 2$ (relatively equal strength), we find that the top three candidates in first-place votes are often an A and two Bs. This gives a one-seat A outcome under plurality. However, in that scenario, the ballots headed by an A candidate often have another A candidate in the second rank, and for all the other voting rules, that is enough to get a second A elected.

[13] See for example Boehmer et al. (2024) for an overview of popular *statistical cultures* (generative models for simulating voter preferences) in the computational social choice literature. Randomizing elections by making all permutations equally likely is still the most popular simulation model, while universally regarded as unrealistic.

7. Acknowledgments. The authors thank Erica Chiang, Daniel Brous and Sujai Hiremath for collaboration on earlier work to define relaxed Arrow axioms. We also thank Edouard Heitzmann, Sophia Guo, Michelle Contreras Catalán, and Kris Tapp for collaborative conversations about slate detection in ranked choice elections.

REFERENCES

ANSCOMBE, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika* **35** 246–254.

ARROW, K. J. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy* **58** 328–346.

BOEHMER, N., FALISZEWSKI, P., JANECZKO, Ł., KACZMARCZYK, A., LISOWSKI, G., PIERCZYŃSKI, G., REY, S., STOLICKI, D., SZUFA, S. and WĄS, T. (2024). Guide to Numerical Experiments on Elections in Computational Social Choice. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)* 7962–7970. https://doi.org/10.24963/ijcai.2024/881

BRADLEY, R. A. and TERRY, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* **39** 324–345.

CARAGIANNIS, I., PROCACCIA, A. D. and SHAH, A. (2013). When Do Noisy Votes Reveal the Truth? In *Proceedings of the 14th ACM Conference on Electronic Commerce (EC '13)* 143–160. ACM, New York, NY. https://doi.org/10.1145/2482540.2482570

CONITZER, V., FREEDMAN, R., HEITZIG, J., HOLLIDAY, W. H., JACOBS, B. M., LAMBERT, N., MOSSÉ, M., PACUIT, E., RUSSELL, S., SCHOELKOPF, H. et al. (2024). Social choice should guide AI alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*.

DATA AND DEMOCRACY LAB (2024). VoteKit: Python package. https://github.com/mggg/VoteKit. GitHub repository.

DELEMAZURE, T., LANG, J. and PIERCZYŃSKI, G. (2024). Independence of irrelevant alternatives under the lens of pairwise distortion. In *Proceedings of the AAAI Conference on Artificial Intelligence* **38** 9645–9652.

DIACONIS, P. (1988). Group representations in probability and statistics. *Lecture notes-monograph series* **11**.

DIACONIS, P. and GRAHAM, R. L. (1977). Spearman's Footrule as a Measure of Disarray. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **39** 262–268. https://doi.org/10.1111/j.2517-6161.1977.tb01624.x

DOUGHERTY, K. L. and HECKELMAN, J. C. (2020). The probability of violating Arrow's conditions. *European Journal of Political Economy* **65** 101936.

GÓMEZ, S., JENSEN, P. and ARENAS, A. (2009). Analysis of community structure in networks of correlated data. *Physical Review E* **80** 016114. https://doi.org/10.1103/PhysRevE.80.016114

HORNISCHER, L. and TERZOPOULOU, Z. (2025). Learning How to Vote With Principles: Axiomatic Insights Into the Collective Decisions of Neural Networks. *Journal of Artificial Intelligence Research*. Preprint also available on arXiv:2410.16170. https://doi.org/10.1613/jair.1.18890

KALAI, G. (2002a). A quantitative version of Arrow's theorem. *Economic Theory* **20** 685–694.

KALAI, G. (2002b). A Fourier-theoretic perspective on the Condorcet paradox and Arrow's theorem. *Advances in Applied Mathematics* **29** 412–426.

LAMBORAY, C. (2006). An axiomatic characterization of the prudent order preference function. *Annales du LAMSADE* **3** 53–71.

MASKIN, E. (2025). Borda's rule and Arrow's independence condition. *Journal of Political Economy* **133** 385–420.

MCCUNE, D. and GRAHAM-SQUIRE, A. (2024). Monotonicity anomalies in Scottish local government elections. *Social Choice and Welfare* **63** 69–101.

MOSSEL, E. (2012). A quantitative Arrow theorem. *Probability Theory and Related Fields* **154** 49–88.

MOSSEL, E., O'DONNELL, R. and SERVEDIO, R. A. (2005). A quantitative Arrow theorem. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing* 556–564.

