How Banzhaf Makes a Victor: Curing a Distortion Paradox in Weighted Voting

JAKOB DE RAAIJ, Harvard University, United States
MOON DUCHIN, University of Chicago, United States
ARIEL D. PROCACCIA, Harvard University, United States
JAMIE TUCKER-FOLTZ, Yale University, United States

Political scientist John Banzhaf brought power indices into the legal and political spotlight in the United States by highlighting problems with naive weighting, like the possibility of "dummy players" who are never pivotal in any coalition. He argued that weighted voting should be framed as an optimization problem, with the objective that the probability of a representative casting a decisive vote is proportional to the size of their constituency. Today, there are examples (such as county government in New York State) where representatives vote with weights derived from exactly this heuristic optimization.

Though there is a massive literature on power indices for weighted voting games, real-world instances also suggest a new kind of problem not frequently discussed in research papers: with supermajority voting quotas, achieving close agreement between populations and powers can require weight vectors starkly out of proportion to both. For instance, not only contrived examples but also realistic ones can force a medium-sized player—like the Town of Victor in Ontario County, NY—to receive an "undeserving" veto, which could undermine the public legitimacy of the system.

We introduce a new power index, built by having players vote yes with a probability equal to the voting quota rather than by flipping a fair coin as in ordinary Banzhaf. Among variants where players independently vote with fixed probability, this choice uniquely avoids systematic weight distortions while retaining the policy-aligned interpretation of Banzhaf power.

Contents

Abstract		0
Contents		0
1	Introduction	1
2	Preliminaries	5
3	Propensities and the Adaptive Banzhaf Power Index	7
4	Powers in Large Weighted Voting Games	7
5	Veto Distortion	12
6	Empirical Results	14
7	Discussion	18
References		19
A	Additional Related Work	21
B	Missing Proofs	22
C	Full Empirical Results	31

Manuscript submitted for review to the 27th ACM Conference on Economics & Computation (EC'26).

1 Introduction

1.1 The measurement of voting power and the inverse power problem

The European Economic Community, formed in 1958, consisted of six Western European countries: Belgium, France, (West) Germany, Italy, Luxembourg, and the Netherlands. It was governed by the Council of Ministers, which included one minister from each member country. When a proposal was sent to the Council from the European Commission, it would be voted on using weighted votes, with weights assigned in a way that was commensurate with population—4 for France, Germany and Italy, 2 for Belgium and the Netherlands, and 1 for Luxembourg—and an overall weight quota of 12 required for a resolution to pass [Mayer, 2018]. The effectiveness of this design depends on whether each country has the *power* it deserves based on its population, where power is commonly understood to be based on playing a pivotal role in collective decisions. From this viewpoint, the Council of Ministers is a textbook design failure: Luxembourg is a "dummy player," meaning that their vote can never change an outcome, no matter how the other countries vote.

Having made the observation that the voting power may be very different from the (relative) voting weight, the *inverse power problem* arises: How do we choose voting weights so that the voting powers match a desired distribution as closely as possible?

The ability to align powers with a target hinges on how voting power itself is quantitatively defined. In the literature, this question is addressed using two canonical measures, one due to Banzhaf [1965]—already implicitly defined by Penrose [1946]—and the other due to Shapley and Shubik [1954]. Underlying both measures is the idea that voter i is pivotal in a coalition if the overall weight of the coalition is under the quota without i and meets or exceeds the quota with i. Under Banzhaf's definition, the power of a voter i is the probability that they are pivotal in a uniformly random coalition that includes i. This is equivalent to assuming that every voter except i votes YES with probability $1/2$, and asking if i's YES or NO vote matters. By contrast, under the Shapley-Shubik power index, it is the *size* of the coalition that is selected uniformly at random. Equivalently, a uniformly random *permutation* of the voters is selected; the power of a voter i is the probability that i is pivotal in the coalition that includes all their predecessors.

Both power indices have theoretical justifications; in particular, both lend themselves to reasonable axiomatic characterizations [Dubey, 1975, Dubey and Shapley, 1979]. The most significant theoretical distinction was pointed out by Felsenthal and Machover [1998], based on an argument by Coleman [1968], arguing that the two power indices measure different types of power. If the voting agents are *policy-seeking*, they form an opinion of the bill at hand and vote accordingly. Their *I-power*, power to *influence* the outcome of the election, is the probability with which their opinion will be pivotal. They argue that by the Principle of Insufficient Reason, any split of the other voters into YES and NO should be assumed *a priori* to be equally likely—as is the case in the Banzhaf power index. In contrast, *office-seeking* voting agents do not hold an intrinsic attitude towards the bill; the winning coalition gains a *prize* that will be split between them. Their *P-power* is their bargaining power in forming a winning coalition—as is measured by Shapley value for cooperative games with transferable utilities (and the Shapley-Shubik power index, its restriction to weighted voting). Felsenthal and Machover [1998] argue that based on the nature of the voting body, either approach may be warranted. The question of which power index is supported by empirical evidence (if any) is more contentious—we elaborate on it in Appendix A.

Our goal is not to settle (or even advance) the Banzhaf vs. Shapley-Shubik debate. Rather, motivated by a case study from New York State, we focus on the Banzhaf power index. We draw attention to a practical shortcoming of the ordinary (i.e., standard) Banzhaf power index that deserves to be called a "paradox": Like the classic "Alabama paradox" of apportionment rules, it produces results that would make many observers cry foul. We introduce a novel variant that

we call *adaptive Banzhaf* that alleviates these issues while preserving the advantages of ordinary Banzhaf over Shapley-Shubik. *If* one is persuaded that the policy-seeking voter behavior justifies the choice of the Banzhaf power index, then our variant is a strict improvement, especially in the setting of the voting quota being fixed and the weights being optimized — the inverse power problem.

1.2 A case study with real-world significance

New York State contains 62 counties, whose local governing structures vary. Some counties elect representatives using equal-population districts, whereas others employ a structure called a *Board of Supervisors*, with one member from each constituent geographical piece, although the populations may not be equal. Ontario County, located just southeast of Rochester, NY, is one of the latter category. Its 21–member Board consists of one member from each of the 16 smaller towns, while the Cities of Canandaigua and Geneva are subdivided by ward into two and three pieces, respectively (see Table 1). From now on, we will use "town" to refer to the geographic area that receives a single representative.[1]

Each town has a different population, yet is represented by a single elected supervisor. To comply with the constitutional principle of "One Person, One Vote," Boards of Supervisors are required to use weighted voting. Today, the legal requirements for implementing weighted voting are surprisingly complex, thanks to a 1967 appeals court decision in *Iannucci vs. Board of Supervisors* [New York Court of Appeals, 1967]. Recognizing the pitfalls of weighted voting—and citing Banzhaf's seminal work specifically—the court ruled:

> "The principle of one man-one vote is violated, however, when the power of a representative to affect the passage of legislation by his vote… does not roughly correspond to the proportion of the population in his constituency."

In other words, voting power, rather than weight, must be calibrated to population. This standard explicitly calls for solving the inverse power problem. On that subject, the court further elaborated:

> "… [measuring power] is impossible without computer analyses, and, accordingly, if the boards choose to reapportion themselves by the use of weighted voting, there is no alternative but to require them to come forward with such analyses and demonstrate the validity of their reapportionment plans."

From then until today, every county using towns as districts does exactly that. In their charter or in local law, they specify language like the following: "The voting power of a supervisor shall be measured by the mathematical possibility of his casting a decisive vote on a particular matter... In preparing each reapportionment, the board of supervisors shall employ an independent computerized mathematical analysis and such other method or methods as shall most nearly equalize the percentage of voting power of each town and city to its percentage of the total county population."[2]

Indeed, we became aware of this state of affairs when one of the authors of this paper was commissioned by Ontario County to generate four sets of voting weights: one to be used for votes requiring a simple majority, and three others for votes requiring supermajority thresholds of $2/3$, $3/5$, and $3/4$. Here, as in many real-world cases, the quota is set by law or constitution to regulate how easy or difficult it is for a voting body to initiate action and pass legislation. Figure 1 shows optimized results for the $3/4$ threshold, which are so bizarre that they rise to the level of a paradox.

[1] New York is one of 12 "township states" that is tiled by its cities, towns, and townships—what the Census Bureau refers to as minor civil divisions, or MCDs—and those have active local governments. By contrast, many other states have much more limited municipal coverage. In New York it is therefore possible for townships to serve as districts for county government. This has seemed appealing because they are well known to residents and fundamentally hard to gerrymander.

[2] This language is drawn from the Nassau County Charter, quoted in law.justia.com.

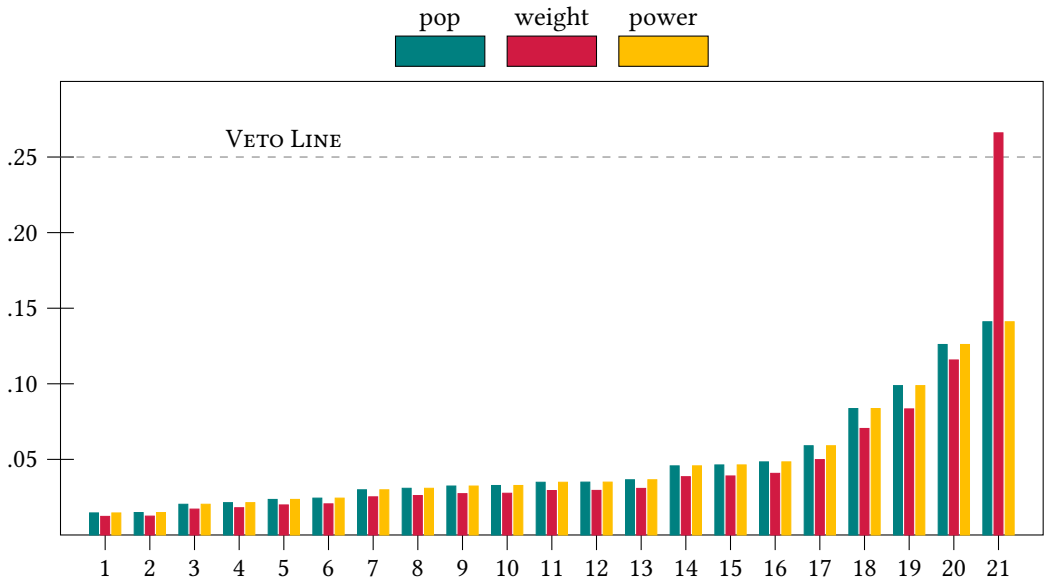


Fig. 1. Populations, (heuristically) optimal weights, and resulting Banzhaf power indices for each of the 21 towns in Ontario County, for a voting quota of 3/4. These weights were found by a randomized local-search algorithm described in Section 6.3.

The Town of Victor, despite only having 14.1% of the overall population—not much larger than the next largest town at 12.6% of the population—was apportioned 26.5% of the weight. This might give any reasonable observer pause, especially because at threshold 3/4, a player with weight over 25% holds the power to unilaterally *veto* a measure: without their support, it cannot pass. From Banzhaf's perspective, this is not a problem, as the plot shows power shares align almost perfectly with population shares—a veto does not in itself contribute to the measurement of power. That a veto player may not be especially powerful is perhaps surprising, and in any event this distribution of weights will likely seem intuitively unfair.

We will use the term *weight distortion* for situations like this in which there is low discrepancy between population and power, but this is achieved with a weight vector far out of proportion to power. A particular kind of weight distortion is *veto distortion*, in which a player's voting weight crosses the veto line while the power lags behind.

1.3 Our proposal: Adaptive Banzhaf

We view the Victor paradox in Ontario County as a symptom of a more fundamental issue with the Banzhaf power index at high voting quotas. By effectively assuming that every voter is equally likely to vote YES or NO, the power computation conditions on the rare event that enough YES votes are cast to reach a threshold much higher than 1/2.

