The Moderating Effect of Instant Runoff Voting

Kiran Tomlinson[1], **Johan Ugander**[2], **Jon Kleinberg**[1]

[1]Cornell University
[2]Stanford University
kt@cs.cornell.edu, jugander@stanford.edu, kleinberg@cornell.edu

Abstract

Instant runoff voting (IRV) has recently gained popularity as an alternative to plurality voting for political elections, with advocates claiming a range of advantages, including that it produces more moderate winners than plurality and could thus help address polarization. However, there is little theoretical backing for this claim, with existing evidence focused on case studies and simulations. In this work, we prove that IRV has a moderating effect relative to plurality voting in a precise sense, developed in a 1-dimensional Euclidean model of voter preferences. We develop a theory of *exclusion zones*, derived from properties of the voter distribution, which serve to show how moderate and extreme candidates interact during IRV vote tabulation. The theory allows us to prove that if voters are symmetrically distributed and not too concentrated at the extremes, IRV cannot elect an extreme candidate over a moderate. In contrast, we show plurality can and validate our results computationally. Our methods provide new frameworks for the analysis of voting systems, deriving exact winner distributions geometrically and establishing a connection between plurality voting and stick-breaking processes.

Introduction

Instant runoff voting (IRV) elections ask voters to rank candidates in order of preference and use a sequence of "instant runoffs" to determine a winner.[1] IRV selects a winner by repeatedly eliminating the candidate with the fewest first-place votes, redistributing those votes to the next-ranked candidate on each ballot, and removing the eliminated candidate from all ballots. The final remaining candidate is declared the winner (equivalently, one can terminate when a majority of the remaining ballots list the winner first). By comparison, in a plurality election the winner is simply the candidate with the most first-place votes. While plurality has historically been the predominant single-winner voting system, IRV is among the most popular alternatives; for instance, Australia and Ireland have used IRV since the early 20th century. In the United States, IRV has recently been gaining traction to address issues with plurality voting (Wang et al. 2021), with three states (Maine, Alaska, and Nevada) voting to adopt IRV for federal elections in the last decade. IRV has also seen increasing adoption in local elections and/or primaries, for instance in San Francisco (since 2004), Minneapolis (since 2009), and New York City (since 2021).

Proponents of IRV claim that it encourages moderation, compromise, and civility, since candidates are incentivized to be ranked highly by as many voters as possible, including by those who do not rank them first (Dean 2016; Diamond 2016). Analyses of campaign communication materials and voter surveys have supported the theory that IRV increases campaign civility (Donovan, Tolbert, and Gracey 2016; John and Douglas 2017; Kropf 2021), with extensive debate about whether this greater civility translates into winners who are also more moderate in their positions (Fraenkel and Grofman 2006a,b; Horowitz 2006, 2007). Analyses of potential moderating effects of IRV have primarily been based on case studies (Fraenkel and Grofman 2004; Mitchell 2014; Reilly 2018) and simulation (Chamberlin and Cohen 1978; Merrill 1984; McGann, Grofman, and Koetzle 2002), as well as empirical evidence for a moderating effect in a related voting system, two-round runoff (Bordignon, Nannicini, and Tabellini 2016). In contrast, there has been almost no theoretical work on the subject; most social choice theory has focused on problems other than moderation, such as minimizing metric distortion and ensuring fairness or representation (Halpern et al. 2023; Aziz et al. 2017; Boutilier et al. 2012; Brill et al. 2022; Ebadian et al. 2022; Gkatzelis, Halpern, and Shah 2020; Kahng, Latifian, and Shah 2023). Two interesting specific exceptions can be found in the works of Grofman and Feld (2004) and Dellis, Gauthier-Belzile, and Oak (2017). Grofman and Feld (2004) show that for single-peaked preferences and four or fewer candidates, IRV is at least as likely as plurality to elect the median candidate. Dellis, Gauthier-Belzile, and Oak (2017) show that in a citizen-candidate model, if the voter distribution is asymmetric then two-party equilibria under plurality can be more extreme than under IRV.

There is clear value in mathematical analyses that identify more general moderating tendencies. At present—beyond the noted exceptions—the arguments for IRV's moderating effects summarized above have tended to point to institutional or behavioral properties of the way candidates run

[1]IRV is also called ranked choice voting in the United States. Other names for IRV include alternative vote, preferential voting, and the Hare method. Multi-winner IRV is also called single transferrable vote. Plurality is also called first-past-the-post.

their campaigns in IRV elections. A natural question, therefore, is whether this picture is complete, or whether there might be something in the definition of IRV itself that leads to outcomes with more moderate winners. Such questions are fundamental to the mathematical theory of voting more generally, where we frequently seek explanations that are rooted in the formal properties of the voting systems themselves, rather than simply the empirical regularities of how candidates and voters tend to behave in these systems. In the case of IRV, what would it mean to formalize a tendency toward moderation in the underlying structure of the voting system? To begin, we must first identify a natural set of definitions under which we can isolate such a property.

Formalizing the moderating effect of IRV. In this paper, we propose such definitions and use them to articulate a precise sense in which IRV produces moderate winners in a way that plurality does not. We work within a standard one-dimensional model of voters and candidates: the positions of voters and candidates correspond to points drawn from distributions on the unit interval $[0, 1]$ of the real line (representing left–right ideology), and voters form preferences over candidates by ranking them in order of proximity. That is, voters favor candidates who are closer to them on the line; this is often called the *1-Euclidean* model, a common model in social choice theory (Coombs 1964; Bogomolnaia and Laslier 2007; Elkind, Lackner, and Peters 2022). We typically assume the voters and candidates are drawn from the same distribution F, but some of our results hold for fixed candidate positions. In addition to its role as one of the classical mathematical models of voter preferences, where it is sometimes called the Hotelling model (Hotelling 1929; Downs 1957), 1-Euclidean preferences arise naturally from higher-dimensional opinions under simple models of opinion updating (DeMarzo, Vayanos, and Zwiebel 2003). There is wide-ranging empirical evidence suggesting that political opinions in the United States are remarkably one-dimensional (Poole and Rosenthal 1984, 1991; Layman, Carsey, and Horowitz 2006; DellaPosta, Shi, and Macy 2015): from a voter's views on any one of a set of issues including tax policy, immigration, climate change, gun control, and abortion, it is possible to predict the others with striking levels of confidence.

