Why Instant-Runoff Voting Is So Resilient to Coalitional Manipulation: Phase Transitions in the Perturbed Culture

François Durand

Abstract

Previous studies have shown that Instant-Runoff Voting (IRV) is highly resistant to coalitional manipulation (CM), though the theoretical reasons for this remain unclear. To address this gap, we analyze the susceptibility to CM of three major voting rules—Plurality, Two-Round System, and IRV—within the Perturbed Culture model. Our findings reveal that each rule undergoes a phase transition at a critical value θ_c of the concentration of preferences: the probability of CM for large electorates converges exponentially fast to 1 below θ_c and to 0 above θ_c. We introduce the Super Condorcet Winner (SCW), showing that its presence is a key factor of IRV's resistance to coalitional manipulation, both theoretically and empirically. Notably, we use this notion to prove that for IRV, $\theta_c = 0$, making it resistant to CM with even minimal preference concentration.

1 Introduction

1.1 Motivation

The Gibbard-Satterthwaite Theorem [16, 31] shows that all non-trivial voting rules are vulnerable to manipulation (strategic voting), even by a single voter. This vulnerability can only worsen when any number of voters with aligned interests can form a coalition to alter the election outcome, a phenomenon called *coalitional manipulation* (CM). Unlike individual manipulation [30, 32], CM remains significant in large-scale elections and raises several concerns, notably creating moral dilemmas [8, Introduction] and power imbalances between strategic and naive voters, thus undermining the "one person, one vote" principle [9, 13].

However, not all voting rules are equally vulnerable: the *CM rate*, i.e., the probability that a voting profile is manipulable by a coalition under a given probabilistic model, can vary significantly between rules. In previous studies, *Instant-Runoff Voting* (IRV) and some of its variants [9] consistently outperform other classical single-winner voting rules in resisting coalitional manipulation, whether the analysis is based on randomly generated profiles [18, 19] [8, Chapters 7–8] or experimental datasets [3, 19] [8, Chapter 9]. This has been confirmed by theoretical calculations in the case of three candidates [25, 27]. This is especially intriguing, as IRV has several theoretical features typically deemed undesirable and seemingly prone to manipulation: most notably, IRV is neither *Condorcet-consistent* nor *monotonic* [2, Definitions 2.8 and 2.10] [11].

In an effort to shed theoretical light on this phenomenon, we will compare IRV to the two other most widely used voting rules in large-scale single-winner political elections: Plurality and the Two-Round System. We will adopt the *Perturbed Culture* model of random voting profiles, first introduced by Williamson and Sargent [34] and later named by Gehrlein [15, Section 4.3.2]. We will focus on the asymptotic behavior as the number of voters tends to infinity, as it offers more mathematical tractability and is relevant for large-scale elections. We will also examine convergence rates to assess how well this limit approximates scenarios with finite electorates. Our approach is similar to the use of the Ising model in physics, which, despite being unrealistic in its microscopic details, has been remarkably effective in explaining the complex macroscopic phenomenon of phase transitions in ferromagnetism [21].

1.2 Contributions

We prove that each of the three voting rules undergoes a *phase transition*, with an abrupt change in behavior based on whether the concentration parameter θ in the Perturbed Culture model exceeds a critical threshold θ_c. Below θ_c, the CM rate tends to 1 for large electorates, while above θ_c, it tends to 0. We compute the critical threshold θ_c for each voting rule as a function of the number of candidates.

We show through simulations how the CM rate curve as a function of θ, which is continuous for a finite number of voters n, converges to a discontinuous curve as n tends to infinity, thereby explaining the phase transition. Additionally, we investigate the critical regime $\theta = \theta_c$, leading to the conjecture that in this case, the CM rate tends to a limit strictly between 0 and 1.

We introduce the concept of a Super Condorcet Winner (SCW), which largely explains IRV's resilience to CM. This leads to one of our most striking results: for IRV, the critical value θ_c is 0, regardless of the number of candidates. This means that IRV is asymptotically resistant to CM as soon as the Perturbed Culture model shows even the slightest preference concentration. Furthermore, using experimental datasets, we show that SCWs are frequent in practice and account for most of IRV's resistance to CM.

Finally, we demonstrate that in non-critical regime, i.e., for $\theta \neq \theta_c$, the convergence of the CM rate toward 0 or 1 is exponentially fast. This implies that our results for $n \to \infty$ quickly become relevant even for finite n. We also study how this speed varies with θ.

This manuscript is a slightly shortened version of our AAMAS paper [10].

1.3 Related Work

In addition to the literature already mentioned, the works most closely related to ours are those studying the CM rate using theoretical tools. Most of them focus on three-candidate elections within models like *Impartial Culture* [26], *Impartial Anonymous Culture* [25, 14], or *Pólya-Eggenberger Urns* [27]. Although they consistently show that IRV is more robust than other rules, they are limited to a specific number of candidates and offer little intuition for IRV's superior performance. Kim and Roush [22] provide key results for large electorates under *Impartial Culture* for Plurality and some other rules (*positional scoring rules* in general, *Maximin*, and *Coombs*) but do not address the Two-Round System or IRV.

Concerning phase transitions, research on this subject is abundant in physics (see Kadanoff [21] for an overview) and in mathematics and computer science[5, 1, 7]. In voting theory, Mossel et al. [28] and Xia [35] also examine phase transitions in coalitional manipulability, focusing on varying numbers of manipulators. In contrast, our study considers the impact of the concentration parameter in the probabilistic distribution of preferences.

1.4 Limitations

The limitations of this work stem from its main assumptions. First, while the Perturbed Culture model is useful, it does not capture the full complexity of real-world preferences. Second, our analysis is limited to three voting rules, and extending this to other systems would be valuable. Finally, the concept of coalitional manipulation may face criticism due to coordination challenges or the lack of binding agreements among coalition members (see Durand [8, Introduction] for a response to these critiques).

1.5 Roadmap

The rest of the paper is organized as follows. Section 2 introduces key definitions and notations. Sections 3, 4, and 5 respectively analyze Plurality, Two-Round System, and IRV. Section 6 explores convergence speed. Section 7 concludes with future work.