We propose a simple alternative model of voting behavior under which this event is not rare: For a matter requiring a supermajority quota of $q > 1/2$, we suppose each voter has a *propensity* to vote YES with probability $p = q$, rather than 1/2. The *adaptive Banzhaf power index* follows Banzhaf's construction exactly, but uses this alternative model of random, independent votes. (There will be many probabilities discussed here, so we will reserve the word "propensity" for the probability of a positive vote.)

Conceptually, we feel that our variant index has multiple advantages. To see why, assume there are many players in a weighted voting game, all voting with equal propensity p. If each votes YES independently with propensity $\frac{1}{2}$, then the distribution of positive weight is tightly concentrated around half. But pivotality occurs when positive weight is near the quota q, so for high quotas, each voter has a negligible chance of being pivotal. The same is true for any propensity p far from q. The setting $p = q$ is the only choice for which elections will likely be on the knife-edge. This was recognized already in Banzhaf's original paper, where he writes that "it would seem that the test of a legislator's power comes only when the other representatives are closely divided and the individual legislator is able to cast a deciding number of votes" [Banzhaf, 1965].

Thus, the adaptive Banzhaf power index hypothesizes a model of voting behavior that is built to emphasize this test of power. In addition, we will provide rigorous proofs that adaptive Banzhaf is the unique form of propensity-based voting that asymptotically eliminates the Victor paradox, along with empirical evidence that it reduces weight distortion in real cases.

1.4 Results and outline

We formally define weighted voting games and power indices following standard conventions in Section 2. We introduce the class of p-propensity Banzhaf power indices, in which every voter is assumed to vote YES independently with probability p, in Section 3. We define the adaptive Banzhaf power index as the p-propensity case where propensity equals voting quota ($p = q$).

In Section 4, we investigate p-propensity Banzhaf powers in various limiting regimes. First, in Section 4.1, we consider a large family of weighted voting games $\mathcal{G}_n$, with the number of players going to infinity and each individual player's weight going to 0. Under mild assumptions, Theorem 4.2 gives a complete solution for the limiting ratio of powers of any two players, in terms of p and q. Computing this ratio in the special case $p = 1/2$ (the ordinary Banzhaf power index) has been an open problem since it was posed in Lindner and Machover [2004]. When $p = q$, we find that the ratio of the powers of any two players tends to the ratio of their weights. We prove that adaptive Banzhaf is the only member of a large class of power indices including the Banzhaf power index (*semivalues* with independent voters) for which this weight-to-power proportionality holds. This is a desirable property of a power index known as the *Penrose Limit Theorem*.

Section 4.2 studies a distinct class of infinite games known as *oceanic games*, where some large players hold a fixed share of the weight each, while the remaining weight is split evenly among a growing "ocean" of small players. We give a precise calculation of the players' powers in the limit in Theorem 4.5. This gives another setting in which the $p = q$ case is special: it is the only choice of propensity for which the power of the large players is bounded away from zero (as long as the quota avoids a finite list of "pitfall points").

In Section 5, we give some negative results on the *veto distortion* of any propensity Banzhaf power indices. Theorem 5.1 states that for any supermajority quota q and any propensity p, there exist examples where a voter of arbitrarily small population share has a veto under an exactly optimal set of weights. However, for $p \approx q$, these examples rely on delicate constructions that are unlikely to arise in real-world instances. We conjecture that weight-to-power proportionality is generic for adaptive Banzhaf power.

We close with empirical results in Section 6, for both synthetic and real-world examples. The predictions made by the theory can be observed with satisfying clarity in the experiments. In particular, at supermajority quotas, the adaptive Banzhaf index leads to significantly more proportional weights than the ordinary Banzhaf power index, resolving the paradox of undeserving veto players that we find to be widespread in real instances.

1.5 Related Work

Below we survey the work most closely related to the current project; additional literature review is found in Appendix A.

Power indices in the limit. Shapiro and Shapley [1978] first considered power indices in the limit, for the Shapley-Shubik index of games with some players of fixed weight and an infinite number of players with negligible weight. Dubey and Shapley [1979] extend their results to the Banzhaf power index. Neyman [1982] studies the Shapley-Shubik index when the weights of all players go to 0. Lindner and Machover [2004], based on unproven claims due to Penrose [1946], define the Penrose Limit Theorem (PLT) as the property that the ratio of two players' powers approaches the ratio of their weights, given that all weights are bounded by a constant. They find that under mild assumptions, this holds for the Shapley-Shubik index, and for the Banzhaf index at quota $1/2$. Chang et al. [2006] empirically, and Lindner and Owen [2007] theoretically, show that the PLT does not hold for the Banzhaf index at quotas other than $1/2$, but leave the question of the actual limit of the ratio of powers open. We resolve this open question.

Distortion for optimized weights. The weight distortion phenomenon seen in Ontario County has been observed elsewhere: Leech [2002c] notes that in weighted voting among the executive directors of the International Monetary Fund at the supermajority quota $q = .85$, the United States has only a 6.5% share of the Banzhaf power, despite holding 17.5% of the voting weight. He estimates that the weight share of the US must be set to 67.5% so that their Banzhaf power hits the target of 17.5%. Leech and Machover [2003] use heuristics to argue that in the Council of Ministers of the European Union at supermajority quotas, Germany's weight would need to be set disproportionally high, converging to 100% as the quota goes to 1, to achieve proportional Banzhaf powers. (We revisit their motivating example of the International Monetary Fund in Appendix C.5 and verify that adaptive Banzhaf cures the weight distortion.)

Voter propensity. The idea of generalizing the Banzhaf power index with independent voting probabilities originates, to the best of our knowledge, with Owen [1972]. Straffin [1977] considers a setting where the p_i (probability of YES vote from voter i) are random variables drawn independently or dependently from a uniform distribution on $[0, 1]$ and shows that this leads to the Banzhaf and the Shapley-Shubik indices, respectively. Puente del Campo [2000] refers to unnormalized power indices where all players vote YES independently all with the same probability p as binomial semivalues and studies them as a basis of the space of semivalues (which will be defined below). Amer and Giménez [2007] give an axiomatic justification for binomial semivalues based on the delegation of power between players. We have not found a theoretical or practical justification in the computational literature for using any propensity other than $1/2$, which seems to be preferred due to its impartial and entropy-maximizing nature. The idea of adaptive Banzhaf, where propensity equals quota, does not appear to have been studied before.

2 Preliminaries

2.1 Weighted voting games

A *weighted voting game* (WVG) is a tuple $\mathcal{W} = (\boldsymbol{w}; q)$ consisting of the player *weights* $\boldsymbol{w} \in \Delta^n$ (the n-dimensional standard simplex) and a *quota* $q \in [0, 1)$. A *coalition* C is a subset of the voter set $N = [n]$. A coalition C is *winning* if the weight of the players voting YES exceeds the quota,

$\sum_{i \in C} w_i > q$, else it is *losing*.[3] We say that a player i is *pivotal* for a coalition C with $i \notin C$ if C is losing but $C \cup \{i\}$ is winning and let $\mathrm{piv}_{\mathcal{W}}(i, C)$ be the corresponding indicator function.

2.2 Semivalues and power indices

Following Felsenthal and Machover [1998], we use the term power *measure* for any function $\hat{\boldsymbol{\varphi}}$ that assigns an unnormalized power vector $\hat{\boldsymbol{\varphi}}(\mathcal{W}) \in \mathbb{R}^n$ to any WVG $\mathcal{W}$ with n players, for all $n \in \mathbb{N}$. We call the function $\boldsymbol{\varphi}$ a power *index* if the power vectors are normalized, that is $\boldsymbol{\varphi}(\mathcal{W}) \in \Delta^n$ for any WVG $\mathcal{W}$ with n players, for all $n \in \mathbb{N}$

A *semivalue* [Weber, 1979] is a power measure $\hat{\boldsymbol{\varphi}}^{\boldsymbol{p}}(\mathcal{W})$ that assigns the voting power of a player i in the WVG $\mathcal{W}$ as the probability with which i is pivotal for a random coalition $C \subseteq N \backslash \{i\}$. It is assumed that the probability of a coalition arising depends only on the size of the coalition but is independent of the labels of the players in the coalition. In particular, for a WVG with n players, a probability vector $\boldsymbol{p} = (p_0, ..., p_{n-1})$ such that $\sum_{i=0}^{n-1} \binom{n-1}{i} p_i = 1$ assigns a probability of $p_{|C|}$ to each coalition $C \subseteq N \backslash \{i\}$ based on its size. The semivalue of player i in WVG $\mathcal{W}$ is

$$\hat{\varphi}_i^{\boldsymbol{p}}(\mathcal{W}) = \sum_{C \subseteq N \backslash \{i\}} \mathrm{piv}_{\mathcal{W}}(i, C) \cdot p_{|C|}.$$

The corresponding *normalized* semivalue is the power index

$$\varphi_i(\mathcal{W}) = \frac{\hat{\varphi}_i(\mathcal{W})}{\sum_{j \in N} \hat{\varphi}_j(\mathcal{W})}.$$

The *Banzhaf power measure* $\hat{\boldsymbol{\beta}}(\mathcal{W})$ [Banzhaf, 1965, Penrose, 1946] is the semivalue in which every coalition $C \subseteq N \backslash \{i\}$ is assumed to be equally likely. Equivalently, each voter in $N \backslash \{i\}$ is assumed to vote YES independently with probability $1/2$. Consequently, $p_k^n = 1/2^{n-1}$ so that

$$\hat{\beta}_i(\mathcal{W}) = \frac{1}{2^{n-1}} \sum_{C \subseteq N \backslash \{i\}} \mathrm{piv}_{\mathcal{W}}(i, C).$$

The *Banzhaf power index* $\boldsymbol{\beta}(\mathcal{W})$ is the Banzhaf power *measure* normalized to sum to 1, i.e.,

$$\beta_i(\mathcal{W}) = \frac{\hat{\beta}_i(\mathcal{W})}{\sum_{j \in N} \hat{\beta}_j(\mathcal{W})}.$$

Example 2.1. Consider $\mathcal{W} = ((0.25, 0.25, 0.5); 0.6)$. The first player, $i = 1$, is pivotal for one coalition, $\{3\}$, so their Banzhaf power measure is $\hat{\beta}_1(\mathcal{W}) = 1/4$; analogously, $\hat{\beta}_2(\mathcal{W}) = 1/4$. The third player, $i = 3$, is pivotal for three coalitions, $\{1\}$, $\{2\}$ and $\{1, 2\}$, so their Banzhaf power measure is $\hat{\beta}_3(\mathcal{W}) = 3/4$. This gives normalized powers $\boldsymbol{\beta} = (1/5, 1/5, 3/5)$.

2.3 Inverse power problem

In the *inverse power problem*, we are given a *target distribution of power* $\boldsymbol{m} \in \Delta^n$, a quota $q \in [0, 1)$, and a desired power index $\boldsymbol{\varphi}$. For a weight vector $w \in \Delta^n$, we define its discrepancy as

$$\mathrm{discr}_{\boldsymbol{m}, q, \boldsymbol{\varphi}}(\boldsymbol{w}) = \|\boldsymbol{m} - \boldsymbol{\varphi}((\boldsymbol{w}; q))\|_1 .$$

The goal is to find one or all weight vectors

$$\boldsymbol{w}^* \in W^*(\boldsymbol{m}, q, \boldsymbol{\varphi}) = \arg\min_{\boldsymbol{w} \in \Delta^n} \mathrm{discr}_{\boldsymbol{m}, q, \boldsymbol{\varphi}}(\boldsymbol{w})$$

[3]It seems to be more common in the literature to denote a coalition as winning when it meets the quota (and not necessarily exceeds), that is, if $\sum_{i \in C} w_i \geq q$. However, we require the coalition to exceed the quota as this seems to be more common for simple majority voting in practice. We note that our theoretical results and proofs hold true regardless of the strict or weak winning convention, except when noted otherwise.

that minimize discrepancy. In the inverse power problem, the quota is prescribed exogenously.