Let's consider a voting system applied to a set of k candidates and a continuum of voters in this setting: we draw k candidates independently from a given distribution F on the unit interval $[0, 1]$, and each candidate gets a vote share corresponding to the fraction of voters who are closest to them (see Figure 1 for examples). The use of a one-dimensional model gives a natural interpretation to the distinction between moderate and extreme candidates: a candidate is more extreme if they are closer to the endpoints of the unit interval $[0, 1]$. We take two approaches to defining a moderating effect in this model, one probabilistic (in the limit of large k) and one combinatorial (for all k). We say that a voting system has a *probabilistic moderating effect* if for some interval $I = [a, b]$ with $0 < a \leq b < 1$, the probability that the winning candidate comes from I converges to 1 as the number of candidates k goes to infinity (since we focus on symmetric voter distributions, we will typically have I symmetric about $1/2$; i.e. $b = 1 - a$). We say that a voting system has a *combinatorial moderating effect* if for all k, the presence of a candidate in I prevents any candidate outside of I from winning; i.e., a moderate candidate (inside I) is guaranteed to win as long as at least one moderate runs. (Note that a combinatorial moderating effect implies a probabilistic one, as long as the candidate distribution F places positive probability mass on I.) We call such an interval I an *exclusion zone* of the voting system, since the presence of a candidate inside this zone precludes outside candidates from winning. In this way, a voting system with a moderating effect will tend to suppress extreme candidates who lie outside a middle portion of the unit interval, while a voting system that does not have at moderating effect will allow arbitrarily extreme candidates to win with positive probability even as the number of candidates becomes large.

Using this terminology, we can state our first main result succinctly: under a uniform voter distribution, IRV has a moderating effect and plurality does not—in both the combinatorial and probabilistic senses. In particular, we prove a novel and striking fact about IRV: when voters and candidates both come from the uniform distribution on $[0, 1]$, the probability that the winning candidate produced by IRV lies outside the interval $[1/6, 5/6]$ goes to 0 as the number of candidates k goes to infinity. In sharp contrast, the distribution of the plurality winner's position converges to uniform as the number of candidates goes to infinity, allowing arbitrarily extreme candidates to win. As part of our analysis, we provide a method for deriving the distribution of plurality and IRV winner positions for finite k and perform this derivation for $k = 3$ candidates. Surprisingly, our analysis of plurality—the simpler voting system—requires much more sophisticated machinery: we establish a connection between plurality voting and a classic model in discrete probability known as the stick-breaking process and develop new asymptotic stick-breaking results for use in our analysis.

Our probabilistic result for IRV follows from a companion fact that is combinatorial in nature and comparably succinct: given any finite set of candidates in $[0, 1]$, and voters from the uniform distribution, if any of the candidates belong to the interval $[1/6, 5/6]$, then the IRV winner must come from $[1/6, 5/6]$; that is, $[1/6, 5/6]$ is an exclusion zone for IRV in the uniform case. Moreover, $[1/6, 5/6]$ is the smallest interval for which this statement is true. Again, the analogue for plurality voting with any proper sub-interval of the unit interval is false: we show that plurality has no exclusion zones.

This first main result therefore gives a precise sense in which the structure of the IRV voting system favors moderate candidates: whenever moderate candidates (in the middle two-thirds of the unit interval) are present as options, IRV will push out more extreme candidates. We then address the more challenging case of non-uniform voter distributions, where we prove that IRV continues to have a moderating effect (in the sense of our formal definitions) even for voter distributions that push probability mass out toward the extremes of the unit interval, up to a specific threshold beyond which the effects cease to hold. Thus, IRV is even able to offset a level of polarization built into the underlying distri-

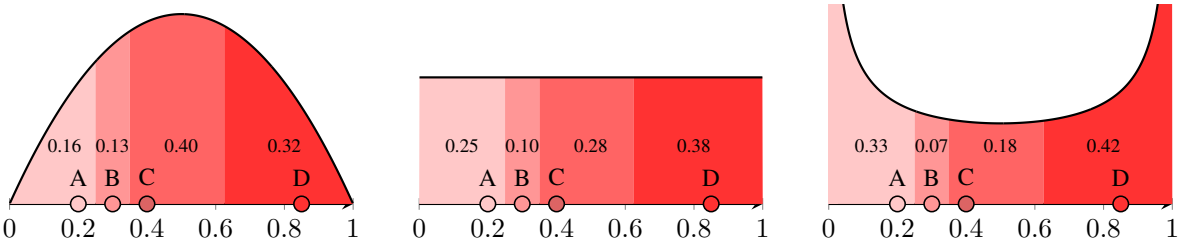

Figure 1: Three example voter distributions in one dimension. Candidates A, B, C, D are placed at 0.2, 0.3, 0.4, and 0.85. The black line shows the density function of the voter distribution. Regions are colored according to the most preferred candidate of voters in that region and annotated with the approximate vote share of that candidate. As an example, the preference ordering of a voter at 0.5 is C, B, A, D. Similarly, a voter at 0.1 has preference ordering A, B, C, D. In the moderate voters example (left), C is both the plurality and IRV winner. In the uniform voters example (center), D is the plurality winner and C is the IRV winner. In the polarized voters example (right), D is the plurality winner and A is the IRV winner.

bution of voters and candidates, although it can only do so up until a certain level of polarization is reached. In contrast, we establish that plurality never has a combinatorial moderating effect for any non-pathological voter distribution.