NEWMAN, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* **103** 8577–8582. https://doi.org/10.1073/pnas.0601602103

PROCACCIA, A. D. and ROSENSCHEIN, J. S. (2007). Junta distributions and the average-case complexity of manipulating elections. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*.

ZHAO, X., WANG, K. and PENG, W. (2024). Measuring the inconsistency of large language models in preferential ranking. *arXiv preprint arXiv:2410.08851*.

APPENDIX A: CHOICES IN THE METRIC CONSTRUCTIONS

For both the UM and IIA metrics, the definitions are fairly natural, but there are still choices to be made. We defined the UM scores using a worst-case construction in M and an arcsin interpolation function. For IIA, we use an averaging convention. Also, the IIA construction is built by disqualifying single candidates rather than larger subsets. In this appendix we justify these design choices.

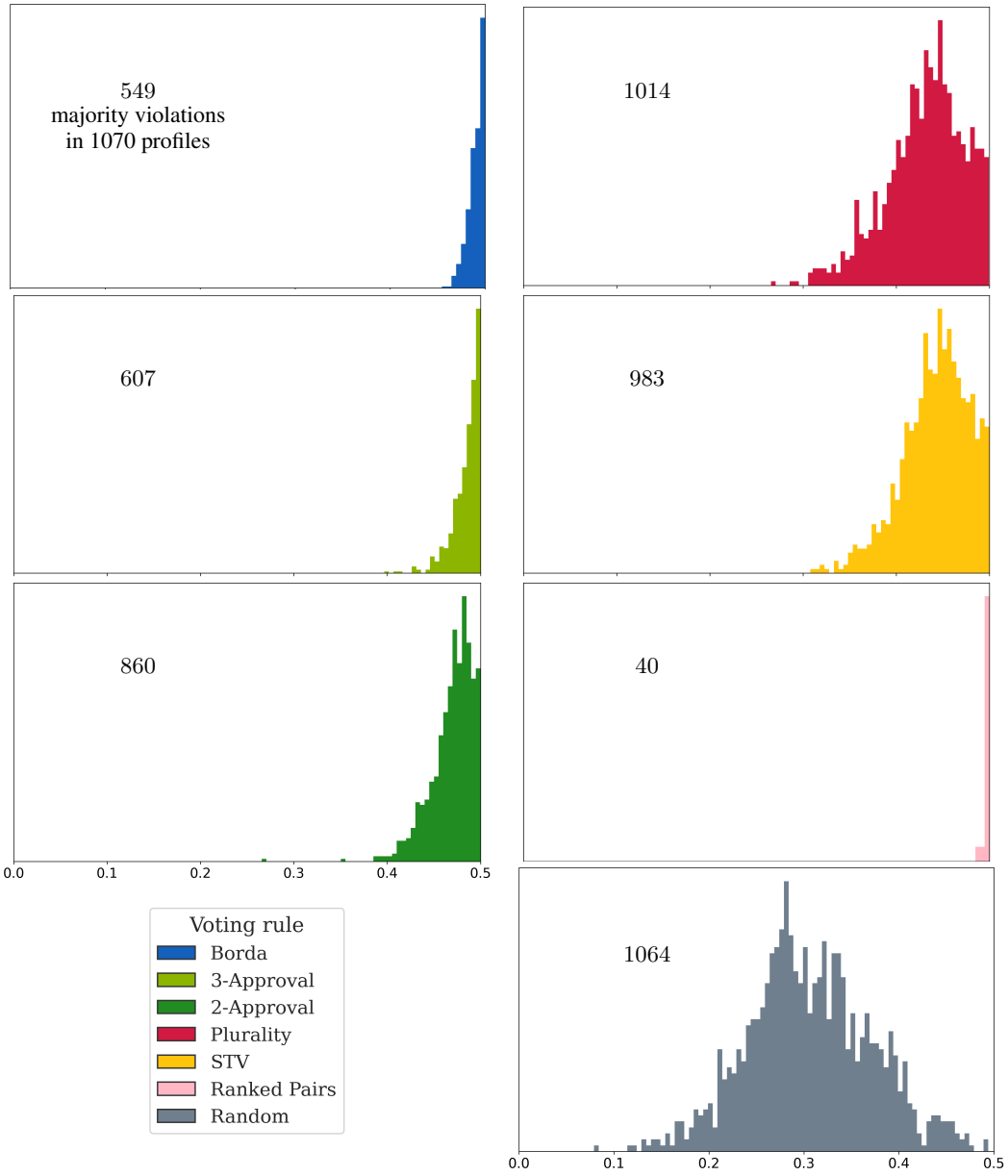


FIG 10. *Each histogram shows the distribution of M values over all profiles in the Scottish dataset for which the voting rule in question has $M < 1/2$, indicating that some pair of candidates is ranked against the majority preference. The number printed in each plot is the total mass in the histogram (i.e, the number of elections out of 1070 with a non-majority pairwise outcome).*