2 Definitions and Notations

2.1 Discrete and Continuous Profiles

A *discrete profile* P consists of three elements: a finite, non-empty set of candidates $\mathcal{C}(P)$ with cardinality $m(P)$; a finite, non-empty set of voters $\mathcal{V}(P)$ with cardinality $n(P)$; and for each voter $v \in \mathcal{V}(P)$, a preference ranking P_v over the candidates in $\mathcal{C}(P)$.

For any preference ranking p, let $w(p, P)$ denote the *weight* of p in P, i.e., the number of voters in P with ranking p. The total weight of a discrete profile is simply the number of voters: $w(P) = \sum_p w(p, P) = n(P)$.

A *continuous profile* is similarly defined by three components: a finite, non-empty set of candidates $\mathcal{C}(P)$ with cardinality $m(P)$; a total weight $w(P) \in (0, \infty)$; and for each ranking p over the candidates, a weight $w(p, P) \in \mathbb{R}$, such that $\sum_p w(p, P) = w(P)$.

For any profile P, whether discrete or continuous, we define the associated *normalized profile* $\bar{P}$ as the continuous profile where the weight of each ranking p is given by $w(p, \bar{P}) = \frac{w(p,P)}{w(P)}$. Viewing a profile as a vector of weights, we can naturally define its *neighborhood* in the usual topological sense.

For any subset $K \subseteq \mathcal{C}(P)$, let P_K be the restriction of P to the candidates in K. For two distinct candidates c and d, let $P^{c \succ d}$ be the restriction to voters who prefer c over d. Similarly, for a candidate c and a position $k \in \{1, \ldots, m(P)\}$, let $P^{r(c)=k}$ be the restriction to voters ranking c in the k-th position. These notations can be combined to restrict the profile both by candidates and voters.

2.2 Voting Rules

A *voting rule* f maps any profile, discrete or continuous, to a candidate from that profile. In this paper, we focus on *homogeneous* voting rules, meaning that $f(P) = f(\bar{P})$ for any profile P. In other words, the outcome depends only on the relative proportions of preference rankings, not the total weight. Each particular voting rule is formally defined at the beginning of its respective section.

2.3 Coalitional Manipulability

When P is a discrete profile, we say that a voting rule f is *coalitionally manipulable* (also abbreviated as CM) in P, or that profile P is CM in rule f, if there exists a *target profile* Q with the same candidates and voters such that $f(Q) \neq f(P)$, and for every voter $v \in \mathcal{V}(P)$, if $Q_v \neq P_v$, then v prefers $f(Q)$ to $f(P)$ based on P_v. In other words, only voters who benefit from the new outcome may alter their ballots, though some may keep their original votes.

An immediate consequence is as follows. If for a ranking p, we have $w(p, Q) < w(p, P)$, then at least one voter with ranking p in P must have changed their ballot in Q, i.e., $Q_v \neq P_v$. By the definition, this implies that $f(Q)$ is preferred to $f(P)$ according to $P_v = p$. This observation will now serve as the basis for defining CM in the continuous case.

For a continuous profile P, we say that a voting rule f is CM in P (or that P is CM in f) if there exists a target profile Q with the same candidates and total weight such that $f(Q) \neq f(P)$ and, for every ranking p, if $w(p, Q) < w(p, P)$, then $f(Q)$ is preferred to $f(P)$ according to p. In other words, only voters (in a continuous sense) who prefer the new outcome can have changed their ballots.

The relationship between the two notions is clarified by:

Lemma 1. *If a homogeneous rule f is CM in a discrete profile P, then f is also CM in the corresponding normalized profile $\bar{P}$. However, the converse is not true.*

The direct implication follows from the definitions, so we will focus on providing a counterexample to show that the converse does not hold. Consider the *positional scoring rule* f with weights $(7, 6, 0, \ldots, 0)$, where each candidate's score is given by $s(c) = 7w(P^{r(c)=1}) + 6w(P^{r(c)=2})$, and the candidate with the highest score wins (using a tie-breaking rule if needed). Now, consider a discrete profile P with 8 voters and 3 candidates:

- 3 voters have the ranking $1 \succ 3 \succ 2$,
- 5 voters have the ranking $2 \succ 1 \succ 3$.

It is straightforward to verify that candidate 1 wins under f, and that the rule is CM in the normalized profile $\bar{P}$, but not in the original discrete profile P. The issue is that the 5 manipulators supporting candidate 2 must carefully distribute their points between candidates 1 and 3, which is impossible in the discrete case because each manipulator must assign their entire vote to one ranking rather than splitting it fractionally.

2.4 Perturbed Culture

Given two positive integers m and n, and a *concentration parameter* $\theta \in (0, 1]$, the *Perturbed Culture* model is defined as follows. A discrete profile P is randomly generated with $\mathcal{C}(P) = \{1, \ldots, m\}$ and $\mathcal{V}(P) = \{1, \ldots, n\}$. Each voter is independently assigned the ranking $(1 \succ \ldots \succ m)$ with probability θ, and a uniformly random ranking with probability $1 - \theta$.

As $\theta \to 0$, this model converges to the classical *Impartial Culture* model, while for $\theta = 1$, it becomes a deterministic culture where all voters share the ranking $(1 \succ \ldots \succ m)$.[1]

Since a profile can be represented as a vector giving the weight of each ranking, we can define the *expected normalized profile* (or simply the *expected profile*) under Perturbed Culture. To simplify notation, we denote it by $\hat{P}$, leaving its dependency on m and θ implicit. In this profile, the ranking $(1 \succ \ldots \succ m)$ has a weight of $\theta + \frac{1-\theta}{m!}$, while each of the other rankings has a weight of $\frac{1-\theta}{m!}$.

2.5 CM Rate

We denote by $\rho(f, m, n, \theta)$ the *CM rate*, i.e., the probability that a voting rule f is CM in a profile drawn from the Perturbed Culture model with m candidates, n voters and concentration θ.

3 Plurality

We will begin our study with the *Plurality* voting rule, which assigns each candidate c in a profile P a score equal to the total weight of voters ranking c first: $s_{\text{Plu}}(c, P) = w(P^{r(c)=1})$. The winner is the candidate with the highest score (using a tie-breaking rule if needed): $\text{Plu}(P) = \arg\max s_{\text{Plu}}(c, P)$. The specific tie-breaking method will not affect our findings.