As a final point, it is worth emphasizing what is and is not a focus of our work here. We examine IRV and plurality because of their widespread use in real-world elections and the fierce debate surrounding the adoption of IRV over plurality. We are not trying to characterize all possible voting systems that give rise to moderation (although we can show that many voting systems not in widespread use have a moderating effect, including the Coombs rule and any Condorcet method; for these systems, any symmetric interval around 0.5 is an exclusion zone). Our interest, instead, is in the following contribution to the plurality–IRV debate: there is a precise mathematical sense in which IRV has a moderating effect and plurality does not. Second, we do not analyze strategic choices by candidates about where to position themselves on the unit interval (Hotelling 1929; Downs 1957; Osborne 1995), but instead derive properties of voting systems that hold for fixed candidate positions, or candidate positions drawn from a distribution. This approach produces results that are robust against the question of whether candidates are actually able to make optimal strategic positioning decisions in practice (Bendor et al. 2011); it also allows us to better understand how the voting systems themselves behave—providing a foundation for future strategic work.

Uniform Voters

The previous section describes our complete model, but it is useful to review it here in the context of some more specific notation. We assume voters and candidates are both drawn from a distribution F on the unit interval $[0, 1]$, representing their ideological position on a left–right spectrum.[2] Voters prefer candidates closer to them (i.e., they have 1-Euclidean preferences). There are k candidates drawn independently from F; suppose that these draws produce candidate positions $x_1 < x_2 < \cdots < x_k$ in order. Some of our results apply regardless of the candidate distribution, relying only on the voter distribution; we will make a note of such cases.

Since we want to model the case of a large population of voters, we do not explicitly sample the voters from F, but instead think of a continuum of voters who correspond to the distribution F itself: that is, under the plurality voting rule, the fraction of voters who vote for candidate x_i is the probability mass of all voters who are closer to x_i than to any other candidate (or, equivalently, it is the probability that a voter randomly chosen according to F would be closer to x_i than any other candidate). In this section, we focus on the case where F is uniform.[3] We use $v(x_i)$ to denote the vote share for candidate x_i. Under IRV, the candidate i with the smallest $v(x_i)$ is eliminated and vote shares are recomputed without candidate i. This repeats until only one candidate remains, who is declared the winner (equivalently, elimination can terminate when a candidate achieves majority). In practice, voters submit a ranking over the candidates and their votes are "instantly" redistributed after each elimination.

IRV's moderating effect: A first result. With uniform 1-Euclidean voters, we now show that IRV cannot elect extreme candidates over moderates—regardless of the distribution of candidates. That is, IRV exhibits an exclusion zone in the middle of the unit interval, where the presence of moderate candidates inside the zone precludes outside extreme candidates from winning. The idea behind the proof is that as moderates get eliminated, the middle part of the interval becomes sparser, granting a higher vote share to any remaining moderates. Consider the moment when only one candidate x remains in the interval $[1/6, 5/6]$ (see Figure 2); extreme candidates near 0 and 1 are then too far away to "squeeze out" x. With uniform voters, the tipping point for squeezing out moderates occurs when extreme candidates are at

[2] We will generally focus on distributions F that are symmetric around $1/2$ and represented by a density function f.

[3] To provide another perspective on the uniform voter assumption, consider the following preference assumption that also produces uniform 1-Euclidean preferences: voters are arbitrarily distributed, but rank candidates according to how many voters are between them and each candidate. That is, voters have 1-Euclidean preferences *in the voter quantile space* and are always uniformly distributed over this space by definition. All of our uniform voter results hold in that setting as well, although stated in terms of voter quantiles rather than absolute positions.

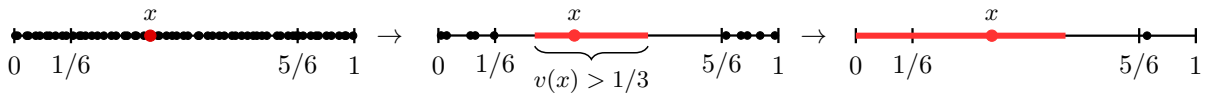

Figure 2: Visual depiction of the proof of Theorem 1. IRV eliminates candidates until a final candidate x remains in the exclusion zone $[1/6, 5/6]$. At this point, x gets more than $1/3$ of the vote share and cannot be eliminated next (regardless of where they are in $[1/6, 5/6]$). Candidates outside of $[1/6, 5/6]$ are thus eliminated until x wins.

1/6 and 5/6. In the next section, we present generalizations of this result for non-uniform voter distributions. All proofs can be found in the extended version of the paper (Tomlinson, Ugander, and Kleinberg 2023b).

Theorem 1. *(Combinatorial moderation for uniform IRV.) Under IRV with uniform voters over $[0, 1]$ and $k \geq 3$ candidates, if there is a candidate in $[1/6, 5/6]$, then the IRV winner is in $[1/6, 5/6]$. No smaller interval $[c, 1-c]$, $c > 1/6$, has this property. If there are no candidates in $[1/6, 5/6]$, then the IRV winner is the one closest to $1/2$.*

In the language of our analysis, $[1/6, 5/6]$ is then the smallest possible exclusion zone of IRV under a uniform voter distribution. See Figure 2 for a visual depiction of the argument. A corollary of Theorem 1 is that if candidates are distributed uniformly at random (for instance, if voters independently and identically decide whether to run for office), then IRV elects extreme candidates with probability going to 0 as the number of candidates grows, since the probability of having no moderate candidates in $[1/6, 5/6]$ is $(1/3)^k$. In the language defined earlier, IRV thus has a probabilistic moderating effect with uniform voters and candidates.

Corollary 1. *(Probabilistic moderation for uniform IRV.) Let R_k be the position of the IRV winner with k candidates distributed uniformly at random and uniform voters.*

$$\lim_{k\to\infty} \Pr(R_k \notin [1/6, 5/6]) = 0. \tag{1}$$

In contrast to IRV, where the presence of candidates with moderate positions (namely, inside $[1/6, 5/6]$) precludes extreme candidates from winning, we now show that no such fact is true for plurality (excluding the extreme points 0 and 1): for any interval $I \subseteq (0, 1)$, there is some configuration of candidates such that the winner is outside of I despite having candidates in I. In other words, plurality voting does not have a combinatorial moderating effect with uniform voters.[4] Later, we generalize this result to non-uniform voter distributions. The idea behind the proof is relatively straightforward: given a set of candidates, keep adding candidates to reduce the vote share of everyone except the desired winner.