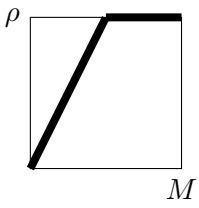
A.1. Worst-case alignment, arcsin interpolation in UM. Recall that ρ_{UM} and σ_{UM} are strictly less than one exactly when $M < 1/2$. This means that defining M by the worst case over pairs of candidates gives an interpretable generalization of the unanimity condition: $\rho_{\text{UM}} < 1$ iff there exists a minority preference that prevails. Using an averaging convention would mean that $\rho_{\text{UM}} < 1$ only when there are severe enough majority violations to bring down the average, which has a different meaning for each number of candidates. Thus the worst-case formulation of M isolates the presence of a violation, which would not occur at any consistent numerical level of $\rho_{\text{UM}}, \sigma_{\text{UM}}$ otherwise.

Next, we seek to construct ρ_{UM} (and likewise σ_{UM}) from M in a manner that flags majority violations when $\rho_{\text{UM}} < 1$ and flags unanimity violations when $\rho_{\text{UM}} = 0$. Worst-case M facilitates this goal, but an interpolation function is needed.

Consider three alternatives.

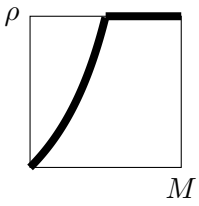
LINEAR

$$\rho_{\text{UM}}(f, P) := \begin{cases} 2M, & M(f, P) < 1/2 \\ 1, & M(f, P) \geq 1/2. \end{cases}$$



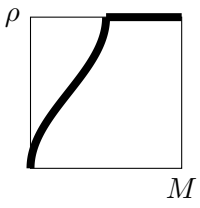
ODDS

$$\rho_{\text{UM}}(f, P) := \begin{cases} \frac{M}{1-M}, & M(f, P) < 1/2 \\ 1, & M(f, P) \geq 1/2. \end{cases}$$



ARCSIN

$$\rho_{\text{UM}}(f, P) := \begin{cases} \frac{2}{\pi} \arcsin \sqrt{2M}, & M(f, P) < 1/2 \\ 1, & M(f, P) \geq 1/2. \end{cases}$$



We might prefer either the odds or the arcsin formulations over the linear one due to their high slope near $M = 1/2$. Practical voting rules applied to real-world preference profiles give many misalignment scores just under $1/2$ (Figure 10). High slope of ρ near $M = 1/2$ makes it much easier to distinguish between close cases.

Figures 11 and 12 show the histograms of M values for synthetic profiles generated with a majority bloc proportion of 0.7. We then transform those M distributions under the arcsin and odds formulations of ρ_{UM}.

Finally, we can confirm that the mean-variance correlation was already fairly low in the Bradley-Terry data, but that arcsin transformation performs as expected, further stabilizing variance (Figure 13).