3 Propensities and the Adaptive Banzhaf Power Index

We study a generalization of the Banzhaf power index that we call *p-propensity Banzhaf index.*

Definition 3.1. The *p-propensity Banzhaf power measure* $\hat{\boldsymbol{\beta}}^p$, also known as *binomial semivalue* with value p, is the semivalue in which each voter in $N\backslash\{i\}$ is assumed to vote YES independently with probability p. Consequently, $p_k^n = p^k(1-p)^{n-k-1}$ so that

$$\hat{\beta}_i^p(\mathcal{W}) = \sum_{C\subseteq N\backslash\{i\}} \mathrm{piv}_{\mathcal{W}}(i,C)\cdot p^{|C|}(1-p)^{n-|C|-1}.$$

The *p-propensity Banzhaf power index* $\boldsymbol{\beta}^p(\mathcal{W})$ is the p-propensity Banzhaf power *measure* normalized to sum to 1

$$\beta_i^p(\mathcal{W}) = \frac{\hat{\beta}_i^p(\mathcal{W})}{\sum_{j\in N}\hat{\beta}_j^p(\mathcal{W})}.$$

The *adaptive Banzhaf power index (measure)* $\hat{\boldsymbol{\rho}}$ is the q-propensity Banzhaf power index (measure), where q is the quota of the WVG $\mathcal{W} = (\boldsymbol{w}, q)$. That is,

$$\rho_i(\mathcal{W}) = \beta_i^q(\mathcal{W}) \quad (\text{and } \hat{\rho}_i(\mathcal{W}) = \hat{\beta}_i^q(\mathcal{W})).$$

Of course, the $1/2$-propensity Banzhaf power index (measure) is simply the Banzhaf power index (measure).

Example 3.2. We return to the WVG $\mathcal{W} = ((0.25, 0.25, 0.5); 0.6)$ from Example 2.1. While the players are still pivotal for the same coalitions, the respective weighting of the coalitions changed; in particular, larger coalitions are assumed to be more likely now. We get that $\hat{\rho}_1(\mathcal{W}) = \hat{\rho}_2(\mathcal{W}) = p(1-p) = q(1-q) = 0.24$ while $\hat{\rho}_3(\mathcal{W}) = 2\cdot q(1-q) + q^2 = 0.84$. This gives normalized powers $\boldsymbol{\rho} = (2/11, 2/11, 7/11)$.

One reason why setting the propensity equal to the quota has not been formally considered so far may be that in this setting, two WVGs with the same set of winning coalitions but different quotas can have different powers. In any finite WVG, the quota q can be slightly increased and/or decreased without changing the set of winning coalitions. For example, one can quickly check that the sets of winning coalitions for WVGs $\mathcal{W} = ((0.25, 0.25, 0.5); 0.6)$ and $\mathcal{W}' = ((0.25, 0.25, 0.5); 0.7)$ are identical. However, $\boldsymbol{\rho}(\mathcal{W}) \neq \boldsymbol{\rho}(\mathcal{W}')$—the players are assigned different powers in the two WVGs, even though the winning coalitions have not changed!

That said, as we will see in subsequent sections, this apparent shortcoming is of no concern in the settings we consider. In the inverse power problem the quota is an exogenous parameter. Furthermore, as the number of players grows, the 'wiggle room' of the quota—and thus that of the powers— goes to 0. Therefore, the adaptive Banzhaf index powers are well-defined in the limit setting.

4 Powers in Large Weighted Voting Games

It is a desirable property of a power index to assign powers proportional to weights in games that are sufficiently 'smooth.' We argue that the fact that the intuitive assumption that voting powers equal voting weights does not hold stems from the discrete nature of the setting: Only a finite number of powers are achievable with coalitions of a finite number of players, so a player's weight cannot in general perfectly correspond to their power. However, this discrete noise should reduce as the number of players grows, given that the weights of all players are of the same order of magnitude. In particular, we posit that in such a setting, the powers of the players should be

proportional to their weights—a deviation from proportionality in a smooth, limiting case may be a symptom of an underlying bias towards larger or smaller weight players by the power index.

In this section, we examine this claim in two different scenarios of weighted voting with an infinite number of players. First, we consider a setting where an infinite number of players with fixed, finite, bounded voting weights are added to a WVG (with weights then normalized to sum to 1). We analytically solve for the p-propensity Banzhaf powers of players in the limit, resolving a long-standing open question by Lindner and Machover [2004] for $p = 1/2$, i.e., normal Banzhaf. We find that the adaptive Banzhaf power index is the unique p-propensity Banzhaf power index that has the desired property of being proportional in the limit.

We then consider a different setting, in which a finite number of players have a fixed amount of weight each, with the remaining weight being split up equally among an infinite number of players. While we cannot expect perfect proportionality in this setting as the discrete noise due to the fixed-sized players persists, we show that the adaptive Banzhaf power index is the unique p-propensity Banzhaf power index that has the property that the powers of the fixed-sized players are always bounded away from 0, except for a finite number of unstable *pitfall* points.

4.1 The Penrose Limit Theorem

Suppose we fix an infinite sequence $\boldsymbol{w} = w_1, w_2, \ldots$ and a quota $q \in [0, 1)$, where the weights w_i are positive integers bounded above by some constant W. Let $W_n = \sum_{i=1}^{n} w_i$ be the partial sums of the weights. For all $n \in \mathbb{N}$, we consider the WVG defined by the first n weights in the sequence, normalized to 1:

$$\mathcal{G}_n = \left(\frac{w_1}{W_n}, \ldots, \frac{w_n}{W_n}; q\right).$$

A much-studied question asks whether the powers assigned to the players by a power index $\boldsymbol{\varphi}$ are proportional to their weights, as n grows large. Since all weights of individual players go to 0 in $\mathcal{G}_n$ as $n \to \infty$, it is natural to look at the ratios of the powers of players.

Definition 4.1 ([Lindner and Machover, 2004, Penrose, 1946]). The *Penrose Limit Theorem (PLT)* holds for a power index $\boldsymbol{\varphi}$, a quota q, and a weight sequence $\boldsymbol{w}$ if for all players i, j,

$$\lim_{n\to\infty} \frac{\varphi_i(\mathcal{G}_n)}{\varphi_j(\mathcal{G}_n)} = \frac{w_i}{w_j}.$$

Remark. Since the Penrose Limit Theorem is concerned with ratios of powers, it does not matter whether we normalize the weights or not—the definitions and theorems in this section could equivalently be stated for power *measures*.

Let us define two useful properties of infinite weight sequences: We say that $\boldsymbol{w}$ is *primitive* if the greatest common divisor of all weights that appear infinitely often is 1. We say $\boldsymbol{w}$ is *regular* if every weight has a well-defined natural density: For all $w \in [W]$, there are values

$$d_w = \lim_{n\to\infty} \frac{|\{i \in [n] : w_i = w\}|}{n}.$$

Lindner and Machover [2004] showed that the PLT holds with respect to the ordinary Banzhaf power index ($q = 1/2$) for any primitive $\boldsymbol{w}$. They then ask whether this holds at other quotas and conjecture that the answer is positive. This was disproved by Lindner and Owen [2007] by giving a class of primitive weight sequences $\boldsymbol{w}$ and a quota $q \neq 1/2$ for which the PLT fails. However, no general formula for the Banzhaf powers at $q \neq 1/2$ or explanation for the behavior of the Banzhaf power index at quotas other than $1/2$ was known. We resolve this open problem.

Theorem 4.2 (Convergence for propensity Banzhaf). *The Penrose Limit Theorem holds for any primitive weight sequence* $\boldsymbol{w}$ *with respect to the adaptive Banzhaf power index.*

Furthermore, if $\boldsymbol{w}$ *is both primitive and regular, then for all players* i, j*,*

$$\lim_{n\to\infty} \frac{\beta_i^p(\mathcal{G}_n)}{\beta_j^p(\mathcal{G}_n)} = \frac{\left(\sum_{k=1}^{w_i} r^{k-1}\right)(1-p+pr^{w_j})}{\left(\sum_{k=1}^{w_j} r^{k-1}\right)(1-p+pr^{w_i})},$$

where r *is the unique positive solution to* $\sum_{w=1}^{W}(d_w \frac{pwr^w}{1-p+pr^w}) = q \cdot \sum_{w=1}^{W}(d_w w)$.

Remark. If $p = q$, it follows that $r = 1$ and thus that the limiting ratio is w_i/w_j. Consequently, the first part of the theorem (PLT if $p = q$) for primitive, regular weight sequences follows from the second.

We defer the proof to Appendix B.1. On a high level, we show that what determines the powers under the p-propensity Banzhaf power index in the limit is the limiting behavior of the distribution of the weight of a random coalition (where each voter joins with probability p) around the quota q. In particular, applying a theorem due to Petrov [1975] that was also used by Lindner and Machover [2004], we show that this distribution is "flat" around q when $p = q$, which implies that the PLT holds. In all other cases, we determine the exponential decay rate of the coalitional weight distribution around q, corresponding to $\ln r$ for the r in the theorem statement. We apply a technique called *exponential tilting* to the distribution to transform it to a distribution to which we can meaningfully apply Petrov's argument to show that it is flat around q. This allows us to describe the limiting behavior of the original distribution in terms of r, which is used to obtain the formula in the theorem statement.

We then solve for r in Theorem 4.2 to get insights into how the ratios of player powers behave in the limit.

Corollary 4.3 (Trends of power ratios). *Assume* $\boldsymbol{w}$ *is a primitive, regular weight sequence. with* $w_i < w_j$ *for some* i, j*. Then* $\lim_{n\to\infty} \beta_i^p(\mathcal{G}_n)/\beta_j^p(\mathcal{G}_n)$ *is "U-shaped" as a function of the quota* q*: It has a unique minimum* $q_{\min}$ *in* $(0, 1)$ *and non-zero derivative at all other points. The derivative of the function at* $q = p$ *is positive when* $p > 1/2$*, negative when* $p < 1/2$*, and zero when* $p = 1/2$*.*

The proof can be found in Appendix B.1. To make the statement easier to parse, we illustrate $\lim_{n\to\infty} \beta_i^p(\mathcal{G}_n)/\beta_j^p(\mathcal{G}_n) > w_i/w_j$ as a function of q in Figure 2.

Corollary 4.3 offers an explanation for why veto players are a frequent occurrence if $q \gg p \geq 1/2$. Since we know that at $q = p$, the limiting ratio is w_i/w_j, we know that for $q > p \geq 1/2$ it holds that $\lim_{n\to\infty} \beta_i^p(\mathcal{G}_n)/\beta_j^p(\mathcal{G}_n) > w_i/w_j$: Players with large weight have disproportionally little power. Thus, if their power share is required to match their population target, their weight share needs to be far greater than this target. Conversely, we get that for $p > 1/2$ and q slightly smaller than p, it holds that $\lim_{n\to\infty} \beta_i^p(\mathcal{G}_n)/\beta_j^p(\mathcal{G}_n) < w_i/w_j$: Players with large weight have disproportionally high power, so in the inverse power problem their weight share will be smaller than their target. We confirm that these trends also hold for weighted voting games with a finite number of players in Section 6.

Corollary 4.3 also highlights that adaptive Banzhaf is the only p-propensity Banzhaf power index for which the powers converge to weights in our setting. Since every semivalue that is anonymous to the voters and assumes they are voting independently is a p-propensity Banzhaf index for some p, this implies that:

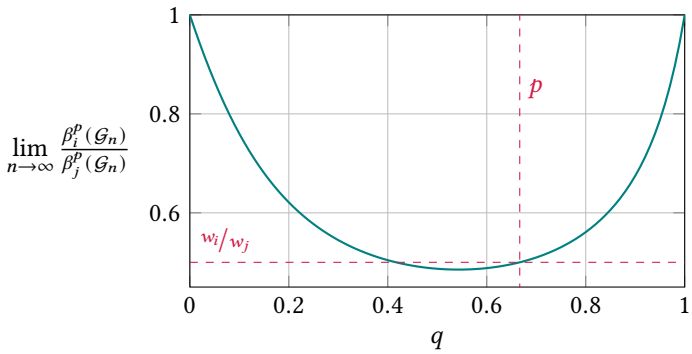


Fig. 2. An illustrative plot of the convergence of the 2/3-propensity Banzhaf powers for a player of weight $w_i = 1$ and a player of weight $w_j = 2$ for the repeating weight sequence $\boldsymbol{w} = 1, 1, 2, 1, 1, 2, 1, 1, 2, \ldots$. At $q = p$, the power ratio limits to the weight ratio. Because the plot is U-shaped, there is a second value $\tilde{q} < p$ with the same limiting ratio $w_i/w_j = 1/2$. This turns out to occur when $r = 1/2$, which gives $\tilde{q} = 5/12$.