Theorem 2. *(No combinatorial moderation for uniform plurality.) Suppose voters are uniformly distributed over $[0, 1]$. Given any set of $\kappa \geq 1$ distinct candidate positions $x_1, \ldots, x_\kappa$ with $x_1 \notin \{0, 1\}$, there exists a configuration of $k \geq \kappa$ candidates (including $x_1, \ldots, x_\kappa$) such that the candidate at x_1 wins under plurality.*

[4] An anonymous reviewer suggested an elegant construction proving this fact for symmetric intervals $I = [c, 1-c]$, which provides counterexamples for every $k \geq 3$: place candidates at $c - \epsilon, 1 - c - \epsilon$, and any others at $1 - c + \epsilon$ (for $\epsilon < c/2$). The candidate at $c - \epsilon$ wins, despite having a candidate in I.

In addition, we prove that the asymptotic distribution of the plurality winner's position is uniform over the unit interval when voters and candidates are positioned uniformly at random. In other words, plurality does not have a probabilistic moderating effect: it does not preclude extreme candidates from winning when there are many moderate candidates to choose from. Note that this result implies plurality also has no combinatorial moderation, but Theorem 2 is considerably easier to prove.

Theorem 3. *(No probabilistic moderation for uniform plurality.) Let P_k be the position of the plurality winner with k candidates distributed uniformly at random and uniform voters. As $k \to \infty$, P_k converges in distribution to Uniform$(0, 1)$; i.e., $\lim_{k\to\infty} \Pr(P_k \leq x) = x$ for $x \in [0, 1]$.*

The proof uses a coupling argument between plurality on the unit interval and plurality on a circle. By rotational symmetry, the plurality winner on a circle is uniformly distributed. We show that as k grows, cutting the circle to transform it into the interval does not change the winner with probability approaching 1, since cutting the circle only affects vote shares of the boundary candidates.

Thus, a key step is deriving the asymptotic distribution of the winning plurality vote share. This vote share distribution may be useful for other asymptotic analyses of plurality voting, so we describe it here. The winning plurality vote share is closely related to a category of probabilistic problems known as *stick-breaking problems*, which focus on the properties of a stick of length 1 broken into n pieces uniformly at random (Holst 1980). Setting $n = k + 1$, these stick pieces can be viewed as the gaps between candidates (equivalently, candidates are the breakpoints of the stick). A classic result in stick-breaking is that the biggest piece will have size B_n almost exactly $\log n/n$ as n grows large (Darling 1953; Holst 1980) and that $nB_n - \log n$ converges to a Gumbel$(1, 0)$ distribution as $n \to \infty$. The plurality vote setting is different, since candidates get vote shares from half of the gap to their left plus half of the gap to their right (except the left- and rightmost candidates). We show that as the number of candidates grows large, the winning vote share V_k with $k = n - 1$ candidates is almost exactly $(\log n + \log\log n)/2n$ and that $nV_k - (\log n + \log\log n)/2$ also converges to Gumbel$(1, 0)$ as $k \to \infty$. Intuitively, the largest pair of adjacent gaps have size $\log n/n$ and $\log\log n/n$, and the candidate between these gaps gets vote shares from half of each gap (more correctly, the total size of the gaps is $(\log n + \log\log n)/n$). This is formalized in the following lemma used to prove Theorem 3.

Lemma 1. *Let V_k be the winning plurality vote share with k candidates distributed uniformly at random over $[0, 1]$ and*

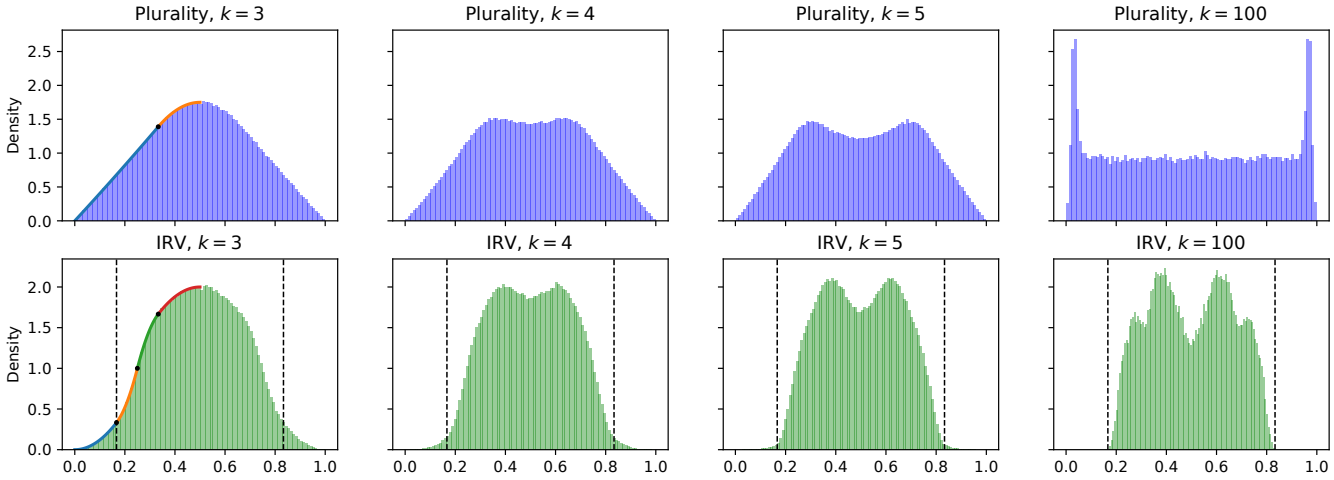

Figure 3: The distributions of the winning position with $k = 3, 4, 5$, and 100 candidates and continuous 1-Euclidean voters (both uniformly distributed) under plurality and IRV. The histograms are from 1 million simulation trials for $k = 3, 4, 5$ and 100,000 trials for $k = 100$, while the curves plotted for $k = 3$ (shown up to 1/2) are the exact density functions given in Propositions 1 and 2, with pieces separated by color. Note that the IRV winner is only at a position $< 1/6$ or $> 5/6$ when no candidates fall in $[1/6, 5/6]$ by Theorem 1; the dashed vertical lines outline this exclusion zone.