3.1 Theoretical Results for Plurality

The intuition behind our theoretical results is as follows. First, we analyze Plurality's behavior in the expected normalized profile $\hat{P}$ as a function of θ. For small θ, Plurality is CM in this profile, but for

[1] We exclude the case $\theta = 0$ from our theoretical analysis to simplify the proofs: as $n \to \infty$, assuming $\theta > 0$ guarantees that candidate 1 wins under sincere voting for all three voting rules considered. Nevertheless, our results hold even in the case $\theta = 0$, and we will also include it in our figures.

large enough θ, it is not. Using the weak law of large numbers, we then show that as $n \to \infty$, the normalized random profile $\bar{P}$ will, with high probability (i.e., with a probability that tends to 1 when $n \to \infty$), be close enough to $\hat{P}$, ensuring that Plurality behaves similarly. Throughout this subsection, we assume $m \geq 2$.

We begin by analyzing the expected profile $\hat{P}$. In this profile, the plurality score for candidate 1 is $s_{\text{Plu}}(1, \hat{P}) = \theta + \frac{1-\theta}{m}$, while the number of voters inclined to manipulate for any candidate $c \neq 1$ is $w(\hat{P}^{c \succ 1}) = \frac{1-\theta}{2}$. If all manipulators vote optimally for c, they succeed if $w(\hat{P}^{c \succ 1}) > s_{\text{Plu}}(1, \hat{P})$, which simplifies to $\theta < \frac{m-2}{3m-2}$. Defining the *critical value* $\theta_c(\text{Plu}, m) = \frac{m-2}{3m-2}$, we conclude that Plurality is CM for $\theta < \theta_c(\text{Plu}, m)$ and not CM for $\theta > \theta_c(\text{Plu}, m)$ (the equality case is not needed for our forthcoming analysis).

We now apply the weak law of large numbers to show that as $n \to \infty$, these results hold with high probability. We start by examining the *supercritical regime* $\theta > \theta_c(\text{Plu}, m)$, relying on the following lemma.

Lemma 2. *Assume there exists a neighborhood of the expected normalized profile $\hat{P}$ where the homogeneous rule f is not CM. Then* $\lim_{n\to\infty} \rho(f, m, n, \theta) = 0$.

Proof. Applying the weak law of large numbers, the following statements hold with high probability: denoting P the random profile, its normalized version $\bar{P}$ lies in the desired neighborhood of $\hat{P}$, hence (by assumption) f is not CM in $\bar{P}$, hence (by Lemma 1) f is also not CM in the random discrete profile P. □

This lemma applies easily to Plurality. For $\theta > \theta_c(\text{Plu}, m)$, we have shown that for every candidate $c \neq 1$, $w(\hat{P}^{c \succ 1}) < s_{\text{Plu}}(1, \hat{P})$. As this is a strict inequality, it holds in a neighborhood of the profile, allowing us to apply Lemma 2. Hence, $\lim_{n\to\infty} \rho(\text{Plu}, m, n, \theta) = 0$.

We now turn to the *subcritical regime* $\theta < \theta_c(\text{Plu}, m)$. Unfortunately, we cannot directly apply the same reasoning: even if the normalized profile $\bar{P}$ is CM near $\hat{P}$, it does not necessarily follow that the discrete profile P is also CM, as Lemma 1 does not hold in the reverse direction.

However, for Plurality, manipulators can always employ a common strategy. Formally, a voting rule f is *unison-manipulable* (UM) in profile P (or P, in f) if manipulation can succeed even when all interested voters cast the same ballot [33, 9].[2] Clearly, UM implies CM. Unlike CM, UM holds equivalently for both a discrete profile P and its normalized profile $\bar{P}$, which leads to the following lemma.

Lemma 3. *Assume there exists a neighborhood of the expected normalized profile $\hat{P}$ where the homogeneous rule f is UM. Then* $\lim_{n\to\infty} \rho(f, m, n, \theta) = 1$.

The proof is similar to Lemma 2: by the weak law of large numbers, with high probability, the normalized random profile $\bar{P}$ is in the desired neighborhood, making it UM, hence the random discrete profile P is also UM, and thus CM. Applied to Plurality, Lemma 3 directly leads to $\lim_{n\to\infty} \rho(f, m, n, \theta) = 1$ for $\theta < \theta_c(\text{Plu}, m)$.

The following theorem summarizes our results so far.

Theorem 1. *Let $\theta_c(\text{Plu}, m) = \frac{m-2}{3m-2}$ with $m \geq 2$.*

- *If $\theta < \theta_c(\text{Plu}, m)$, then* $\lim_{n\to\infty} \rho(\text{Plu}, m, n, \theta) = 1$.
- *If $\theta > \theta_c(\text{Plu}, m)$, then* $\lim_{n\to\infty} \rho(\text{Plu}, m, n, \theta) = 0$.

[2] The term *unison* was introduced by Walsh [33] but we follow the slightly different definition proposed by Durand [9].

For $m = 2$, the theorem indicates $\theta_c(\text{Plu}, 2) = 0$, which is expected, as Plurality cannot be manipulated with only two candidates. Similarly, for $m = 1$, we would reach the same conclusion by conventionally setting $\theta_c(\text{Plu}, 1) = 0$. The theorem becomes more interesting for $m \geq 3$, where it describes a *phase transition* around $\theta_c(\text{Plu}, m)$, meaning a sudden change in behavior as the parameter crosses this threshold. This raises key questions: What causes this discontinuity, and how do we approach it as n increases? What happens when θ is equal to or near the critical value?

3.2 Simulations for Plurality

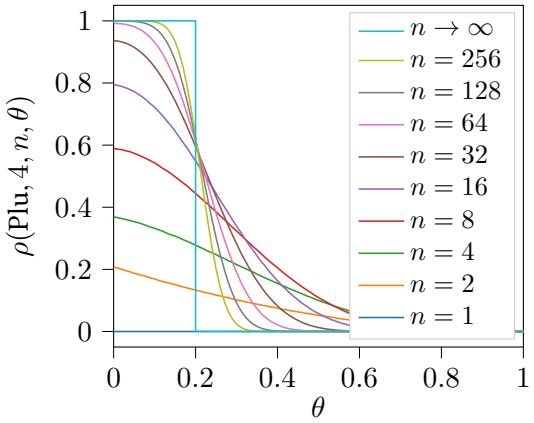

Figure 1: CM rate of Plurality as a function of θ for different values of n with $m = 4$. Curves for finite n are based on Monte Carlo simulations with 1,000,000 profiles per point. The limiting curve as $n \to \infty$ follows from Theorem 1.