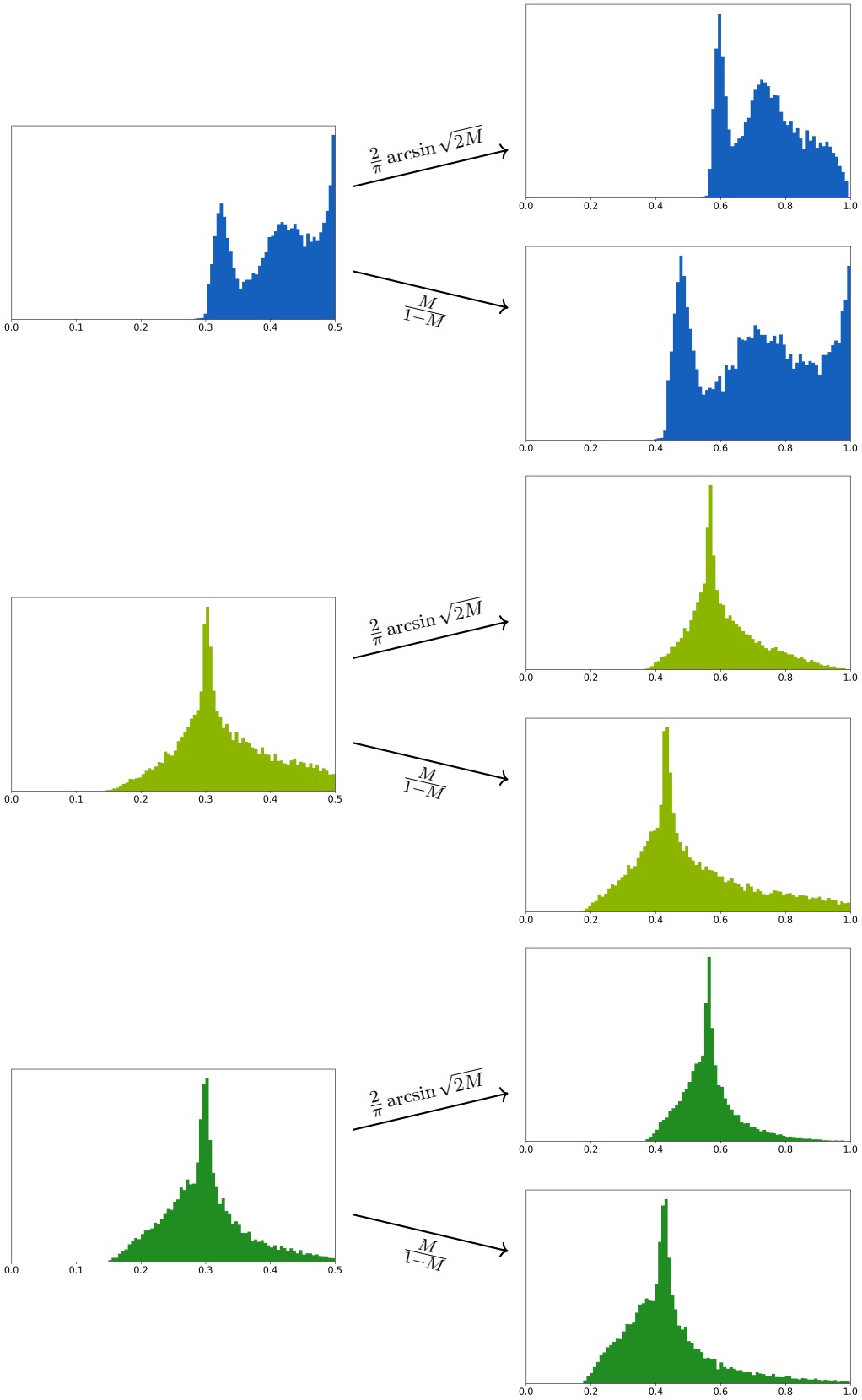


FIG 11. LEFT: *The histogram of M values for various voting rules applied to 16,800 profiles where A voters make up 70% of the electorate. A high spike at 0.3 indicates instances in which some $B_j \succ A_i$.* RIGHT: *The histogram of ρ_{UM} values after applying either the arcsin or odds transformation.*