Corollary 4.4 (Uniqueness of adaptive Banzhaf). *The adaptive Banzhaf power index is the only normalized semivalue that assumes voters are voting independently and satisfies the Penrose Limit Theorem for all primitive, regular weight sequences $\boldsymbol{w}$ and all quotas q.*

The proof of Corollary 4.4 can be found in Appendix B.1.

To conclude the section, we consider if there is hope to relax the conditions of primitivity and regularity for the weight sequence. Lindner and Machover [2004] already point out that without primitivity, the PLT can fail. For example, consider the weight sequence $w_1 = 1$ and $w_i = 2$ for all $i \geq 2$ with quota $1/2$. For any odd number of players, player 1 is pivotal in no coalition, so their power is 0. For any even number of players, it is not hard to verify that player 1 is pivotal if and only if any of the players of weight 2 is pivotal, so the powers of all players are equal. Thus, the limit $\beta_1^p(\mathcal{G}_n)/\beta_2^p(\mathcal{G}_n)$ does not exist, as the ratio jumps between 0 and 1.

Let us now consider regularity. If the limits of the fraction of voters that have a given weight, i.e., the natural densities, do not exist and $q \neq p$, the limit $\beta_1^p(\mathcal{G}_n)/\beta_2^p(\mathcal{G}_n)$ may not exist either. We give an informal argument: Consider the finite weight sequences $\boldsymbol{v}_1 = 1, 2, 1$ and $\boldsymbol{v}_2 = 1, 2, 2$, quota 4/5, and the standard (i.e., 1/2-propensity) Banzhaf power index. By Theorem 4.2, the ratio β_1^p/β_2^p, i.e., of powers of a player with weight 1 to a player with weight 2, approaches $1 : 1.662\ldots$ for $\boldsymbol{w} = \boldsymbol{v}_1, \boldsymbol{v}_1, \ldots$ and approaches the different ratio $1 : 1.754\ldots$ for $\boldsymbol{w} = \boldsymbol{v}_2, \boldsymbol{v}_2, \ldots$. We can now create a weight sequence $\boldsymbol{w}$ consisting of alternating blocks of just $\boldsymbol{v}_1$ and just $\boldsymbol{v}_2$, with the blocks increasing in size. By making each block sufficiently much longer than all preceding blocks, we know that after the end of each $\boldsymbol{v}_1$ block, β_1^p/β_2^p is arbitrarily close to $1 : 1.662\ldots$, while after each $\boldsymbol{v}_2$ block, β_1^p/β_2^p is arbitrarily close to $1 : 1.754\ldots$. Thus, β_1^p/β_2^p does not converge, the limit as $n \to \infty$ does not exist. Note that in this example, the natural densities do not exist: At the end of a $\boldsymbol{v}_1$ block, d_1 will be arbitrarily close to 2/3, while after a $\boldsymbol{v}_2$ block, it will be arbitrarily close to 1/3. Also, note that this argument does not work if $q = p$, since both under $\boldsymbol{w} = \boldsymbol{v}_1, \boldsymbol{v}_1, \ldots$ and $\boldsymbol{w} = \boldsymbol{v}_2, \boldsymbol{v}_2, \ldots$, the ratio of powers ρ_1/ρ_2 approaches $1 : 2$.

4.2 Oceanic games

A second setting of weighted voting in the limit that has been extensively discussed in the literature is the setting where some *large* players have a fixed weight while the remaining weight is split up evenly among a growing number of *small* players. In particular, we can fix weights $w_1, \ldots, w_\ell$ of ℓ

large players summing to no more than 1 and a quota $q \in (0, 1)$. Let $\alpha = 1 - \sum_{i=1}^{\ell} w_i$ be the leftover weight. We now consider the WVGs

$$\mathcal{G}_n = (w_1, \ldots, w_\ell, \underbrace{\frac{\alpha}{n}, \ldots, \frac{\alpha}{n}}_{n}; q),$$

defined for $n = 1, 2, \ldots$, where the small players divide up the leftover weight. In slight abuse of notation, n here denotes the number of small players in the game, so that the total number of players is $|N| = n + \ell$.

Theorem 4.5 (Oceanic propensity Banzhaf). *For $i \in [\ell]$,*

$$\lim_{n\to\infty} \beta_i^p(\mathcal{G}_n) = \begin{cases} \beta_i^p(\mathcal{G}_0^p) & \text{if } q - \alpha p \in [0, 1-\alpha] \setminus A_0 \\ 0 & \text{else,} \end{cases}$$

where $\mathcal{G}_0^p = (\frac{w_1}{1-\alpha}, \ldots, \frac{w_\ell}{1-\alpha}; \frac{q-\alpha p}{1-\alpha})$ is a game restricted to the large players and $A_0 = \{\sum_{i\in S} w_i \mid S \subseteq [\ell]\}$ are the weights achievable by large players.

The special case of Theorem 4.5 for the standard Banzhaf power index ($p = 1/2$) was established by Dubey and Shapley [1979]. The formal proof of this generalization is given in Appendix B.2 and follows the same proof strategy.

The theorem states that there are two different scenarios for the powers of the large players, $i \in [\ell]$, in the limit, depending on the relation of the quota to α, p, and A_0. If $\alpha p \le q \le 1 - \alpha + \alpha p$ and $q - \alpha p \notin A_0$, the total power of the large players approaches 1; in all other cases, the total power of the large players goes to 0. It is noteworthy that the latter case stems from two regimes that differ in their stability: Following Dubey and Shapley [1979], we define a set of *pitfall points* for the quota as $P = \alpha p + A_0$. A quota taking a value at the pitfall points leads to the limit of the game being *unstable*: A small perturbation to q, p or the large player weights will change the behavior in the limit so that the total large player power goes to 1 (instead of 0). By contrast, for q outside the range $[\alpha, 1 - \alpha + \alpha p]$, the limit is *stable*: the fact that the total power of the large players goes to 0 is robust to perturbations. We illustrate these regions in Figure 3.

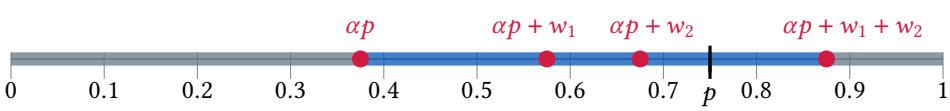


Fig. 3. The different regions for q that determine player power in the oceanic limit with $\ell = 2$ large players of weight $w_1 = 0.2$ and $w_2 = 0.3$ at propensity $p = 3/4$. Here, $\alpha = 0.5$ is the share of weight made up by the small players. The blue regions correspond to the interval $\alpha p + [0, 1 - \alpha]$. The weights achievable by large players are $A_0 = \{0,\ w_1,\ w_2,\ w_1 + w_2\}$ and the "pitfall points" $P = \alpha p + A_0$ are marked in red. As long as the quota falls in the blue region, the large players retain power as the small players shrink.

A key observation to understand these cases is that in the limit, the fraction of small voters who vote yes is tightly concentrated around p, so the weight that they contribute to the coalition is tightly concentrated around αp. Now, if the quota is between αp and $\alpha p + (1 - \alpha)$, some large players are going to be pivotal at these most-likely weight sums. If there exists a coalition of large players $S \subseteq [\ell]$ so that $q = \alpha p + \sum_{i\in S} w_i$ (i.e., if $q \in P$), then small players become pivotal as well in the limit. In the proof, we show that in both the case that $q \in P$ (where both large and small voters are pivotal in the most-likely coalitions) and $q \notin (\alpha p, 1 - \alpha(1 - p))$ (where neither large nor

small voters are pivotal in the most-likely coalitions), the sum of pivotal probabilities is dominated by the small players, causing the large players to have shrinking power, going to 0 in the limit.

At all other quotas, where only large players are pivotal in the most-likely coalitions, the sum of pivotality probabilities for all small players goes to zero, while the pivotality probability of each large player tends to the limit derived from the reduced WVG $\mathcal{G}_0^p = (\frac{w_1}{1-\alpha}, ..., \frac{w_\ell}{1-\alpha}; \frac{q-\alpha p}{1-\alpha})$. This particular weighted voting game arises from assuming that p share of the small players always vote YES and the remaining small players always vote NO.

In this family of games, we see the special position that the quota-equals-propensity case occupies in the family of propensity-Banzhaf power indices. In particular, it is not hard to verify that $p = q$ is the only value of p for which $0 \leq q - \alpha p \leq 1 - \alpha$ for all $\alpha \in (0, 1)$. Thus, $p = q$ is the only value of p for which the large players retain non-zero power for all $\alpha \in (0, 1)$ (besides the unstable pitfall points). Furthermore, we note that $p = q$ is the only value of p for which the quota $\frac{q-\alpha p}{1-\alpha}$ in the reduced game $\mathcal{G}_0^p$ is equal to the original quota q. Thus, only at $p = q$ are the powers as if the small players didn't exist and all large-player weights were scaled up equally to sum to 1, with the quota unchanged:

Corollary 4.6 (adaptive Banzhaf in the oceanic case). *It holds that*

$$\lim_{n\to\infty} \rho_i(\mathcal{G}_n) = \rho_i\big((\tfrac{w_1}{1-\alpha}, ..., \tfrac{w_l}{1-\alpha}; q)\big),$$

unless there exists $S \subseteq [l]$ for which $\sum_{i\in S} w_i = q(1-\alpha)$.

We conclude this section with two remarks: First, note that there are obstructions to obtaining a result in the style of the Penrose Limit Theorem (see Definition 4.1) for the ratio of the players' powers. In the first case of Theorem 4.5, $q \in [\alpha p, 1 - \alpha(1-p)] \setminus P$, pairs of large players will violate the PLT, while in the second case, a pairing of large player and small player will violate the PLT (with their power ratio converging to a constant). Finally, we note that in the literature the term "oceanic" games is used for the greater class of WVGs where the weight of all small players goes to 0, but they are not necessarily all equal. However, Dubey and Shapley [1979] show that in this more general setting, the powers of the large players under the (1/2-propensity) Banzhaf power index do not necessarily converge, even at quota 1/2. Therefore our narrower definition of oceanic games is a reasonable place to look for positive results.

5 Veto Distortion

A player i in a WVG $\mathcal{W} = (\boldsymbol{w}, q)$ is called a *veto player* if $w_i \geq 1 - q$, or equivalently, i is included in every winning coalition.[4] Intuitively, veto players have great power: Since no coalition of voters excluding them is winning, their approval is necessary for any motion to pass.

As observed in Section 1.2, it can happen in an instance of the inverse power problem that the discrepancy-minimizing weights for a power index include a veto player. Generally, this is not necessarily concerning: If a player's target power m_i (for example, their population share) exceeds the *veto threshold* of $1 - q$, it is not unexpected that also their weight w_i^* (in an optimal weight distribution $\boldsymbol{w}^* \in W^*(\boldsymbol{m}, q, \boldsymbol{\varphi})$ for power index $\boldsymbol{\varphi}$) exceeds the veto threshold, making them a veto player. In contrast, it *is* concerning if a player whose target power m_i is far below the veto threshold becomes a veto player. We call such a player i with $m_i \leq 1 - q$ but $w_i^* > 1 - q$ an *undeserving* veto player.