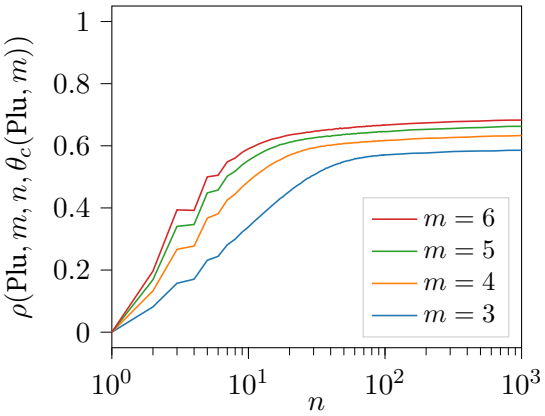

Figure 2: CM rate of Plurality as a function of n for different values of m with $\theta = \theta_c(\text{Plu}, m)$. Monte Carlo simulations with 1,000,000 profiles per point.

To understand the origin of the discontinuity, Figure 1 shows the CM rate of Plurality as a function of θ for various n with $m = 4$. Curves for finite n are based on Monte Carlo simulations,[3] with 1,000,000 profiles per point, leading to error margins of $\frac{1}{\sqrt{1000000}} = 0.1\%$. The limiting curve for $n \to \infty$ is derived from Theorem 1. For finite n, the curve is continuous. As n increases, it becomes sigmoid-shaped and steepens, ultimately converging to a step function as $n \to \infty$.

The observed behavior mirrors what occurs in physics: since a finite combination of continuous functions remains continuous, non-analyticity can only arise in an infinite system [21, Section 11.6]. As in physics, a phase transition occurs beyond a certain level of disorder: while a ferromagnetic metal loses its magnetization above the Curie temperature [4], Plurality loses its resistance to coalitional manipulation below the critical value of the concentration parameter θ.

Theorem 1 describes the behavior in the subcritical and supercritical regimes, but what happens in the *critical regime*, i.e., when $\theta = \theta_c(\text{Plu}, m)$? Figure 2 shows the CM rate in that case as a function of the number of voters n, for different values of m. This leads to several conjectures:

- The *critical CM rate* $\rho(\text{Plu}, m, n, \theta_c(\text{Plu}, m))$ converges to a limit as $n \to \infty$.
- This limit is strictly less than 1.
- This limit increases with m.

It is beyond the scope of this paper to theoretically prove these results. In the study of phase transitions, analyzing the critical behavior is often challenging [29, 6, 24, 23]. With that, we conclude our study of Plurality and proceed to the Two-Round System.

[3] The code is available at https://github.com/francois-durand/irv-cm-aamas-2025.

4 Two Round System

The *Two-Round System* (TR) is as follows. In the first round, each candidate c receives a score $s^1_{\text{TR}}(c, P) = s_{\text{Plu}}(c, P)$, and the set K of the two candidates with the highest scores advances to the second round. These two candidates then receive scores $s^2_{\text{TR}}(c, P) = s_{\text{Plu}}(c, P_K)$, and the candidate with the highest score wins. A tie-breaking rule is applied if necessary.[4]

4.1 Theoretical Results for the Two-Round System

For $n = 2$, the Two-Round System is equivalent to Plurality, so we focus on the case $m \geq 3$.

As with Plurality, we begin by examining the expected normalized profile $\hat{P}$. Candidate 1 clearly wins the election, with the second-round opponent determined by the tie-breaking rule. For a manipulation to succeed in favor of a candidate $c \neq 1$, candidate c must reach the second round. However, if candidate 1 also advances, the manipulation will fail. Therefore, the second round must involve a candidate $d \notin \{1, c\}$, which is possible since we assumed $m \geq 3$. Now, consider the portion of the first-round scores for candidates 1, c, and d coming from sincere voters:

$$\begin{cases} s^1_{\text{TR}}(1, \hat{P}^{1>c}) = \theta + \frac{1-\theta}{m}, \\ s^1_{\text{TR}}(c, \hat{P}^{1>c}) = 0, \\ s^1_{\text{TR}}(d, \hat{P}^{1>c}) = \frac{1-\theta}{2m}, \end{cases}$$

where, for example, $s^1_{\text{TR}}(d, \hat{P}^{1>c})$ denotes the first-round score, in the two-round system, of candidate d in the restriction of the expected normalized profile $\hat{P}$ to the voters who prefer candidate 1 to candidate c (i.e., "sincere" voters).

For both candidates c and d to surpass candidate 1's score, at least $\theta + \frac{1-\theta}{m}$ manipulators must vote for c, while $\theta + \frac{1-\theta}{m} - \frac{1-\theta}{2m}$ must vote for d. Therefore, the total number of manipulators, given by $w(\hat{P}^{c\succ 1}) = \frac{1-\theta}{2}$, must be at least the sum of these two quantities. Simplifying, the necessary condition becomes $\theta \leq \frac{m-3}{5m-3}$. In other words, coalitional manipulation is impossible for $\theta > \frac{m-3}{5m-3}$. In our original paper [10], we easily show that, conversely, when $\theta < \frac{m-3}{5m-3}$, manipulation is possible in the expected profile $\hat{P}$.

To extend this result to a random profile P with high probability, we follow the same general strategy as for Plurality, but with a technical glitch: unison manipulation is generally insufficient under the Two-Round System. To overcome this, we introduce a notion of stability in the manipulation outcome around a given profile (see [10] for more details). This leads to the following theorem.

Theorem 2. *Let* $\theta_c(\text{TR}, m) = \frac{m-3}{5m-3}$ *with* $m \geq 3$.

- *If* $\theta < \theta_c(\text{TR}, m)$, *then* $\lim_{n\to\infty} \rho(\text{TR}, m, n, \theta) = 1$.
- *If* $\theta > \theta_c(\text{TR}, m)$, *then* $\lim_{n\to\infty} \rho(\text{TR}, m, n, \theta) = 0$.