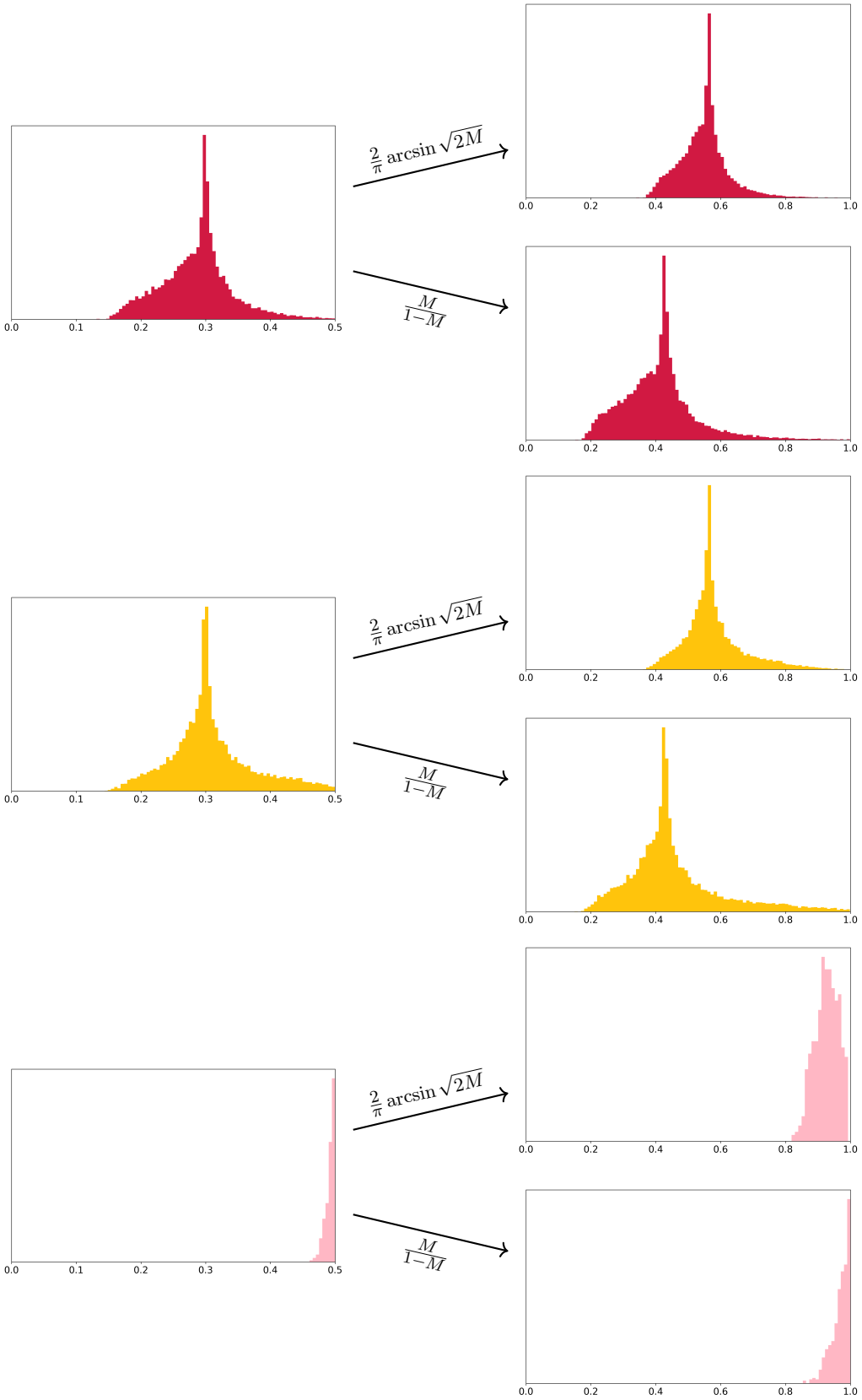


FIG 12. LEFT: *The histogram of M values for various voting rules applied to 16,800 profiles where A voters make up 70% of the electorate. A high spike at 0.3 indicates instances in which some $B_j \succ A_i$. (This essentially never occurs in the Ranked Pairs rule.)* RIGHT: *The histogram of ρ_{UM} values after applying either the arcsin or odds transformation.*