In this spirit, it is natural to investigate *how* undeserving a veto player is: How small can the m_i of a player be that receives a veto for some optimal weights $\boldsymbol{w}^*$? To answer this question, we define the *veto distortion* of an inverse power problem instance with target distribution $\boldsymbol{m}$, quota q, and

[4]In the case of a non-strict quota, the veto player condition becomes $w_i > 1 - q$.

power index $\boldsymbol{\varphi}$ as the ratio between the veto threshold of $1-q$ and the smallest target distribution value m_i of a player i that ended up as a veto player in some discrepancy-minimizing weight vector $\boldsymbol{w}^*$. If this fraction is less than one, no undeserving player of weight less than $1-q$ was made a veto player in any discrepancy-minimizing weight vector, so we set the distortion to 1.[5] That is,

$$\text{veto-dist}(\boldsymbol{m}, q, \boldsymbol{\varphi}) = \max\left\{\frac{1-q}{\min_{\boldsymbol{w}^* \in W^*(\boldsymbol{m},q,\boldsymbol{\varphi})} \min_{i \in N: w_i \geq 1-q} m_i}, 1\right\}.$$

We let the *veto distortion* of power index $\boldsymbol{\varphi}$ for fixed quota q be the worst-case veto distortion of this power index for any target distribution $\boldsymbol{m}$. With a slight abuse of notation, we write

$$\text{veto-dist}_q(\boldsymbol{\varphi}) = \max_{\boldsymbol{m} \in \Delta^{\mathbb{N}}} \text{veto-dist}(\boldsymbol{m}, q, \boldsymbol{\varphi}).$$

Theorem 5.1 (Arbitrary veto distortion). *For any $q \in (1/2, 1)$ and any $p \in (0,1)$, the veto distortion veto-dist$_q(\boldsymbol{\beta}^p)$ is unbounded.*

The key idea of the proof is to construct, for any quota q and propensity p, a family $\{\mathcal{G}_n\}$ of WVGs with weights $\{\boldsymbol{w}^n\}$ and quota q that all have a veto player whose p-propensity Banzhaf power can be made arbitrarily small for sufficiently large n. We then define the target distributions $\{\boldsymbol{m}^n\}$ to be precisely the p-propensity Banzhaf powers of those WVGs $\{\mathcal{G}_n\}$. Now, since $\boldsymbol{m}^n$ are the p-propensity Banzhaf powers of WVG $\mathcal{G}_n$, we know that by definition the weights $\boldsymbol{w}^n$ are going to have discrepancy 0 and thus be optimal. However, this already implies Theorem 5.1: By our assumption on $\{\mathcal{G}_n\}$, we can make the power of the veto player $\beta_i^p(\mathcal{G}_n)$, and thus their target distribution value m_i^n, arbitrarily small, while ensuring that the weights $\boldsymbol{w}^n$ for which i is a veto player are optimal weights. We can make the denominator of the expression defining veto distortion arbitrarily small, leading to unbounded veto distortion. We give the construction of these $\{\mathcal{G}_n\}$, which is based on Theorem 4.5, and a formal proof in Appendix B.3.

For $q = 1/2$ and a non-strict quota, the veto distortion behaves a lot more nicely:

Theorem 5.2. *For any $p \in (0,1)$, if the quota is not strict (so that weights $\geq q$ are winning), then there is no distortion at simple majority: veto-dist$_{1/2}(\boldsymbol{\beta}^p) = 1$.*

The proof of Theorem 5.2 can be found in Appendix B.3. The key idea is that for non-strict quota $1/2$, any veto player is also a *dictator*: Any coalition that includes them is winning. From this, we can deduce that the power of any veto player is 1, so it suffices to prove that the power distribution $(1, 0, ..., 0)$ is never optimal if no player's target distribution value exceeds the veto threshold.

There remains the case $q = 1/2$ with strict quota (i.e., weights $> q$ are winning). That situation can be distinguished from the non-strict case by an example. For the target distribution $\boldsymbol{m} = (5/12, 5/12, 2/12)$, one can easily confirm that in this case, veto-dist$_{1/2}(\boldsymbol{\beta}^p) \geq 6/5$.

The results in this section may at first seem bleak. However, the proof of Theorem 5.1 relies on engineering games that put p, q in an unstable "pitfall" relationship from Theorem 4.5. Due to this instability, we conjecture that for $q \approx p$, a random target distribution will have little to no veto distortion with probability 1. As partial confirmation, the empirical results below in Section 6 show that undeserving veto players are frequent for $q \ll p$ or $q \gg p$, while they are rare in $p \approx q$ cases.

[5]We only focus on the issue of a player that does not deserve a veto (as $m_i \leq 1-q$) becoming a veto player. A related question is whether it can happen that a player that would deserve a veto ends up without a veto in the optimal weights. We believe the latter phenomenon is significantly less concerning; it is conceivable that some player of large target distribution (e.g., population share) may need to forgo their veto for a discrepancy-minimizing power distribution.

6 Empirical Results

6.1 Predictions from theory

As we have seen, the objective to minimize the discrepancy of the normalized vectors $\boldsymbol{m}$ and $\boldsymbol{\beta}$ can lead to optimized weights $\boldsymbol{w}$ with $\boldsymbol{w} \not\approx \boldsymbol{\beta}$. In particular, Corollary 4.3 suggest that at large quotas (larger than the propensity), the power of the largest player will lag far behind their weight. Thus, to get the power to hit a target will require the allocation of a large weight. This is exactly what has been observed for the ordinary Banzhaf power index ($p = 1/2$) at supermajority quotas: this was noted for the United States at the IMF at $q = 85\%$ [Leech, 2002c], for Germany in the Council of Ministers of the EU at quotas exceeding 75% [Leech and Machover, 2003], and in our consulting work for Ontario County.

On the other hand, our theoretical results from Section 4.1 tell us that in several limiting constructions, the adaptive Banzhaf power index leads to proportional optima

$$\boldsymbol{m} \approx \boldsymbol{w} \approx \boldsymbol{\rho}.$$

We now design a range of experiments to test whether the asymptotic predictions are already observable in small finite games.

6.2 Experiment setup

In the experiments, we focus on two forms of *distortion*: In an instance of the inverse power problem with target distribution $\boldsymbol{m}$ such that $m_1 \leq \ldots \leq m_n$, quota q, and power index $\boldsymbol{\varphi}$, the largest-player distortion of a weight vector $\boldsymbol{w}$ is the (signed) difference between the weight and target of the largest player:[6]

$$\text{large-player-distortion}_{\boldsymbol{m},q,\boldsymbol{\varphi}}(w) = w_n - m_n.$$

Similarly, we define the total distortion to be the L^1 distance between the weights and the target distribution,

$$\text{total-distortion}_{\boldsymbol{m},q,\boldsymbol{\varphi}}(w) = \|\boldsymbol{w} - \boldsymbol{m}\|_1 .$$

We study these notions of distortion for optimized weights in real-world instances: counties in New York State, especially Ontario, and executive directors of the International Monetary Fund.

Unless otherwise noted, we find optimized weights giving powers close to the desired target using a simple Markov chain method. Weights are initialized to be proportional to populations. Then, in every iteration, we sample a player with probability proportional to the absolute difference between that player's power index and population share (the power target). The weight of that player is then adjusted in the direction that would reduce the discrepancy by a small, random step size. The search terminates after 1000 consecutive steps in which the L^1 discrepancy did not improve. We empirically observe that this approach stabilizes, with multiple runs eventually converging to roughly the same near-optimal solution, with tiny values for the objective function, often with $|\boldsymbol{m} - \boldsymbol{\varphi}| < .001$ for vectors with entries summing to 1. Figure 8 in Appendix C.2 shows the progression of one such run as it converged to optimized weights for quota $3/4$ and propensity $1/2$ (i.e., the exact setting described in Section 1.2).

To confirm that the results are not just artifacts of our heuristic algorithm, we employ an integer linear programming (ILP) approach due to Kurz [2012] to find *globally* optimal weights for small instances, confirming our findings. Exact results are presented in Appendix C.3.

[6]It is not hard to check that this largest-target player will also be the player with the largest weight in an optimal solution.

6.3 Ontario County

Data. Recall that there are 21 members of the Board of Supervisors in Ontario County, who pass measures by weighted voting. The sizes of the populations they represent, as well as the weights optimized for quotas $1/2$, $3/5$, $2/3$, $3/4$, and $4/5$, can be found in the supplementary material in Table 1. The largest player in this weighted voting game is the Town of Victor with a population share of 14.1%.

Experiments. For every combination of propensity p and quota q in the set $\{50\%, 51\%, \ldots, 99\%\}$, we found heuristically optimized weights under p-propensity Banzhaf power index with a Markov chain local search as described above. Victor's weight distortion is shown in Figure 4, while the heatmap showing the total distortion is deferred to Appendix C.

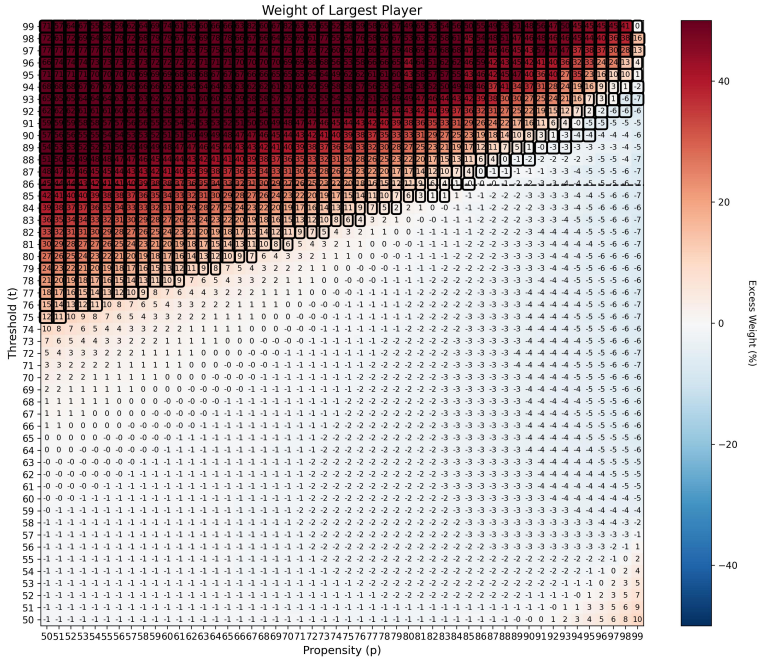


Fig. 4. Largest-player distortion for the Town of Victor, where $\boldsymbol{m}$ is the (fixed) vector of population shares in Ontario County and $\boldsymbol{w}$ has been heuristically optimized for each (p, q) pair. Outlined boxes represent situations where the weight boost makes Victor a veto player. The dashed line represents the threshold at which Victor would be a veto player with naive weights $\boldsymbol{w} = \boldsymbol{m}$.

These results empirically confirm that the predictions of Corollary 4.3 hold in a 21-town instance, rather than merely asymptotically. In particular, the U-shape of the limiting ratio is visible. When the quota is much larger than the propensity, the largest town has power lagging far behind weight, so a near-optimal solution calls for massively high weights so that the power can hit its target. When the propensity is roughly equal to the quota, the weight of the largest town is roughly proportional. For quotas less than the propensity, the optimized weight of the largest player first decreases, before increasing again in the bottom right corner. We note that the line of zeroes, where weight and power match, is not exactly along the main diagonal $p = q$ (adaptive Banzhaf) as the asymptotic theory predicts. However, along this main diagonal, Victor becomes a veto player right around the line $q = 1 - m$ where its population is large enough to "deserve" it.

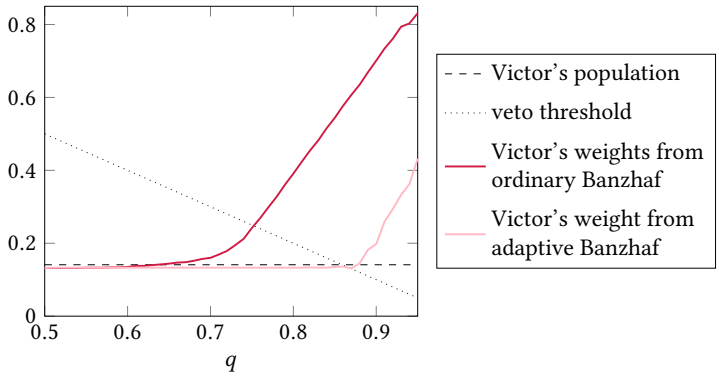


Fig. 5. In Ontario, adaptive Banzhaf causes Victor's weight to stay close to proportional until they become a deserving veto player, after which their weight grows quickly.