uniform voters. Setting $n = k + 1$,

$$\lim_{k\to\infty} \Pr\left(V_k \leq \frac{\log n + \log\log n + x}{2n}\right) = e^{-e^{-x}}. \quad (2)$$

Plurality and IRV Winner Distributions

Given these results about the asymptotic distributions of the plurality and IRV winner positions P_k and R_k, asymptotic in the number of candidates k, a natural follow-on question is whether we can say anything about these distributions for fixed values of k. Indeed, we show that case analysis can in principle be used to derive the exact density functions of P_k and R_k and we perform the derivations for $k = 3$. However, the number of cases grows exponentially in k for plurality and super-exponentially for IRV. In the analysis, we consider a configuration of candidates as a point in k-dimensional space and integrate over the region where a particular candidate wins, which we show is a union of convex polytopes.

Proposition 1.

$$f_{P_3}(x) = \begin{cases} x^2/2 + 4x, & x \in [0, 1/3] \\ -13x^2 + 13x - 3/2, & x \in (1/3, 1/2] \\ f_{P_3}(1 - x), & x \in (1/2, 1]. \end{cases} \quad (3)$$

Proposition 2.

$$f_{R_3}(x) = \begin{cases} 12x^2, & x \in [0, 1/6] \\ 48x^2 - 12x + 1, & x \in (1/6, 1/4] \\ -48x^2 + 36x - 5, & x \in (1/4, 1/3] \\ -12x^2 + 12x - 1, & x \in (1/3, 1/2] \\ f_{R_3}(1 - x), & x \in (1/2, 1]. \end{cases} \quad (4)$$

See Figure 3 for a visualization of P_3 and R_3. Details of the derivations can be found in the extended version (Tomlinson, Ugander, and Kleinberg 2023b). In Proposition 2, the integral of the density $f_{R_3}(x)$ on $[0, 1/6]$ is exactly equal to half the probability that the $k - 1$ losing candidates did not appear inside $[x, 1-x]$ (scaled by k to account for relabeling symmetry), since we know by Theorem 1 that a candidate can only win outside $[1/6, 5/6]$ if they are the most moderate candidate. For general $k > 3$ we can derive the density on $[0, 1/6]$ and $[5/6, 1]$ by generalizing this argument: $f_{R_k}(x) = k(2x)^{k-1}$ on $[0, 1/6]$ (with the right tail being mirrored). Note that the integral of $f_{R_k}(x)$ over $[0, 1/6]$ goes to 0 as $k \to \infty$, a limit that furnishes another way of establishing a probabilistic moderating effect for IRV.

Connecting our results to related work, while the distribution of the winner's position is challenging to derive, the expected plurality vote share at each point is more tractable. This distribution was discovered in another context: a guessing game where the goal is to be closest to an unknown target distributed uniformly at random, against k players who guess uniformly at random (Drinen, Kennedy, and Priestley 2009). The target can be thought of as a random voter and the guesses as candidate positions. The guessing game and plurality winner position distributions are similar in shape, with two prominent bumps that move outward as k grows; and both converge to uniform distributions. However, the point with the max *expected plurality vote share* (and max guessing game win probability) is not quite the same as the point with the maximum *plurality win probability*, since a candidate's position influences other candidates' vote shares.

Non-Uniform Voters

Given our understanding of the uniform voter case, we now broaden our scope and show that IRV exhibits exclusion zones more generally. We find that the same "squeezing" argument can be applied to any symmetric voter distribution. The generalized result hinges on a specific condition