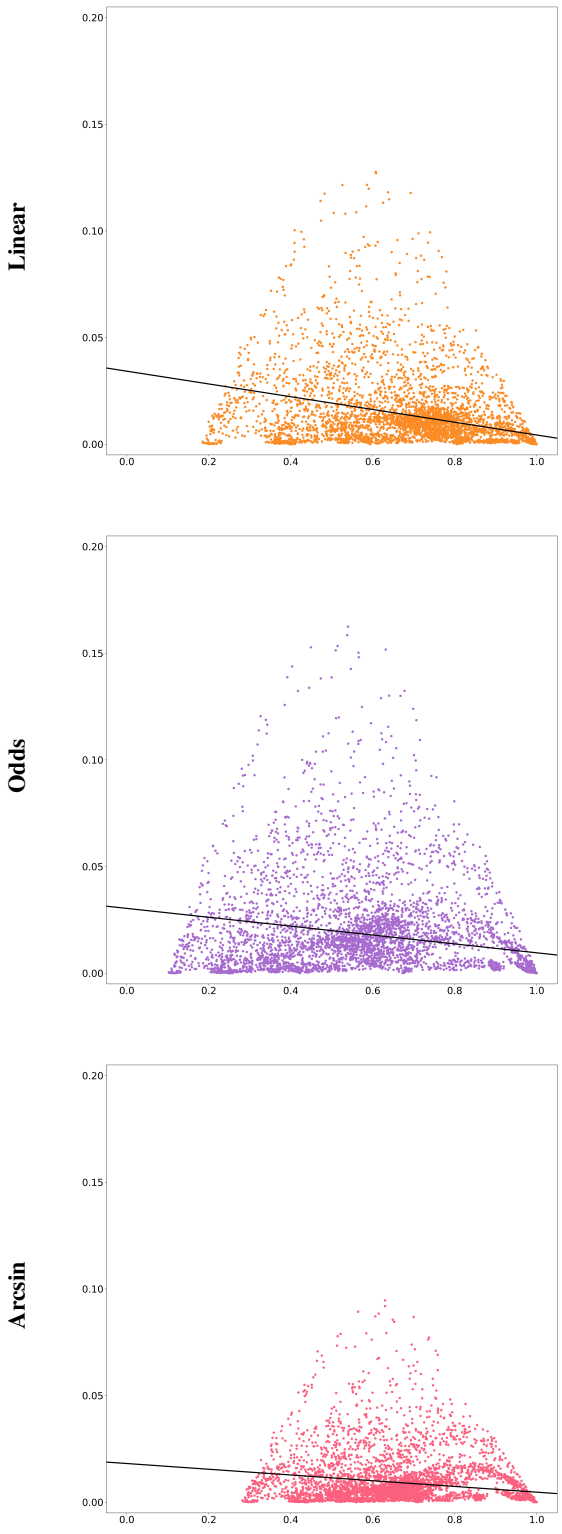


FIG 13. *For each of the 840 parameter tuples and each of the six ranking-based voting rules f, we show the mean and the variance of $\rho_{\mathrm{UM}}(f, P)$ over 100 random profiles P made with those parameters. This produces 5040 data points in each scatterplot.*

A.2. Single disqualification, averaging in IIA. As we did in the the UM case, we seek to construct definitions of ρ_{IIA} and σ_{IIA} that prioritize interpretability and are as parallel as possible (between the ranking case and the winner-set case). This time, we will also need to fold in concerns about computational tractability.

In Definition 3.1, we check IIA compliance by considering whether $f\left(P^C\right) = f(P)^C$ for all candidates $C \in \mathcal{C}$. IIA failure is witnessed by the existence of a profile and candidate for which this fails. In the definition, we could equally well have chosen to test $f\left(P^S\right) = f\left(P\right)^S$ over all subsets S of the candidates. Either definition exactly captures the classic IIA axiom via $\rho_{\text{IIA}} \equiv 1$. For completeness, we include a brief proof for the single-disqualification definition. Recall that classical IIA is usually phrased as the requirement that when two profiles P, P' with the same voters and candidates have the property that A, B have the same relative ranking from all voters, then $f(P)$ and $f(P')$ must have the same relative ranking of A, B.

PROPOSITION A.1. *A ranking rule f satisfies classical IIA iff*

$$f(P^C) = f(P)^C \quad \text{for all profiles } P \text{ and candidates } C.$$

PROOF. If rankings are stable under single-candidate disqualifications, then they are also stable under setwise disqualifications (just by removing candidates one at a time). Thus letting $S = \mathcal{C} \setminus \{A, B\}$, and assuming P, P' rank A, B the same way in every ballot, we have $P^S = (P')^S$, so $f(P)^S = f(P^S) = f(P'^S) = f(P')^S$, as needed.

For the other direction, suppose classical IIA holds for f. Consider P^C for a candidate C. This has all the same pairwise comparisons as P for any other two candidates A, B, so IIA implies the same relative rankings for every pair, which means $f(P^C) = f(P)^C$. □

So single-disqualification and all-subsets-disqualification capture the Arrow axiom equally well. There are two reasons to avoid the all-subsets option. One problem is that it adapts much less cleanly from rankings to winner sets. Recall Def 3.2:

$$\sigma_{\text{IIA}}(f^{(k)}, P) := \frac{1}{m}\left[\sum_{C \in f^{(k)}(P)} \frac{|f^{(k-1)}(P^C) \cap f^{(k)}(P)|}{k-1} + \sum_{C \notin f^{(k)}(P)} \frac{|f^{(k)}(P^C) \cap f^{(k)}(P)|}{k}\right].$$

(In this notation, the superscript decoration $f^{(k)}$ is a reminder that the general rule f is being applied to output k winners.) We have split this definition up by whether C was a winner or loser, which controls whether $f^{(k)}(P) \setminus \{C\}$ has $k-1$ or k elements. The cases for the size of $f^{(k)}(P) \setminus S$ would proliferate if disqualifying larger subsets, and would make for a messier definition with more arbitrary choices to make. In addition, an all-subsets definition requires checking a number of cases that grows exponentially rather than linearly in $m = |\mathcal{C}|$, which is manageable for low numbers of candidates but scales poorly.