Recall that for $m = 2$, the same conclusions hold by setting $\theta_c(\text{TR}, 2) = 0$, since Two-Round is equivalent to Plurality in this case. For $m = 3$, the theorem also gives $\theta_c(\text{TR}, 3) = 0$, which is remarkable: according to the Gibbard-Satterthwaite theorem, manipulability becomes an issue from $m = 3$, yet Two-Round avoids this with high probability in Perturbed Culture as soon as $\theta > 0$.

Since Theorems 1 and 2 share similar structures, a natural question arises: does every voting rule f have a critical parameter $\theta_c(f, m)$ with similar properties? The answer is no. Consider a rule f that

[4] For simplicity, we consider an "instant" version of TR, where voters cast their ballots once. In most actual implementations, voters participate in two rounds. While this is equivalent for sincere voting, the instant version restricts some manipulation strategies [8, Table 1.1]. However, our results apply to both variants.

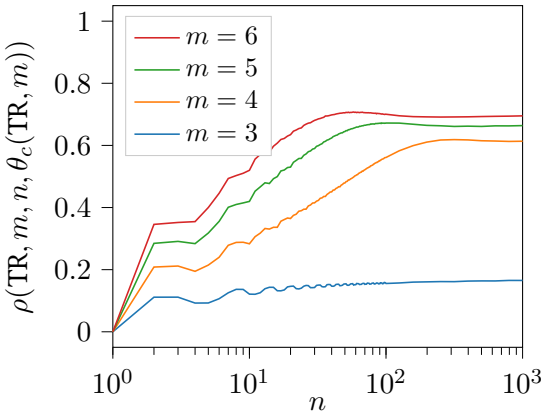

Figure 3: CM rate of the Two-Round System as a function of n for different values of m with $\theta = \theta_c(\mathrm{TR}, m)$. Monte Carlo simulations with 1,000,000 profiles per point.

uses Plurality when n is even and Two-Round when n is odd.[5] From Theorems 1 and 2, it follows that for $\theta < \frac{m-3}{5m-3}$, the CM rate converges to 1, while for $\theta > \frac{m-2}{3m-2}$, it converges to 0. However, for $\theta \in \left(\frac{m-3}{5m-3}, \frac{m-2}{3m-2}\right)$, the CM rate tends to 1 for even n and to 0 for odd n: overall, it does not converge.

It is still possible to define a lower critical value $\theta_l(f, m)$ and an upper critical value $\theta_u(f, m)$ as, respectively, the largest and smallest values in $[0, 1]$ such that:

- If $\theta < \theta_l(f, m)$, then $\lim_{n\to\infty} \rho(f, m, n, \theta) = 1$,
- If $\theta > \theta_u(f, m)$, then $\lim_{n\to\infty} \rho(f, m, n, \theta) = 0$.

We can then define $\theta_c(f, m)$ as their common value when it exists. With this convention, Theorems 1 and 2 are summarized as:

$$\theta_c(\mathrm{Plu}, m) = \frac{m-2}{3m-2}, \quad \theta_c(\mathrm{TR}, m) = \frac{m-3}{5m-3}.$$

4.2 Simulations for the Two-Round System

The Two-Round equivalent of Figure 1 is similar, so we proceed directly to the counterpart of Figure 2: Figure 3, showing the critical CM rate as a function of n for different m. We use SVVAMP 0.12.0 [12], a Python package for studying the manipulability of voting rules. As for Plurality, the critical CM rate appears to converge to a limit in $(0, 1)$ that increases with m.

5 IRV (Instant-Runoff Voting)

Let us now proceed to *Instant-Runoff Voting* (IRV), where the winner is determined through multiple rounds. In each round, the candidate with the lowest Plurality score is eliminated, until only one remains. Formally, let $K(r, P)$ be the set of remaining candidates at the start of round r and $\ell(r, P)$ be the candidate losing at round r. We have:

$$\begin{cases} K(1, P) = \mathcal{C}(P), \\ \ell(r, P) = \arg\min s_{\mathrm{Plu}}(c, P_{K(r,P)}), \\ K(r+1, P) = K(r, P) \setminus \{\ell(r, P)\}, \end{cases}$$

[5] For simplicity, this counter-example involves a non-homogeneous rule.

using a tie-breaking rule for elimination when necessary. The winner IRV(P) is the last remaining candidate in $K(m(P), P)$.

5.1 Theoretical Results for IRV

As usual, we start by examining the expected normalized profile $\hat{P}$. Since $\theta > 0$, candidate 1 clearly wins. Now suppose that IRV is CM in $\hat{P}$ to a target profile Q, where candidate $c \neq 1$ wins. Candidate 1 must be eliminated in some round r. For conciseness, denote $K = K(r, Q)$ and $k = |K|$. Obviously c must belong to K. The sincere voters' contribution to candidate 1's score at this round is:

$$s_{\text{Plu}}(1, \hat{P}_K^{1 \succ c}) = s_{\text{Plu}}(1, \hat{P}_K) = \theta + \frac{1-\theta}{k}.$$

Thus, $s_{\text{Plu}}(1, Q_K) \geq \theta + \frac{1-\theta}{k}$, and since $k \geq 2$, this is strictly greater than $\frac{1}{k}$. Therefore, the score of candidate 1 exceeds the average score at this round, hence it cannot be minimal. This contradiction proves that IRV is not CM in $\hat{P}$.

In this reasoning, IRV's resistance to coalitional manipulation stems from the fact that in any subset of candidates K containing candidate 1, this candidate has a Plurality score that exceeds the average score. This motivates the following definition:

Definition 1. *A candidate c is a* Super Condorcet Winner *(SCW) in a profile P if, for every subset of candidates K containing c, the following holds:*

$$s_{\text{Plu}}(c, P_K) > \frac{w(P)}{|K|}.$$

This concept strengthens the classical notion of a *Condorcet Winner*, which only requires the condition to hold for subsets K of size 2. We summarize its relevance to IRV as follows:

Lemma 4. *If c is an SCW in profile P, then* IRV$(P) = c$ *and IRV is not CM in P.*

The same result easily extends to several IRV variants, such as Exhaustive Ballot [9], Condorcet-IRV [19, 11], Benham rule, Tideman rule, Smith-IRV, and Woodall rule [17]. However, the converse is not true: there exists profiles without an SCW where IRV is still not CM (see Durand [8, Table 1.1] for an example).