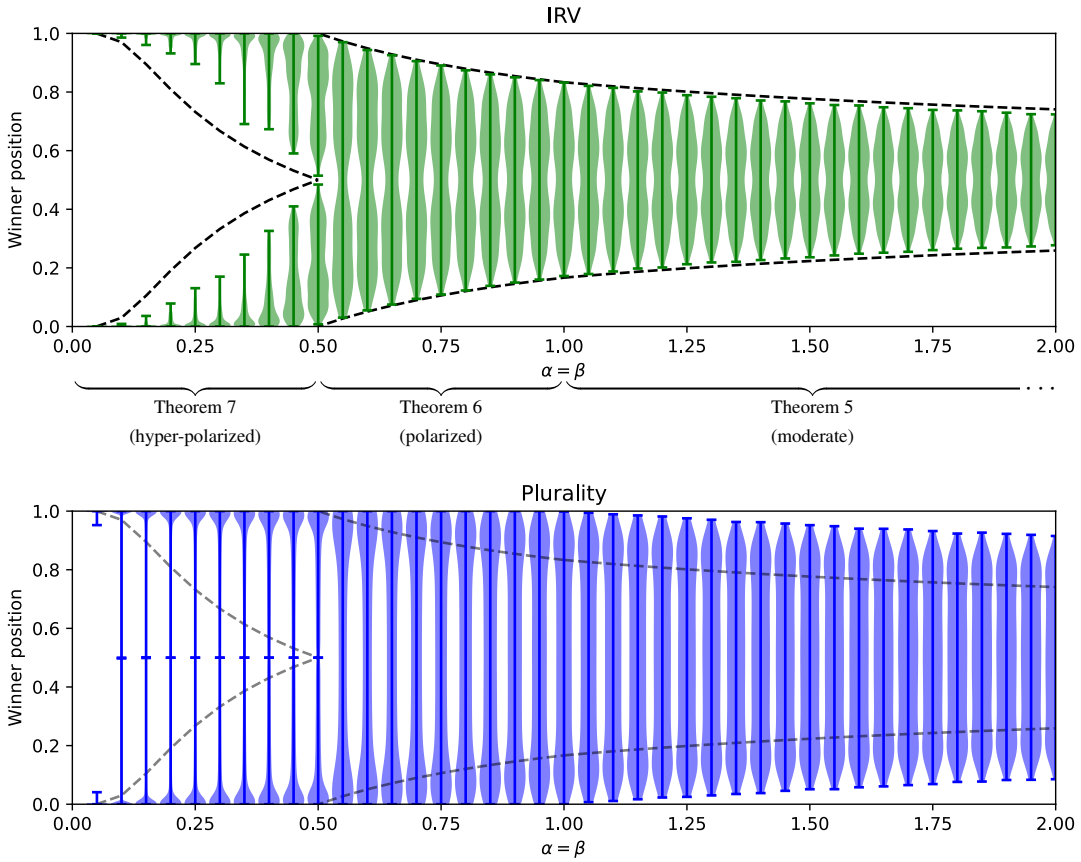

Figure 4: IRV (top) and plurality (bottom) winner positions with Beta(α, α)-distributed voters and candidates. The violin plots show empirical distributions from 100,000 simulation trials with $k = 30$ candidates at each α value, with whiskers marking extrema. The dashed lines show the bounds from Theorems 5 to 7 in the annotated ranges. The red PDFs below show the shape of the voter distribution in each range. As long as voters are not too polarized, IRV prevents extreme candidates from winning. Plurality, on the other hand, allows arbitrarily extreme candidates to win for $\alpha = 1$, when the voter distribution is uniform.

on the cumulative distribution function, Equation (5), which intuitively captures when, no matter where the last moderate candidate is, they cannot be squeezed out by the most moderate extremists. This condition is not always possible to satisfy non-trivially. After first giving the general statement, we present special cases where the condition is simple to state and satisfy—specifically, when the voter density is monotonic over $[0, 1/2]$. If the voter distribution is sufficiently highly polarized, the condition becomes impossible to satisfy. In this *hyper-polarized* regime, the exclusion zone of IRV actually flips, and IRV cannot elect moderate candidates over extreme ones. First, we present the general moderating effect of IRV for symmetric voter distributions.

Theorem 4. *(General combinatorial moderation for IRV.) Let f be symmetric over $[0, 1]$ with CDF F and let $c \in (0, 1/2)$. If for all $x \in [c, 1/2]$,*

$$F\left(\frac{x+1-c}{2}\right) - F\left(\frac{c+x}{2}\right) > 1/3, \qquad (5)$$

then if there is at least one candidate in $[c, 1-c]$, the IRV winner must be in $[c, 1-c]$.

We now consider two cases where Condition (5) can be greatly simplified: when the voter distribution is moderate (f increases over $[0, 1/2]$; Theorem 5) and when voters are polarized (f decreases over $[0, 1/2]$ but $F(1/4) < 1/3$; Theorem 6). The proofs in these cases follow the same structure, but differ in where moderate candidates are easiest to squeeze out (nearer or farther from 1/2). As another note, just as with Corollary 1, we immediately see from Theorem 4 (and the special cases below) that IRV has a probabilistic moderating effect with symmetric voter and candidate distributions (as long as they place positive mass on $[c, 1-c]$): as the number of candidates goes to infinity, the probability that the winner comes from $[c, 1-c]$ goes to 1.

Theorem 5. *(Moderate voter distribution.) Let f be symmetric over $[0, 1]$ and non-decreasing over $[0, 1/2]$. For any $c \leq F^{-1}(1/6)$, if there is a candidate in $[c, 1-c]$, then the IRV winner is in $[c, 1-c]$.*

Theorem 6. *(Polarized voter distribution.) Let f be symmetric over $[0, 1]$, non-increasing over $[0, 1/2]$, and let $F(1/4) < 1/3$. For any $c \leq 2(F^{-1}(1/3) - 1/4)$, if there is a candidate in $[c, 1-c]$, then the IRV winner is in $[c, 1-c]$.*

The uniform distribution is the unique distribution whose density is both non-increasing and non-decreasing over $[0, 1/2]$. Indeed, for uniform $F(x) = x$, $1/6 = 2(F^{-1}(1/3) - 1/4) = F^{-1}(1/6)$. Note that for polarized voter distributions, Theorem 6 requires $F(1/4) < 1/3$ (i.e., less than 1/3 of voters are left of 1/4). If the population is hyper-polarized ($F(1/4) > 1/3$), we can prove that IRV cannot elect moderates if both extremes are represented.

Theorem 7. *(Hyper-polarized voter distribution.) Let f be symmetric over $[0, 1]$ and let $F(1/4) > 1/3$. For any $c \geq 2F^{-1}(1/3)$, if there is at least one candidate in $[0, c]$ and at least one candidate in $[1 - c, 1]$, then the IRV winner must be in $[0, c]$ or $[1 - c, 1]$.*

Finally, we saw in Theorem 2 that plurality has no exclusion zones for uniform voters. We now show that plurality has no exclusion zones regardless of the voter distribution (given mild conditions), except the points 0 and 1.

Theorem 8. *(No combinatorial moderation for plurality.) Let f be continuous and strictly positive over $(0, 1)$. Given any set of $\kappa \geq 1$ distinct candidate positions $x_1, \ldots, x_\kappa$ with $x_1 \notin \{0, 1\}$, there exists a configuration of $k \geq \kappa$ candidates (including $x_1, \ldots, x_\kappa$) such that the candidate at x_1 wins under plurality. If $x_1 \in \{0, 1\}$, then there exist voter distributions f where x_1 cannot win under plurality.*

Figure 4 illustrates Theorems 5 to 8, showing empirical IRV and plurality winner positions when voters (and $k = 30$ candidates) are distributed according to symmetric Beta(α, α) distributions. This family of Beta distributions is polarized for $\alpha < 1$, uniform for $\alpha = 1$, and moderate for $\alpha > 1$. Theorem 5 thus applies for $\alpha \geq 1$. The crossover point between Theorems 6 and 7 (polarized to hyper-polarized) occurs at $\alpha = 1/2$ (i.e., for Beta(1/2, 1/2), $F^{-1}(1/3) = 1/4$). Figure 4 also shows the positions of plurality winners for these voter distributions, consistent with our analysis of plurality in Theorem 8. Our code is available at https://github.com/tomlinsonk/irv-moderation.