APPENDIX B: NETWORK MODULARITY AND SLATE OPTIMIZATION

Many authors have contributed to the topic of network modularity, surveyed by Mark Newman in (Newman, 2006). We assess the quality of a clustering of network data via

$$Q = \frac{1}{4m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta_{ij},$$

where δ_{ij} is the indicator for whether nodes i and j are in the same cluster, k_i is the degree of vertex i, and m is the number of edges so that $2m$ is the sum of entries in the adjacency matrix A. The interpretation uses a null model that is randomly "re-wired": we compare the graph to other graphs where each vertex maintains the same degree, but the connections are randomly permuted.

This is easily modified for directed graphs. We have

$$Q = \frac{1}{w} \sum_{i,j} \left[A_{ij} - \frac{k_i^{\text{out}} k_j^{\text{in}}}{w} \right] \delta_{ij},$$

where this time w is the entry sum of A (and A is no longer symmetric).

In our context, we regard weighted edges as a generalization of multi-edges, and we treat positive weight and negative weight separately in the boost matrix B:

$$Q = \frac{1}{p+n} \cdot \sum_{i,j} \left[B_{ij} - \frac{p_i^{\text{out}} p_j^{\text{in}}}{p} + \frac{n_i^{\text{out}} n_j^{\text{in}}}{n} \right] \delta_{ij},$$

Here, $B = P - N$ is a splitting of the B matrix as a difference of a positive part and a negative part, so P, N both have non-negative entries, with p and n as the respective matrix sums. This is quite close to the approach of Gómez, Jensen and Arenas (2009) for signed weighted networks, generalized here to the directed setting. We now run a local-search Markov chain to optimize Q, which will reward us for positive boosts within a cluster and penalize us for negative boosts. (Note that the rewiring convention means it would be essentially redundant to also reward and penalize the signs of connections between clusters.)

In our four Portland districts, a simple 1000 step optimization chain always gives the same answer, which divides the candidates into the clusters shown in Figure 6. Since the number of viable candidates being bipartitioned is quite small in each district (7, 9, 5, and 8 candidates, respectively), it is quite easy to check this against global optimization over bipartitions. We thus confirm that the solutions found by the Markov chain search were indeed globally optimal in each case. We nonetheless emphasize the Markov chain approach here because it scales very efficiently to large numbers of candidates, while brute force search of bipartitions does not.

APPENDIX C: BOOTSTRAP UNCERTAINTY QUANTIFICATION

If we assume that each observed election profile is generated from an unknown population distribution μ over preference profiles, we can use standard bootstrapping techniques to get confidence intervals for estimation of the mean values of our scores.

Our target estimand is the population mean

$$\theta_{g,f,\mu} = \mathbb{E}_{P\sim\mu}\big[g(f,P)\big] \qquad \text{for } g = \rho_{\text{IIA}}, \rho_{\text{UM}}, \sigma_{\text{IIA}}, \sigma_{\text{UM}}.$$

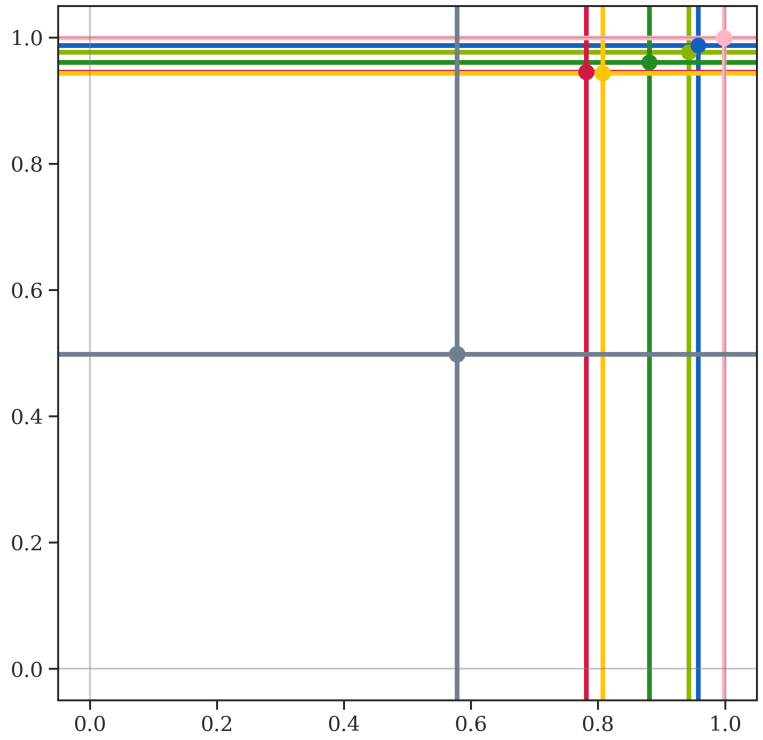


FIG 14. *This plot shows the empirical mean value of ρ_{UM} and ρ_{IIA} for each voting rule over the Scottish dataset.*

The raw mean for each score over the Scottish dataset is shown in Figure 14. We can then run a percentile bootstrap to get confidence intervals, shown in Table 4.