Now, consider the neighborhood of the expected profile $\hat{P}$. Since the SCW condition involves a finite number of strict inequalities that depend continuously on the profile's coefficients, candidate 1 is an SCW not only in $\hat{P}$ but also in its neighborhood. From here, we can follow two proof strategies that only differ in the order of their steps.

One approach is to first apply Lemma 4 to deduce that IRV is not CM in this neighborhood. Using Lemma 2 (based on the weak law of large numbers), we then deduce $\lim_{n\to\infty} \rho(\text{IRV}, m, n, \theta) = 0$. Alternatively, we could first use the weak law of large numbers to show that candidate 1 is an SCW with high probability, then apply Lemma 4 to show $\lim_{n\to\infty} \rho(\text{IRV}, m, n, \theta) = 0$.

Since this holds for any $\theta > 0$, we obtain a remarkable result:

Theorem 3. *For IRV, the critical value of the concentration parameter in Perturbed Culture is*

$$\theta_c(\text{IRV}, m) = 0.$$

In summary, within the Perturbed Culture model, IRV has the smallest possible critical value. Even a slight concentration of preferences favoring candidate 1 is enough for IRV to become resistant to coalitional manipulation with high probability. And for the same reasons, this also holds for the IRV variants mentioned earlier.

5.2 Simulations for IRV

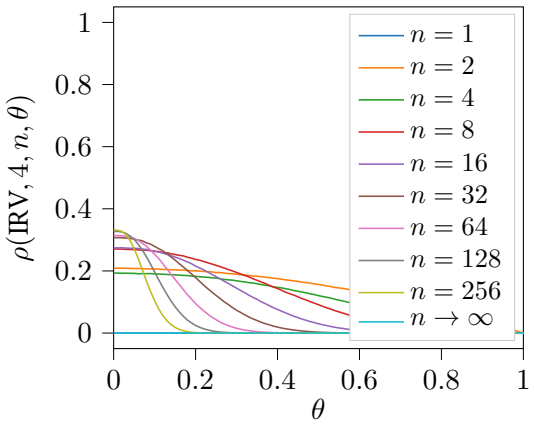

Figure 4: CM rate of IRV as a function of θ for different values of n with $m = 4$. Curves for finite n are based on Monte Carlo simulations with 1,000,000 profiles per point. The limiting curve as $n \to \infty$ follows from Theorem 3.

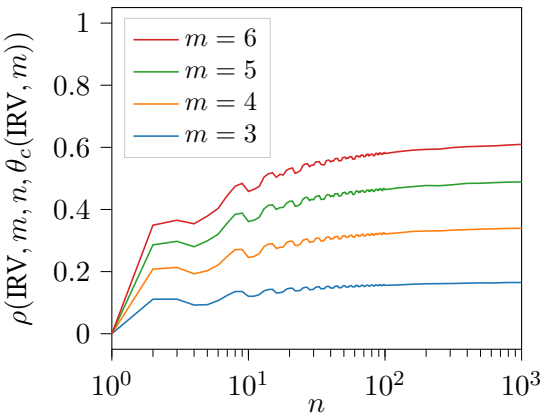

Figure 5: CM rate of IRV as a function of n for different values of m with $\theta = \theta_c(\text{IRV}, m) = 0$ (Impartial Culture). Monte Carlo simulations with 1,000,000 profiles per point.

For IRV, as for Two-Round, our simulations for finite n are carried out using SVVAMP. Similar to Figure 1 for Plurality, Figure 4 shows the CM rate as a function of θ for different n. For large n, the curve takes on a sigmoidal shape that converges to the theoretical curve from Theorem 3. The behavior near $\theta = 0$ suggests that in the Impartial Culture model, the CM rate converges to a limit within $(0, 1)$, as proven for $m = 3$ by Lepelley and Valognes [26], conjectured in the general case by Durand [8, Conjecture 7.8], and further supported by Figure 5. This second figure also indicates that, as with Plurality and Two-Round, the limit CM rate in the critical regime appears to increase with the number of candidates m.

5.3 Empirical Results for IRV

In the Perturbed Culture model, the presence of an SCW explains IRV's resistance to coalitional manipulation. However, as noted earlier, this is not a necessary condition: IRV can be non-manipulable in profiles without an SCW. This raises the question of whether the presence of an SCW often accounts for IRV's non-manipulability in realistic scenarios.

	Netflix dataset	*FairVote dataset*
Profiles	11,215	10,044
— with a CW	99.30%	99.98%
— where IRV is not CM (a)	95.87%	96.30%
— with an SCW (b)	94.05%	96.20%
Ratio (b) / (a)	98%	> 99%*

* We omit the next digit of the raw result (99.9%), not significant given the sample size.

Table 1: Empirical study of Super Condorcet Winners (SCW) and IRV in two datasets, with the presence of a Condorcet Winner (CW) included as a reference.

To investigate this, Table 1 analyzes the Netflix and FairVote datasets [9], which respectively contain 11,215 profiles derived from slight perturbations of 2,243 empirical profiles and 10,044 profiles based

on 162 empirical profiles. It provides two key insights. First, an SCW is very common in real-world datasets, here appearing in 94% or 96% of profiles. Second, in most cases where IRV resists CM, this can be explained by the presence of an SCW—98% in the Netflix dataset and over 99% in the FairVote dataset. This confirms that SCWs are a crucial factor in IRV's resilience to manipulation.

The frequent appearance of SCWs may seem surprising, but it becomes intuitive when revisiting the definition. For a candidate c and a subset K of candidates that includes c, if preferences were perfectly balanced, we would expect $s_{\text{Plu}}(c, P_K) = \frac{w(P)}{|K|}$. The condition for being an SCW is simply to exceed this average. Therefore, even a slight bias in favor of c makes it likely for c to be an SCW.

6 Convergence Speed

We will now study the convergence speed, to assess how fast the results found for $n \to \infty$ become relevant for finite values of n.