Discussion

We began by considering a contrast between IRV and plurality voting when the positions of voters and candidates are drawn from the uniform distribution on the unit interval: in this case, IRV (unlike plurality) has a moderating effect, with the probability that the winner comes from the interval $[1/6, 5/6]$ converging to 1 as the number of candidates goes to infinity. This moderating effect persists (with different sub-intervals) even as the distribution of voters and candidates becomes more polarized, with an increasing amount of probability mass near the endpoints of the interval, until a specific threshold of hyper-polarization is reached. Our analysis also provides methods for determining the exact distribution of winner positions in certain cases, enabling more fine-grained comparisons between IRV and plurality.

It would be interesting to consider extensions of our work in a number of directions, and here we highlight three of these. First, we did not consider strategic analyses (e.g., of Nash equilibria, as in Dellis, Gauthier-Belzile, and Oak (2017)), and were instead motivated by bounded rationality (Bendor et al. 2011) and a need to better understand the underlying voting system, focusing on the non-strategic setting where candidate positions are fixed. For instance, how might candidates behave strategically given an understanding of IRV exclusion zones or the winner position distribution of IRV? Behavioral evidence for bounded rationality indicates that people tend to operate at a low strategic depth (Stahl and Wilson 1995; Colman 2003; Ohtsubo and Rapoport 2006). In this framework, level-0 players act randomly, level-1 players calculate best responses to level-0 players, and so on. Our analysis therefore corresponds to level-0 strategic reasoning, and can be used as a starting point for analysis of higher-order strategy.

Second, we modeled voting populations as symmetric continuous distributions in one dimension, with preferences arising strictly from distances in this dimension. Considering higher-dimensional preference spaces would also be a natural extension of our analysis. Asymmetric voter distributions would also be valuable to consider, although the notion of a *moderate* may need to be revisited in this case (perhaps based on the median voter). Using the same squeezing argument, IRV should also exhibit exclusion zones with asymmetric voter distributions, although their forms may not be as tidy as the ones we derive. Other possible extensions include non-linear voter preferences (for instance, where a voter ranks all candidates on their right before all candidates on their left, regardless of distance), probabilistic voting, and voter abstention. Practical considerations of IRV could also be taken into account; for instance, real-world elections often ask for top-truncated preferences rather than full rankings, which can then affect the outcome (Tomlinson, Ugander, and Kleinberg 2023a). Does IRV with truncated ballots still exhibit a moderating effect?

Finally, as we noted earlier, there are voting systems that always select the most moderate candidate with symmetric 1-Euclidean voters. This is true for any system that satisfies the Condorcet criterion, selecting the Condorcet winner whenever one exists (e.g., the minimax, Condorcet-Hare, Copeland, and Dodgson methods, among many others (Black 1958; Richelson 1975; Green-Armytage, Tideman, and Cosman 2016)); it is also true for some other voting systems that do not in general satisfy the Condorcet criterion, like the Coombs rule (Coombs 1964; Grofman and Feld 2004). There are a variety of practical and historical reasons why these methods are not widely used for political elections. For instance, Dodgson's method is NP-hard to compute (Bartholdi, Tovey, and Trick 1989) and the Coombs rule is sensitive to incomplete ballots, which are common in practice. As we are motivated by ongoing debates about IRV and plurality, our attention has been restricted to these systems. However, a broader understanding of moderating effects would be valuable. There has been some theoretical work on moderating effects of score-based voting systems (like Borda count and approval voting) with strategic voters and candidates (Dellis 2009). However, it is an open question, with some computational evidence to support it (Chamberlin and Cohen 1978), whether other voting systems like Borda count exert a moderating effect in the setting we study, with fixed voter and candidate distributions.

Acknowledgments

This work was supported in part by ARO MURI, a Simons Investigator Award, a Simons Collaboration grant, a grant from the MacArthur Foundation, the Koret Foundation, and NSF CAREER Award #2143176. We thank Robert Kleinberg and Spencer Peters for suggesting the circle-cutting argument used to prove Theorem 3.

References

Aziz, H.; Brill, M.; Conitzer, V.; Elkind, E.; Freeman, R.; and Walsh, T. 2017. Justified representation in approval-based committee voting. *Social Choice and Welfare*, 48(2): 461–485.

Bartholdi, J.; Tovey, C. A.; and Trick, M. A. 1989. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare*, 6: 157–165.

Bendor, J.; Diermeier, D.; Siegel, D. A.; and Ting, M. 2011. A behavioral theory of elections. In *A Behavioral Theory of Elections*. Princeton University Press.

Black, D. 1958. *The theory of committees and elections*. Springer.

Bogomolnaia, A.; and Laslier, J.-F. 2007. Euclidean preferences. *Journal of Mathematical Economics*, 43(2): 87–98.

Bordignon, M.; Nannicini, T.; and Tabellini, G. 2016. Moderating political extremism: single round versus runoff elections under plurality rule. *American Economic Review*, 106(8): 2349–70.

Boutilier, C.; Caragiannis, I.; Haber, S.; Lu, T.; Procaccia, A. D.; and Sheffet, O. 2012. Optimal social choice functions: A utilitarian view. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, 197–214.

Brill, M.; Israel, J.; Micha, E.; and Peters, J. 2022. Individual representation in approval-based committee voting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 4892–4899.

Chamberlin, J. R.; and Cohen, M. D. 1978. Toward applicable social choice theory: A comparison of social choice functions under spatial model assumptions. *American Political Science Review*, 72(4): 1341–1356.

Colman, A. M. 2003. Depth of strategic reasoning in games. *Trends in Cognitive Sciences*, 7(1): 2–4.

Coombs, C. H. 1964. *A theory of data*. Wiley.

Darling, D. A. 1953. On a class of problems related to the random division of an interval. *The Annals of Mathematical Statistics*, 239–253.

Dean, H. 2016. How to Move Beyond the Two-Party System. *The New York Times*. 10/7/16.