Rule	ρ_{IIA}	ρ_{UM}	σ_{IIA}	σ_{UM}
Borda	0.9879 ± 0.0014	0.9579 ± 0.0029	0.9876 ± 0.0021	0.9913 ± 0.0017
3-Approval	0.9767 ± 0.0025	0.9429 ± 0.0038	0.9806 ± 0.0039	0.9890 ± 0.0020
2-Approval	0.9609 ± 0.0028	0.8808 ± 0.0048	0.9644 ± 0.0041	0.9620 ± 0.0041
Plurality	0.9454 ± 0.0026	0.7814 ± 0.0057	0.9389 ± 0.0035	0.8804 ± 0.0074
STV	0.9434 ± 0.0028	0.8075 ± 0.0056	0.9429 ± 0.0032	0.9148 ± 0.0063
Ranked Pairs	0.9996 ± 0.0002	0.9985 ± 0.0005	0.9992 ± 0.0004	0.9993 ± 0.0004
Random	0.4984 ± 0.0045	0.5780 ± 0.0058	0.4943 ± 0.0086	0.6284 ± 0.0079

TABLE 4

Confidence intervals for percentile bootstrap with 10,000 samples from Scottish elections across all voting rules.

To obtain joint confidence bounds for the pairs of estimates, we apply a Bonferroni correction to the bootstrap intervals.

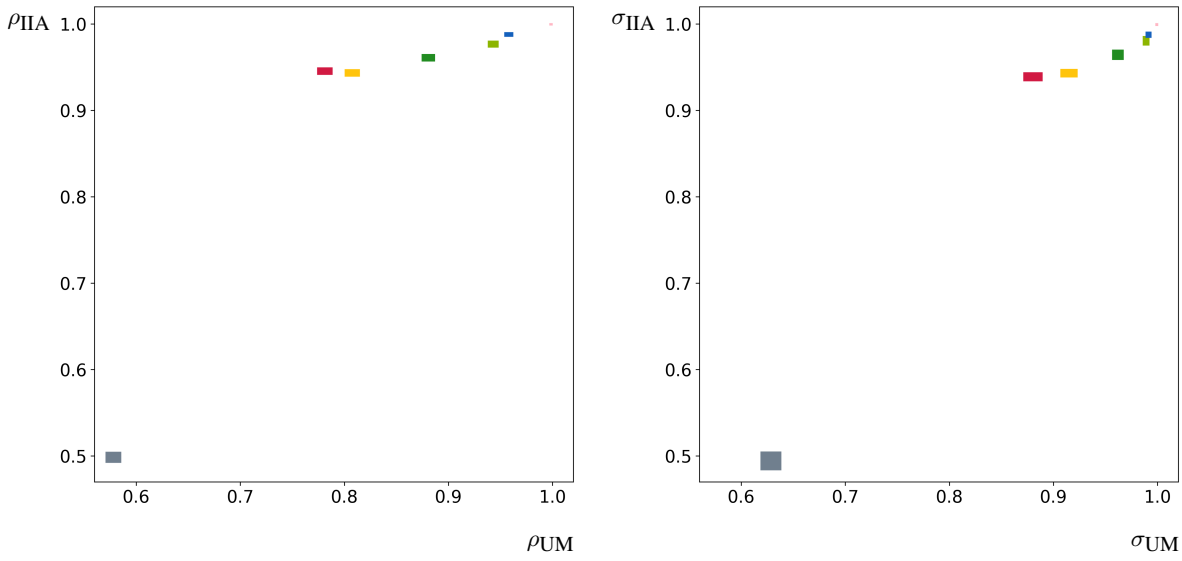


FIG 15. *Bonferroni confidence rectangles for ranking scores (left) and winner-set scores (right).*

This yields an ordering for the joint estimates at the 95% confidence level.

ρ_{IIA}: RP > Bor > 3A > 2A > {STV, Plu}

ρ_{UM}: RP > Bor > 3A > 2A > STV > Plu

σ_{IIA}: RP > {Bor, 3A} > 2A > {STV, Plu}

σ_{UM}: RP > {Bor, 3A} > 2A > STV > Plu