By contrast, ordinary Banzhaf is observed along the left-hand edge of the square, at propensity $p = 1/2$. There, the optimized weight of Victor increases almost linearly with the quota, for quotas greater than 70%. Figure 5 shows another view.

6.4 Counties in New York State

Data. In each of the 16 counties in New York State that use weighted voting, we obtain the voting weights for the Board of Supervisors from the local laws in the counties; the populations of the towns they represent are pulled from the 2020 Decennial Census [U.S. Census Bureau, 2020].

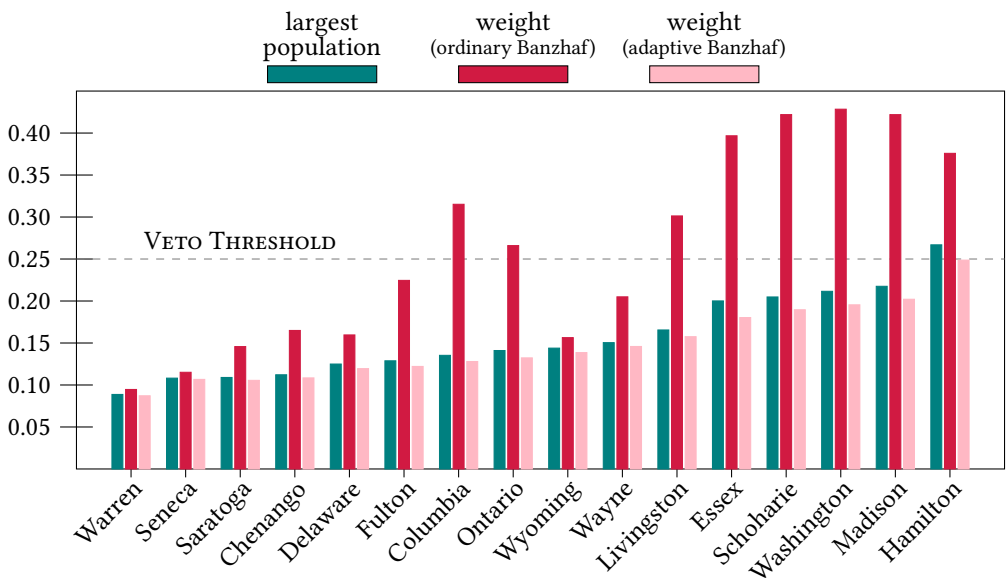


Fig. 6. For each of the 16 counties in New York State that use weighted voting, we show the largest player's population and their weight in the optimized weights under the Banzhaf power index and adaptive Banzhaf power index at quota $q = 3/4$.

Experiment and Results. For each county, we find optimized weights under both the ordinary Banzhaf index and the adaptive Banzhaf index at quota $q = 75\%$. We find that in all 16 counties, the largest player distortion is negative for the adaptive Banzhaf index and positive for the ordinary Banzhaf index, with the latter dominating in magnitude. In seven counties, thus almost half of the analyzed counties, the largest player has a population share (i.e., target) below the veto threshold, but under the Banzhaf power index receives weight above the veto threshold, making them an undeserving veto player. In some cases, the largest player's weight is more than twice their proportional share. In one county, Hamilton, the largest town's population just exceeds the veto threshold, while their optimized adaptive Banzhaf weight is just below: a town "deserving" of a veto does not get one. We plot the largest player population and weight, from which the largest player distortion can easily be inferred, in Figure 6. Furthermore, in all 16 counties, the total distortion of the Banzhaf index is significantly larger than that of the adaptive Banzhaf index. We plot this in Figure 7.

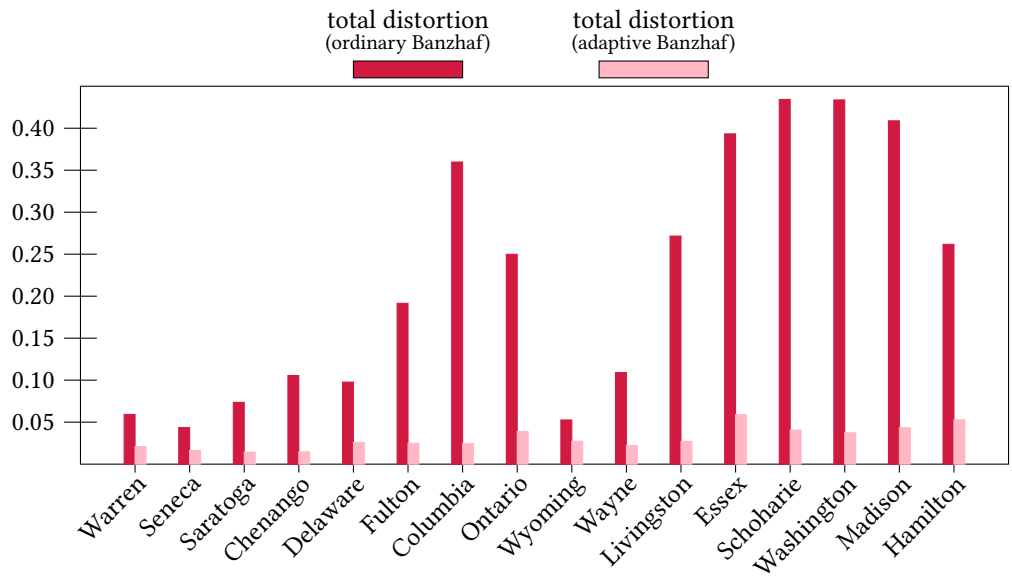


Fig. 7. For each of the 16 counties in New York State that use weighted voting, we show the total distortion, the L^1 distance between optimized weights and populations, under the Banzhaf power index and the adaptive Banzhaf power index at quota $q = 3/4$.

6.5 Additional experiments

We analyze voting in the International Monetary Fund in Appendix C.5. Weight distortion for the largest player (the United States) at supermajority quotas was identified as paradoxical and problematic in Leech [2002c]. Here again, optimizing weights for the adaptive Banzhaf index works in practice, alleviating the distortion paradox.

Heatmaps displaying the total distortion and largest-player distortion for all combinations of propensity p and quota q in the set $\{50\%, 51\%, \ldots, 99\%\}$ for Ontario County, Livingston County, and a set of synthetic examples can be found in Appendix C.4. We consider two synthetic populations that resemble a finite version of the oceanic case and find that the pitfall points are clearly visible from the optimized weight vectors.

7 Discussion

We have argued that the properties of the adaptive Banzhaf power index make it an appealing alternative to the ordinary Banzhaf power index, leading to more proportional weights in the inverse power problem while still abiding by the essential assumptions that players vote anonymously and independently. Due to the independence of the voters, it is a policy-seeking power index in the formulation of Felsenthal and Machover [1998], measuring a fundamentally different notion of power than the office-seeking Shapley-Shubik index. Thus, it conserves the known conceptual and theoretical advantages of the Banzhaf index over the Shapley-Shubik index for measuring voting power.

There is another body of work giving signs in favor of the adaptive Banzhaf index over the ordinary Banzhaf index and Shapley-Shubik index. Leech [2002a] argues that shareholders in publicly traded companies offer real-world approximations to oceanic games, with few large players and a large number small players present. Drawing on classic work from Berle and Means [1932], Leech argues that a knowledgeable analysis of the powers should indicate that the practical control over decisions (i.e., the power) of the large players greatly exceeds their voting weight. Therefore, a power index that reflects this reality, when applied to publicly traded companies, should in most cases assign disproportionately much power to the few players with large voting weight. Analyzing data from publicly traded companies in the UK, Leech [2002a] finds this is the case for the Banzhaf index but not for Shapley-Shubik index, at a quota of $1/2$. In the light of our results for oceanic games (Theorem 4.5), this is not surprising: Unless the large player weights lead to a (very unlikely) pitfall point, the ordinary Banzhaf power index at quota $q = 1/2$ will lead to the large players having disproportionally high power. Interestingly, this no longer holds for the Banzhaf power index at any quota other than $1/2$, when α is sufficiently close to 1. Instead, by Theorem 4.5, we see that the adaptive Banzhaf power index is the only p-propensity Banzhaf power index that satisfies Leech's criterion for all quotas q and any α. We see this as intriguing evidence on the side of adaptive Banzhaf.

Let us close where we started, in Victor, New York. Our results show that the seeming necessity of the Victor veto for supermajority voting is merely a byproduct of contestible modeling assumptions. Adaptive Banzhaf power can serve as a tool for institutional design, aligning mathematical notions of power with normative expectations and avoiding distortions that undermine the legitimacy of real-world voting systems.

References

N. Alon and P. H. Edelman. 2010. The Inverse Banzhaf Problem. *Social Choice and Welfare* 34, 3 (2010), 371–377.

R. Amer and J. M. Giménez. 2007. Technical Note: Characterization of Binomial Semivalues Through Delegation Games. *Naval Research Logistics* 54, 6 (2007), 702–708.

H. Aziz, M. Paterson, and D. Leech. 2008. Efficient Algorithm for Designing Weighted Voting Games. In *Proceedings of the 11th IEEE International Multitopic Conference*. 1–6.

Y. Bachrach, E. Markakis, E. Resnick, A. D. Procaccia, J. S. Rosenschein, and A. Saberi. 2010. Approximating Power Indices: Theoretical and Empirical Analysis. *Autonomous Agents and Multi-Agent Systems* 20 (2010), 105–122.

J. F. Banzhaf. 1965. Weighted Voting Doesn't Work: A Mathematical Analysis. *Rutgers Law Review* 19 (1965), 317–343.

J. F. Banzhaf. 1968. One Man, 3,312 Votes: A Mathematical Analysis of the Electoral College. *Villanova Law Review* 13 (1968), 304–332.

A. Berle and G. Means. 1932. *The Modern Corporation and Private Property*. Commerce Clearing House.

S. Brenner. 1978. The Shapley-Shubik Power Index and the Supreme Court. *American Political Science Review* 72, 2 (1978), 463–470.

P.-L. Chang, V. Chua, and M. Machover. 2006. L. S. Penrose's Limit Theorem: Tests by Simulation. *Mathematical Social Sciences* 51, 1 (2006), 90–106.

J. S. Coleman. 1968. Control of Collectivities and the Power of a Collectivity to Act. RAND Corporation.

J. Deegan and E. W. Packel. 1978. A New Index of Power for Simple n-Person Games. *International Journal of Game Theory* 7, 2 (1978), 113–123.

I. Diakonikolas, D. M. Kane, J. Peebles, and A. Stewart. 2022. Efficient Approximation Algorithms for the Inverse Semivalue Problem. *Mathematics of Operations Research* 47, 3 (2022), 1908–1939.

I. Diakonikolas and C. Pavlou. 2019. On the Complexity of the Inverse Semivalue Problem for Weighted Voting Games. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*.

P. Dubey. 1975. On the Uniqueness of the Shapley Value. *International Journal of Game Theory* 4 (1975), 131–139.

P. Dubey and L. S. Shapley. 1979. Mathematical Properties of the Banzhaf Power Index. *Mathematics of Operations Research* 4, 2 (1979), 99–131.

D. S. Felsenthal and M. Machover. 1998. *The Measurement of Voting Power*. Edward Elgar Publishing.

A. Gelman, J. N. Katz, and J. Bafumi. 2004. Standard Voting Power Indexes Don't Work: An Empirical Analysis. *British Journal of Political Science* 34, 4 (2004), 657–674.