DellaPosta, D.; Shi, Y.; and Macy, M. 2015. Why do liberals drink lattes? *American Journal of Sociology*, 120(5): 1473–1511.

Dellis, A. 2009. Would letting people vote for multiple candidates yield policy moderation? *Journal of Economic Theory*, 144(2): 772–801.

Dellis, A.; Gauthier-Belzile, A.; and Oak, M. 2017. Policy Polarization and Strategic Candidacy in Elections under the Alternative-Vote Rule. *Journal of Institutional and Theoretical Economics*, 565–590.

DeMarzo, P. M.; Vayanos, D.; and Zwiebel, J. 2003. Persuasion bias, social influence, and unidimensional opinions. *The Quarterly Journal of Economics*, 118(3): 909–968.

Diamond, L. 2016. The Second-Most Important Vote On Nov. 8. *Foreign Policy*. 10/13/16.

Donovan, T.; Tolbert, C.; and Gracey, K. 2016. Campaign civility under preferential and plurality voting. *Electoral Studies*, 42: 157–163.

Downs, A. 1957. *An economic theory of democracy*. Harper & Row.

Drinen, D.; Kennedy, K. G.; and Priestley, W. M. 2009. An optimization problem with a surprisingly simple solution. *The American Mathematical Monthly*, 116(4): 328–341.

Ebadian, S.; Kahng, A.; Peters, D.; and Shah, N. 2022. Optimized distortion and proportional fairness in voting. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, 563–600.

Elkind, E.; Lackner, M.; and Peters, D. 2022. Preference Restrictions in Computational Social Choice: A Survey. *arXiv preprint: arXiv:2205.09092*.

Fraenkel, J.; and Grofman, B. 2004. A neo-Downsian model of the alternative vote as a mechanism for mitigating ethnic conflict in plural societies. *Public Choice*, 487–506.

Fraenkel, J.; and Grofman, B. 2006a. Does the alternative vote foster moderation in ethnically divided societies? The case of Fiji. *Comparative Political Studies*, 39(5): 623–651.

Fraenkel, J.; and Grofman, B. 2006b. The failure of the alternative vote as a tool for ethnic moderation in Fiji: A rejoinder to Horowitz. *Comparative Political Studies*, 39(5): 663–666.

Gkatzelis, V.; Halpern, D.; and Shah, N. 2020. Resolving the optimal metric distortion conjecture. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, 1427–1438. IEEE.

Green-Armytage, J.; Tideman, T. N.; and Cosman, R. 2016. Statistical evaluation of voting rules. *Social Choice and Welfare*, 46: 183–212.

Grofman, B.; and Feld, S. L. 2004. If you like the alternative vote (aka the instant runoff), then you ought to know about the Coombs rule. *Electoral Studies*, 23(4): 641–659.

Halpern, D.; Kehne, G.; Procaccia, A. D.; Tucker-Foltz, J.; and Wüthrich, M. 2023. Representation with incomplete votes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 5657–5664.

Holst, L. 1980. On the lengths of the pieces of a stick broken at random. *Journal of Applied Probability*, 17(3): 623–634.

Horowitz, D. L. 2006. Strategy takes a holiday: Fraenkel and Grofman on the alternative vote. *Comparative Political Studies*, 39(5): 652–662.

Horowitz, D. L. 2007. Where have all the parties gone? Fraenkel and Grofman on the alternative vote–yet again. *Public Choice*, 133(1): 13–23.

Hotelling, H. 1929. Stability in Competition. *The Economic Journal*, 39(153): 41–57.

John, S.; and Douglas, A. 2017. Candidate civility and voter engagement in seven cities with ranked choice voting. *National Civic Review*, 106(1): 25–29.

Kahng, A.; Latifian, M.; and Shah, N. 2023. Voting with Preference Intensities. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Kropf, M. 2021. Using campaign communications to analyze civility in ranked choice voting elections. *Politics and Governance*, 9(2): 280–292.

Layman, G. C.; Carsey, T. M.; and Horowitz, J. M. 2006. Party polarization in American politics. *Annual Review of Political Science*, 9: 83–110.

McGann, A. J.; Grofman, B.; and Koetzle, W. 2002. Why party leaders are more extreme than their members: Modeling sequential elimination elections in the US House of Representatives. *Public Choice*, 113(3-4): 337–356.

Merrill, S. 1984. A comparison of efficiency of multicandidate electoral systems. *American Journal of Political Science*, 23–48.

Mitchell, P. 2014. The single transferable vote and ethnic conflict: the evidence from Northern Ireland. *Electoral Studies*, 33: 246–257.

Ohtsubo, Y.; and Rapoport, A. 2006. Depth of reasoning in strategic form games. *The Journal of Socio-Economics*, 35(1): 31–47.

Osborne, M. J. 1995. Spatial models of political competition under plurality rule: A survey of some explanations of the number of candidates and the positions they take. *Canadian Journal of Economics*, 261–301.

Poole, K. T.; and Rosenthal, H. 1984. The polarization of American politics. *The Journal of Politics*, 46(4): 1061–1079.

Poole, K. T.; and Rosenthal, H. 1991. Patterns of congressional voting. *American Journal of Political Science*, 228–278.

Reilly, B. 2018. Centripetalism and electoral moderation in established democracies. *Nationalism and Ethnic Politics*, 24(2): 201–221.

Richelson, J. 1975. A comparative analysis of social choice functions. *Behavioral Science*, 20(5): 331–337.

Stahl, D. O.; and Wilson, P. W. 1995. On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1): 218–254.

Tomlinson, K.; Ugander, J.; and Kleinberg, J. 2023a. Ballot Length in Instant Runoff Voting. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*.

Tomlinson, K.; Ugander, J.; and Kleinberg, J. 2023b. The Moderating Effect of Instant Runoff Voting. arXiv:2303.09734.

Wang, S. S.-H.; Cervas, J.; Grofman, B.; and Lipsitz, K. 2021. A systems framework for remedying dysfunction in US democracy. *Proceedings of the National Academy of Sciences*, 118(50): e2102154118.