6.1 Theoretical Bound

We will first show that in the non-critical regime, the CM rate converges exponentially fast as $n \to \infty$. Next, we will bound this speed of convergence depending on the parameter θ.

As an example, consider Lemma 2, where we assume the existence of a neighborhood where the rule f is not CM. By definition, there exists $\epsilon > 0$ such that this neighborhood contains an open ball of diameter ϵ for the infinity norm. Our approach is to apply Hoeffding's inequality [20] to bound the probability that the normalized random profile $\bar{P}$ falls outside this ball. Since Hoeffding's inequality applies to scalar random variables, we use the union bound to extend it to the weight vector representing a voting profile. Formally, denoting by $\mathbb{P}$ the probability:

$$\begin{aligned}\rho(f, m, n, \theta) &\leq \mathbb{P}(d(\bar{P}, \hat{P}) \geq \epsilon), \\ &\leq \textstyle\sum_p \mathbb{P}(|w(p, \bar{P}) - w(p, \hat{P})| \geq \epsilon) \quad \text{(union bound)}, \\ &\leq 2m! e^{-2\epsilon^2 n} \quad \text{(Hoeffding's inequality)}. \end{aligned} \tag{1}$$

Thus, the convergence is exponentially fast in n, and we can quantify the rate if the size of the neighborhood is known. By the same reasoning, similar results hold for Lemma 3 and Theorems 1, 2, and 3. In the literature on phase transitions, this is known as *sharp* transitions, meaning that the limiting curve quickly approximates the behavior even for finite n.

Let us now bound the speed of convergence more precisely. For example, consider Plurality in the supercritical regime. If a profile P lies within an open ball of radius ϵ centered at $\hat{P}$, the score of candidate 1 is bounded from below: $s_{\text{Plu}}(1, P) > \theta + \frac{1-\theta}{m} - m!\epsilon$, and the number of manipulators is bounded from above: $w(P^{2\succ 1}) < \frac{1-\theta}{2} + m!\epsilon$. To ensure $s_{\text{Plu}}(1, P) > w(P^{2\succ 1})$, we set $\epsilon = \frac{(3m-2)\theta - (m-2)}{2m!}$, which can be rewritten as $\epsilon = \big(\theta - \theta_c(\text{Plu}, m)\big)\frac{3m-2}{2m!}$. Using our bound (1), there exists a coefficient $A^+(\text{Plu}, m)$—which we could explicitly compute—such that:

$$\rho(\text{Plu}, m, n, \theta) = O\left(e^{-A^+(\text{Plu},m)(\theta - \theta_c(\text{Plu},m))^2}\right).$$

This reasoning for Plurality in the supercritical regime generalizes to the subcritical regime, with a coefficient $A^-(\text{Plu}, m)$, and to the other voting rules in this paper: since all relevant quantities (scores, numbers of manipulators) are linear in the profile weights, we can take a value of ϵ that depends linearly on $\theta - \theta_c$, leading to a term in $(\theta - \theta_c)^2$ via Hoeffding's inequality. Thus, we obtain:

- Supercritical regime: $\rho = O\big(e^{-A^+(f,m)(\theta-\theta_c)^2 n}\big)$,

- Subcritical regime: $\rho = 1 - O\left(e^{-A^-(f,m)(\theta_c-\theta)^2 n}\right)$,

where $\rho = \rho(f, m, n, \theta)$ and $\theta_c = \theta_c(f, m)$.

6.2 Simulation Study of the Convergence Speed

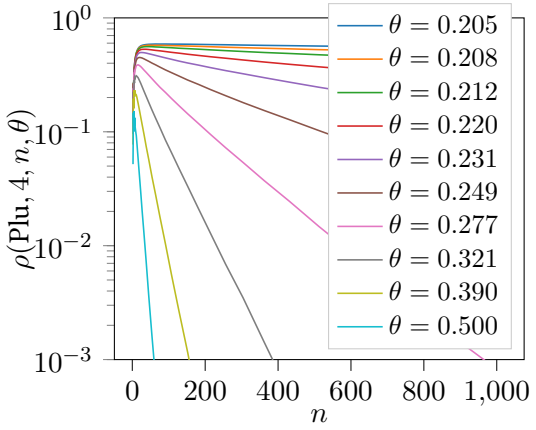

Figure 6: CM rate of Plurality as a function of n for different supercritical values of θ with $m = 4$. Monte Carlo simulations with 1,000,000 profiles per point.

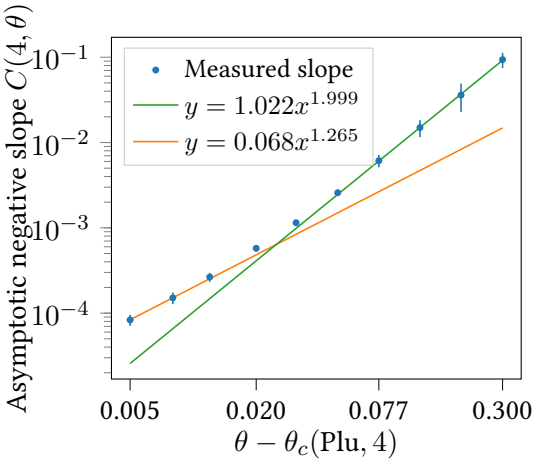

Figure 7: Asymptotic negative slope $C(4, \theta)$ from Figure 6, plotted as a function of $\theta - \theta_c(\text{Plu}, 4)$. The vertical blue lines represent error margins.[6]

Figure 6 shows the CM rate of Plurality as a function of n for various values of θ (whereas Figure 1 does the reverse). Each curve has an oblique asymptote on a semi-log scale, indicating not only that it is bounded by a decreasing exponential (as predicted by theory), but that it follows the form $\rho \sim_{n\to\infty} B(m, \theta)e^{-C(m,\theta)n}$, with $m = 4$ here. This figure also allows us to measure the slopes of the asymptotes, providing the values of $C(4, \theta)$ for each θ.

To analyze how convergence speed varies with θ, Figure 7 plots the measured asymptotic slopes against $\theta - \theta_c$ in log-log scale (the values of θ were specifically chosen to be evenly spaced in that figure). When θ is far from θ_c, the dependency is in $(\theta - \theta_c)^2$, in line with the upper bound found previously. However, close to θ_c, the dependency seems to involve a smaller exponent (estimated at 1.265). In the terminology of phase transition, this is called the *critical exponent* of the convergence speed.