C. M. Grinstead and J. L. Snell. 1998. *Introduction to Probability* (2 ed.). American Mathematical Society.

B. Grofman. 1981. Fair Apportionment and the Banzhaf Index. *Amer. Math. Monthly* 88, 1 (1981), 1–5.

R. J. Johnston. 1978. On the Measurement of Power: Some Reactions to Laver. *Environment and Planning A: Economy and Space* 10, 8 (1978), 907–914.

S. Kurz. 2012. On the Inverse Power Index Problem. *Optimization* 61, 8 (2012), 989–1011.

D. Leech. 2002a. An Empirical Comparison of the Performance of Classical Power Indices. *Political Studies* 50 (2002), 1–22.

D. Leech. 2002b. *Power Indices as an Aid to Institutional Design: The Generalised Apportionment Problem*. Economic Research Papers. University of Warwick.

D. Leech. 2002c. *Voting Power in the Governance Of The International Monetary Fund*. Economic Research Papers. University of Warwick.

D. Leech. 2003. Computing Power Indices for Large Voting Games. *Management Science* 49, 6 (2003), 831–838.

D. Leech and M. Machover. 2003. Qualified Majority Voting: The Effect of the Quota. In *European Governance*, M. J. Holler, H. Kliemt, D. Schmidtchen, and M. E. Streit (Eds.). 127–143.

I. Lindner and M. Machover. 2004. L.S. Penrose's limit theorem: Proof of Some Special Cases. *Mathematical Social Sciences* 47, 1 (2004), 37–49.

I. Lindner and G. Owen. 2007. Cases Where the Penrose Limit Theorem Does Not Hold. *Mathematical Social Sciences* 53, 3 (2007), 232–238.

T. Matsui and Y. Matsui. 2000. A Survey of Algorithms for Calculating Power Indices of Weighted Majority Games. *Journal of the Operations Research Society of Japan* 43 (2000), 71–86.

Y. Matsui and T. Matsui. 2001. NP-completeness for Calculating Power Indices of Weighted Majority Games. *Theoretical Computer Science* 263, 1 (2001), 305–310.

A. Mayer. 2018. Luxembourg in the Early Days of the EEC: Null Player or Not? *Games* 9, 2 (2018).

New York Court of Appeals. 1967. *Ianucci v. Board of Supervisors*.

J. Neyman. 1982. Renewal Theory for Sampling Without Replacement. *Annals of Probability* 10, 2 (1982), 464–481.

A. P. Ostrow. 2016. One Person, One Weighted Vote. *Florida Law Review* 68, 5 (2016), 1611–1660.

G. Owen. 1972. Multilinear Extensions of Games. *Management Science* 18, 5 (1972), 64–79.

L. S. Penrose. 1946. The Elementary Statistics of Majority Voting. *Journal of the Royal Statistical Society* 109, 1 (1946), 53–57.

V. V. Petrov. 1975. *Sums of Independent Random Variables*. Springer.

M. A. Puente del Campo. 2000. *Aportaciones a la Representabilidad de Juegos Simples y al Cálculo de Soluciones de Esta Clase de Juegos.* Ph. D. Dissertation. Universitat Politècnica de Catalunya.

N. Z. Shapiro and L. S. Shapley. 1978. Values of Large Games, I: A Limit Theorem. *Mathematics of Operations Research* 3, 1 (1978), 1–9.

L. S. Shapley and M. Shubik. 1954. A Method for Evaluating the Distribution of Power in a Committee System. *American Political Science Review* 48, 3 (1954), 787–792.

P. D. Straffin. 1977. Homogeneity, Independence, and Power Indices. *Public Choice* 30 (1977), 107–118.

U.S. Census Bureau. 2020. DP1: Profile of General Population and Housing Characteristics. Decennial Census, DEC Demographic Profile (Table DECENNIALDP2020.DP1). Accessed 2026-02-09.

R. J. Weber. 1979. Subjectivity in the Valuation of Games. *Econometrica* 47, 5 (1979), 1115–1130.

A Additional Related Work

Voting power. The formal study of voting power was initiated by Penrose [1946] studying the probability that a voter is on the 'winning side' of a YES/NO vote, when all other voters vote independently and uniformly at random. Independently, Shapley and Shubik [1954] developed the first power index, now known as the Shapley-Shubik index, by applying the Shapley value from cooperative game theory to weighted voting. After Banzhaf's paper made a splash in the 1960s, Weber [1979] generalized both the Shapley-Shubik and the Banzhaf index to a class of anonymous, probabilistic power indices called semivalues. Some less-well-known power indices, which do not fall into the class of semivalues, were defined by Deegan and Packel [1978] and Johnston [1978]. More information about these power indices and their respective justifications and shortcomings can be found in the survey by Felsenthal and Machover [1998].

Law and political science. Banzhaf's investigation of voting power came in the wake of major U.S. Supreme Court decisions in *Baker v. Carr* (1962) and *Reynolds v. Sims* (1964) interpreting the constitution to mandate equalization of voting weight with a strong basis in raw population. He argued against naive weighting (with weights presumed to be proportional to power) in [Banzhaf, 1965], with a follow-up in [Banzhaf, 1968]. Subsequent coverage in law reviews and political science journals has not been extensive but includes at least work by Brenner [1978], Gelman et al. [2004], Grofman [1981], and Ostrow [2016].

Complexity of computing the Banzhaf power index. Matsui and Matsui [2001] show that deciding whether a voter's Banzhaf (and Shapley-Shubik) power is non-zero is **NP**-complete, thus proving the problems of computing the Banzhaf power index or measure, and of even getting an multiplicative approximation to them, to be intractable. Many papers propose algorithms for additive approximations to the Banzhaf power index or measure in large instances [Bachrach et al., 2010, Leech, 2003, Matsui and Matsui, 2000].

The inverse power problem. The problem of finding weights giving powers close to a desired distribution is as old as the question of how to measure voting power. While the problem is known to be intractable [Diakonikolas and Pavlou, 2019], there exist many approximation algorithms [Aziz et al., 2008, Diakonikolas et al., 2022, Leech, 2002b]. Alon and Edelman [2010] prove that there exist certain power distributions for which no weighted voting game, no matter the number of players, leads to Banzhaf powers close to it. Kurz [2012] extends this analysis to the Shapley-Shubik index and proposes an approach based on integer linear programming for solving the inverse problem exactly.

Empirical evidence for Banzhaf vs. Shapley-Shubik. Clearly, we cannot expect either power index to actually coincide with observed coalitions in real-world settings. Since the power indices aim to make an *a priori* measurement—the power distribution in a voting body without consideration of the players' nature—they must be agnostic to ideological similarities between voters. Gelman et al. [2004] examine which of the two indices' assumptions on the distribution of coalition sizes matches data from states in U.S. presidential elections by looking at the vote share the winning party received. They find that while coalition sizes around the threshold of $1/2$ are more likely than extreme coalition sizes (which provides evidence against the Shapley-Shubik assumption), the distribution of the winning party's relative vote share around the threshold seems to be independent of the number of voters (which conflicts with the Banzhaf assumption). Leech [2002a] investigates the voting among shareholders in publicly traded companies in the UK and finds that the Banzhaf power index aligns with the qualitative comparison of practical power described in the literature much better than Shapley-Shubik.

B Missing Proofs

B.1 The Penrose Limit Theorem

An important building block is the following local limit theorem for lattice distribution, proved in Petrov [1975]. For us, it plays the role of a lemma.

Lemma B.1 (Local Limit for Lattice Distribution). *Let $\{\mathcal{D}_i\}_{i\in[W]}$ be a finite number of distributions over the integers, each with finite variance. The* span H_i *of distribution $\mathcal{D}_i$ is the smallest integer h for which there exists an integer a so that* $\Pr_{x\sim\mathcal{D}_i}[x \equiv a \pmod h] = 1$. *Let $X_1, X_2, \ldots$ be a sequence of independent random variables where each X_j is distributed according to one of the $\mathcal{D}_i$. Let $\mathcal{D}_1^*, \ldots, \mathcal{D}_{W^*}^*$ be all distributions according to which infinitely many X_i in the sequence are distributed and let their spans be $H_1^*, \ldots H_{W^*}^*$. If* $\gcd\left(H_1^*, \ldots, H_{W^*}^*\right) = 1$, *then the distribution is asymptotically normal:*

$$\sup_{x\in\mathbb{Z}}\left|\sigma_n \Pr\left[\sum_{j=1}^{n} X_j = x\right] - \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu_n}{\sigma_n}\right)^2\right)\right| \xrightarrow{n\to\infty} 0,$$

where $\mu = \mathbb{E}\left[\sum_{j=1}^{n} X_j\right]$ and $\sigma_n = \sqrt{Var\left(\sum_{j=1}^{n} X_j\right)}$.

Proof of Theorem 4.2. Without loss of generality, we'll assume that players i, j are players of weight w_1 ('player 1') and of weight w_2 ('player 2'), and that $w_1 < w_2$. Going forward, we'll thus reuse i and j as variables, no longer to denoting the 2 players in question. Furthermore, we'll refer to the WVGs in their unnormalized form $\mathcal{G}^{(n)} = \left(w_1, \ldots, w_r; qW^{(n)}\right)$ since this simplifies notation (note that this does not change which coalitions are winning and thus doesn't change the powers).

For $i \in \{3, \ldots, n\}$, we let $X_i \sim \text{Bernoulli}(p)$ be independent random variables representing how the ith player votes. Thus, $S_n = \sum_{i=3}^{n} w_i X_i$ denotes the weight of a random coalition, conditioned on excluding the first two voters.

We let $T_n = \left\lfloor qW^{(n)} \right\rfloor$ be the integer voting threshold: Any coalition with more than T_n weight is winning, any coalition with T_n or less weight is losing.[7] Thus, we can observe:

- Player 1 is pivotal if and only if $S_n \in \{T_n - w_1 - w_2 + 1, \ldots, T_n - w_2\}$ and player 2 votes yes or $S_n \in \{T_n - w_1 + 1, \ldots, T_n\}$ and player 2 votes no. Thus,

$$\hat{\beta}_1^p(\mathcal{G}_n) = p\Pr[S_n \in \{T_n - w_1 - w_2 + 1, \ldots, T_n - w_2\}] + (1-p)\Pr[S_n \in \{T_n - w_1 + 1, \ldots, T_n\}].$$

- Player 2 is pivotal if and only if $S_n \in \{T_n - w_1 - w_2 + 1, \ldots, T_n - w_2\}$ and player 1 votes yes, $S_n \in \{T_n - w_2 + 1, \ldots, T_n - w_1\}$ regardless of player 1's vote, or $S_n \in \{T_n - w_1 + 1, \ldots, T_n\}$ and player 1 votes no. Thus,

$$\begin{aligned}\hat{\beta}_2^p(\mathcal{G}_n) = p\Pr[S_n \in \{T_n - w_1 - w_2 + 1, \ldots, T_n - w_2\}] + \\ + \Pr[S_n \in \{T_n - w_2 + 1, \ldots, T_n - w_1\}] + (1-p)\Pr[S_n \in \{T_n - w_1 + 1, \ldots, T_n\}].\end{aligned}$$

Let's first consider adaptive Banzhaf. Note that $\frac{\rho_1(\mathcal{G}_n)}{\rho_2(\mathcal{G}_n)} = \frac{\beta_1^q(\mathcal{G}_n)}{\beta_2^q(\mathcal{G}_n)} = \frac{\hat{\beta}_1^q(\mathcal{G}_n)}{\hat{\beta}_2^q(\mathcal{G}_n)}$. To calculate the limit of

$$\frac{\hat{\beta}_1^q(\mathcal{G}_n)}{\hat{\beta}_2^q(\mathcal{G}_n)} = \frac{p\sum_{i=w_2}^{w_1+w_2-1}\Pr[S'_n = T_n - i] + (1-p)\sum_{i=0}^{w_1-1}\Pr[S'_n = T_n - i]}{p\sum_{i=w_2}^{w_1+w_2-1}\Pr[S'_n = T_n - i] + \sum_{i=w_1}^{w_2-1}\Pr[S'_n = T_n - i] + (1-p)\sum_{i=0}^{w_1-1}\Pr[S'_n = T_n - i]}, \quad (1)$$

[7] Note that the proof works identically if we set $T_n = \left\lceil qW^{(n)} \right\rceil - 1$, corresponding to the case when a coalition is winning if it meets the quota (but does not necessarily need to exceed it).

we'll first determine the ratio $\Pr[S_n = T_n - c]/\Pr[S_n = T_n]$ for $n \to \infty$ and a constant c.