We repeated this for $m \in \{5, 6, 7\}$, the subcritical regime, and the other voting rules, with similar results but various critical exponents. This suggests a long-range dependency in $|\theta - \theta_c|^2$ but smaller critical exponents near the critical regime. This intriguing behavior will deserve further theoretical investigation.

7 Future Work

A natural direction for future work is to compute the critical parameter θ_c for other voting systems. Another key area of research would be a deeper analysis of the critical regime, including the calculation of the limiting CM rate at $\theta = \theta_c$ and the asymptotic behavior of the slope of the sigmoid $\rho(\theta)$ at $\theta = \theta_c$, which is linked to a finer analysis of the convergence speed in the non-critical regime. Expanding the study to other models, such as Mallows, is also promising. Preliminary analysis shows that the qualitative results observed in this paper, particularly the key finding that IRV's limit CM rate drops to zero with even slight concentration of preferences, also hold true under the Mallows model.

[6]The computation of the error margins is detailed in the code repository https://github.com/francois-durand/irv-cm-aamas-2025.

References

[1] Béla Bollobás and Oliver Riordan. *Percolation*. Cambridge University Press, 2006.

[2] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016.

[3] John Chamberlin, Jerry Cohen, and Clyde Coombs. Social choice observed: Five presidential elections of the american psychological association. *The Journal of Politics*, 46(2):479–502, 1984.

[4] Sóshin Chikazumi. *Physics of ferromagnetism*. Oxford University Press, 1997.

[5] Kim Christensen. Percolation theory. *Imperial College London*, 1:87, 2002.

[6] Hugo Duminil-Copin. Lectures on the ising and potts models on the hypercubic lattice. In *PIMS-CRM Summer School in Probability*, pages 35–161. Springer, 2017.

[7] Hugo Duminil-Copin. Sixty years of percolation. In *Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018*, pages 2829–2856, 2018.

[8] François Durand. *Towards less manipulable voting systems*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2015.

[9] François Durand. Coalitional manipulation of voting rules: Simulations on empirical data. *Constitutional Political Economy*, 34(3):390–409, 2023.

[10] François Durand. Why instant-runoff voting is so resilient to coalitional manipulation: Phase transitions in the perturbed culture. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025)*, 2025.

[11] François Durand, Fabien Mathieu, and Ludovic Noirie. Can a Condorcet rule have a low coalitional manipulability? In *European Conference on Artificial Intelligence (ECAI)*, volume 285, pages 707–715, 2016.

[12] François Durand, Fabien Mathieu, and Ludovic Noirie. Svvamp: Simulator of various voting algorithms in manipulating populations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[13] Andrew Eggers and Nick Vivyan. Who votes more strategically? *American Political Science Review*, 114(2):470–485, 2020.

[14] Pierre Favardin, Dominique Lepelley, and Jérôme Serais. Borda rule, Copeland method and strategic manipulation. *Review of Economic Design*, 7:213–228, 2002.

[15] William Gehrlein. *Condorcet's Paradox*. Springer, 2006.

[16] Allan Gibbard. Manipulation of voting schemes: A general result. *Econometrica*, 41(4):587–601, 1973.

[17] James Green-Armytage. Four Condorcet-Hare hybrid methods for single-winner elections. *Voting matters*, 29(1):1–14, 2011.

[18] James Green-Armytage. Strategic voting and nomination. *Social Choice and Welfare*, 42(1):111–138, 2014.

[19] James Green-Armytage, Nicolaus Tideman, and Rafael Cosman. Statistical evaluation of voting rules. *Social Choice and Welfare*, 46:183–212, 2016.

[20] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

[21] Leo Kadanoff. *Statistical physics: Statics, dynamics and renormalization*. World Scientific, 2000.

[22] K.H. Kim and F.W. Roush. Statistical manipulability of social choice functions. *Group Decision and Negotiation*, 5:263–282, 1996.

[23] Zohar Komargodski and David Simmons-Duffin. The random-bond ising model in 2.01 and 3 dimensions. *Journal of Physics A: Mathematical and Theoretical*, 50(15):154001, 2017.

[24] Filip Kos, David Poland, David Simmons-Duffin, and Alessandro Vichi. Precision islands in the ising and o(n) models. *Journal of High Energy Physics*, 2016(8):1–16, 2016.

[25] Dominique Lepelley and Boniface Mbih. The vulnerability of four social choice functions to coalitional manipulation of preferences. *Social Choice and Welfare*, 11:253–265, 1994.

[26] Dominique Lepelley and Fabrice Valognes. On the Kim and Roush voting procedure. *Group*

Decision and Negotiation, 8:109–123, 1999.

[27] Dominique Lepelley and Fabrice Valognes. Voting rules, manipulability and social homogeneity. *Public Choice*, 116:165–184, 2003.

[28] Elchanan Mossel, Ariel D Procaccia, and Miklós Z Rácz. A smooth transition from powerlessness to absolute power. *Journal of Artificial Intelligence Research*, 48:923–951, 2013.

[29] Mark Newman and Robert Ziff. Efficient monte carlo algorithm and high-precision results for percolation. *Physical Review Letters*, 85(19):4104, 2000.

[30] Bezalel Peleg. A note on manipulability of large voting schemes. *Theory and Decision*, 11:401–412, 1979.

[31] Mark Satterthwaite. 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, 1975.

[32] Arkadii Slinko. On asymptotic strategy-proofness of classical social choice rules. *Theory and Decision*, 52:389–398, 2002.

[33] Toby Walsh. An empirical study of the manipulability of single transferable voting. In *European Conference on Artificial Intelligence (ECAI)*, pages 257–262. 2010.

[34] Oliver Williamson and Thomas Sargent. Social choice: A probabilistic approach. *The Economic Journal*, 77(308):797–813, 1967.

[35] Lirong Xia. The impact of a coalition: Assessing the likelihood of voter influence in large elections. *arXiv preprint arXiv:2202.06411*, 2022.

François Durand
Nokia Bell Labs France
Massy, France
Email: francois.durand@nokia-bell-labs.com