The X_i are independent random variables from at most W different distributions (one for each possible value of w_i in $[W]$). The distribution of X_i has finite variance and span w_i. Thus, primitivity of the weight sequence implies that the greatest common divisor of the span of all distributions that appear infinitely often as distributions of X_i is 1, so we can apply Lemma B.1 to get that we can write

$$\Pr[S_n = x] = \frac{1}{\sigma_n}\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu_n}{\sigma_n}\right)^2\right) + \xi_n(x)\right)$$

for $\mu_n = q(W^{(n)} - w_1 - w_2)$, $\sigma_n = \sqrt{\sum_{i=3}^n q(1-q)w_i^2}$, and some error $\xi_n(x)$ where $\sup_{x\in\mathbb{Z}} |\xi_n(x)| \to 0$ as $n \to \infty$. Thus,

$$\frac{\Pr[S_n = T_n - c]}{\Pr[S_n = T_n]} = \frac{\exp\left(-\frac{1}{2}\left(\frac{T_n - c - \mu_n}{\sigma_n}\right)^2\right) + \xi_n(T_n - c)}{\exp\left(-\frac{1}{2}\left(\frac{T_n - \mu_n}{\sigma_n}\right)^2\right) + \xi_n(T_n)}.$$

Since both $|T_n - \mu_n|$ and $|T_n - c - n\mu|$ are bounded above by a constant independent of n but $\sigma_n \to \infty$ as $n \to \infty$, we get that both $\exp\left(-\frac{1}{2}\left(\frac{T_n - n\mu}{\sigma\sqrt{n}}\right)^2\right)$ and $\exp\left(-\frac{1}{2}\left(\frac{T_n - c - n\mu}{\sigma\sqrt{n}}\right)^2\right)$ go to $e^0 = 1$ as $n \to \infty$. In particular, since the limit of the denominator is non-zero, we get that for any constant c,

$$\frac{\Pr[S_n = T_n - c]}{\Pr[S_n = T_n]} \to 1$$

as $n \to \infty$. This is the key finding for the proof of the first part of the theorem: The distribution of the weight of a random coalition, S_n, is flat around the threshold T_n. Thus, the players powers are proportional to the size of their corresponding interval of pivotality, i.e. their weight. Formally, plugging into Equation (1), this tell us that

$$\frac{\hat{\beta}_1^q(\mathcal{G}_n)}{\hat{\beta}_2^q(\mathcal{G}_n)} \to \frac{pw_1 + (1 - pw_1)}{pw_1 + (w_2 - w_1) + (1-p)w_1} = \frac{w_1}{w_2}.$$

Thus, the PLT holds for adaptive Banzhaf at any quota $q \in (0, 1)$.

Let's now assume that the weight sequence is regular, i.e., that all the natural densities d_w for $w \in [W]$ are well defined, and consider the general case where $p \in (0, 1)$, not necessarily equal to the quota. We cannot directly use the same trick as in the $p = q$ case, since the ratio $\Pr[S_n = T_n - c]/\Pr[S_n = T_n]$ as $n \to \infty$ will not be 1 for all constant c (in terms of n) but instead depend on c. To overcome this, we use a technique that is standard in probability theory called *exponential tilting*, modifying S_n to make its expectation equal to $qW^{(n)}$: Let

$$\kappa_{S_n}(s) = \ln \mathbb{E}[e^{sS_n}] = \sum_{i=3}^n \mathbb{E}[e^{sw_iX_i}] = \sum_{i=3}^n \kappa_{X_i}(s)$$

be the cumulant generating function of S_n, where $\kappa_{X_i}(s) = \ln(1 - p + pe^{sw_i})$ is the cumulant generating function of X_i. We let s_n^* be the unique solution to

$$\kappa'_{S_n}(s_n^*) = \sum_{i=3}^n \frac{pw_ie^{w_is_n^*}}{1 - p + pe^{w_is_n^*}} = qW^{(n)}.$$

Note that there exists exactly one such s_n^* since $\kappa'_{S_n}(s)$ is continuous, $\kappa''_{S_n}(s) = \sum_{i=3}^n \frac{(1-p)pw_i^2e^{w_is}}{(1-p+pe^{w_is})^2} > 0$ for all s, $\lim_{s\to-\infty} \kappa'_{S_n}(s) = 0$, and $\lim_{s\to\infty} \kappa'_{S_n}(s) = W^{(n)}$.

We define the exponentially tilted random variables X_i' for $i \in \{3, ..., n\}$ by

$$\Pr[X_i' = x] = e^{s_n^* x - \kappa_{X_i}(s_n^*)} \Pr[X_i = x],$$

and $S_n' = \sum_{i=3}^n X_i'$ so that

$$\Pr[S_n' = x] = e^{s_n^* x - \kappa_{S_n}(s_n^*)} \Pr[S_n = x].$$

It are well-known facts about exponential tilting that the X_i and S_n' are well-defined random variables and that $\mathbb{E}[S_n'] = \kappa_{S_n}'(s_n^*) = qW^{(n)}$ and $\mathrm{Var}(S_n') = \kappa_{S_n'}''(s_n^*)$.

We can now write $\frac{\beta_1^p(\mathcal{G}_n)}{\beta_2^p(\mathcal{G}_n)} = \frac{\hat{\beta}_1^p(\mathcal{G}_n)}{\hat{\beta}_2^p(\mathcal{G}_n)} =$

$$\frac{p \sum_{i=w_2}^{w_1+w_2-1} e^{is_n^*} \Pr[S_n' = T_n - i] + (1-p) \sum_{i=0}^{w_1-1} e^{is_n^*} \Pr[S_n' = T_n - i]}{p \sum_{i=w_2}^{w_1+w_2-1} e^{is_n^*} \Pr[S_n' = T_n - i] + \sum_{i=w_1}^{w_2-1} e^{is_n^*} \Pr[S_n' = T_n - i] + (1-p) \sum_{i=0}^{w_1-1} e^{is_n^*} \Pr[S_n' = T_n - i]}. \quad (2)$$

Now we can apply the same technique as in the previous case of $p = q$ to determine the ratio $\Pr[S_n' = T_n - c]/\Pr[S_n' = T_n]$ for $n \to \infty$ and any constant c: The X_i' are still independent random variables from at most W different distributions (one for each possible value of w_i in $[W]$); the distribution of X_i' has finite variance and span w_i. Thus, primitivity of the weight sequence implies that the greatest common divisor of the span of all distributions that appear infinitely often as distributions of X_i' is 1, so we can apply Lemma B.1 to get that we can write

$$\Pr[S_n' = x] = \frac{1}{\sigma_n'}\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x - \mu_n'}{\sigma_n'}\right)^2\right) + \xi_n(x)\right)$$

for $\mu_n' = \mathbb{E}[S_n'] = qW^{(n)}$, $\sigma_n' = \sqrt{\kappa_{S_n'}''(s_n^*)}$, and some error $\xi_n(x)$ where $\sup_{x\in\mathbb{Z}} |\xi_n(x)| \to 0$ as $n \to \infty$. By the exact same argument as above, we get that for any constant c,

$$\frac{\Pr[S_n = T_n - c]}{\Pr[S_n = T_n]} \to 1$$

as $n \to \infty$. We get that

$$\lim_{n\to\infty} \frac{\beta_1^p(\mathcal{G}_n)}{\beta_2^p(\mathcal{G}_n)} = \lim_{n\to\infty} \frac{p \sum_{i=w_2}^{w_1+w_2-1} e^{is_n^*} + (1-p) \sum_{i=0}^{w_1-1} e^{is_n^*}}{p \sum_{i=w_2}^{w_1+w_2-1} e^{is_n^*} + \sum_{i=w_1}^{w_2-1} e^{is_n^*} + (1-p) \sum_{i=0}^{w_1-1} e^{is_n^*}}. \quad (3)$$

To simplify notation, we set $r_n = e^{s_n^*}$. Recall from the definition of s_n^* that r_n thus is the unique positive root of $\frac{1}{n}\sum_{i=3}^n (\frac{pw_i r^{w_i}}{1-p+pr^{w_i}} - qw_i)$, which converges to $\sum_{w=1}^C d_w(\frac{pwr^w}{1-p+pr^w} - qw)$ as $n \to \infty$. It is not hard to verify that convergence also holds for the unique positive root so that $r = \lim_{n\to\infty} r_n$.

Plugging into Equation (3), we get that when $r \neq 1$

$$\lim_{n\to\infty} \frac{\beta_1^p(\mathcal{G}_n)}{\beta_2^p(\mathcal{G}_n)} = \frac{pr^{w_2}\frac{1-r^{w_1}}{1-r} + (1-p)\frac{1-r^{w_1}}{1-r}}{pr^{w_2}\frac{1-r^{w_1}}{1-r} + r^{w_1}\frac{1-r^{(w_2-w_1)}}{1-r} + (1-p)\frac{1-r^{w_1}}{1-r}} = \frac{(1-r^{w_1})(1-p+pr^{w_2})}{(1-r^{w_2})(1-p+pr^{w_1})}$$

and when $r = 1$,

$$\lim_{n\to\infty} \frac{\beta_1^p(\mathcal{G}_n)}{\beta_2^p(\mathcal{G}_n)} = \frac{pw_1 + (1-p)w_1}{pw_1 + (w_2 - w_1) + (1-p)w_1} = \frac{w_1}{w_2}.$$

To finish the proof, we note that it holds that $r = 1$ if $p = q$. Since $\kappa_{S_n}'(\ln(r))$ is strictly monotonically increasing in r and p, it follows that $r = 1$ if and only if $p = q$. □

Proof of Corollary 4.3. To depart from the informal term"U-shaped", we will say that a function f is *strictly negatively unimodal* on an interval I if there exists a value $r^* \in I$ such that f is strictly monotonically decreasing on $(i\infty, r^*] \cap I$ and strictly monotonically increasing on $[r^*, \infty) \cap I$. Thus, to prove the corollary it suffices to show that $\lim_{n\to\infty} \beta_i^p(\mathcal{G}_n)/\beta_j^p(\mathcal{G}_n)$ as a function of q is strictly negatively unimodal on $(0, 1)$.

First, note that if $g : I \to J$ is strictly increasing and its image is all of J, then by the chain rule f is strictly negatively unimodal on J if and only if $f \circ g$ is strictly negatively unimodal on I.

As established in the proof of Theorem 4.2, for any fixed $p \in (0, 1)$ the right-hand side of the equation defining r, namely $\kappa'_{S_n}(\ln(r)) = \sum_{w=1}^{W} d_w \frac{pwr^w}{1-p+pr^w}$, is strictly increasing in r on $\mathbb{R}_{\geq 0}$. Thus, the equation $\sum_{w=1}^{W} d_w \frac{pwr^w}{1-p+pr^w} = q \cdot \sum_{w=1}^{W} d_w w$ has a unique solution $r = r(q)$ and $r(q)$ is a strictly increasing function of q. Furthermore, $r : (0, 1) \to \mathbb{R}_{>0}$ has image $\mathbb{R}_{>0}$ since $\lim_{q\to 0} r(q) = 0$ and $\lim_{q\to 1} r(q) = \infty$. Therefore, we get that to prove the theorem, it suffices to show that the function

$$f(r) = \lim_{n\to\infty} \frac{\beta_i^p(\mathcal{G}_n)}{\beta_j^p(\mathcal{G}_n)} = \frac{\left(\sum_{k=1}^{w_i} r^{k-1}\right)(1 - p + pr^{w_j})}{\left(\sum_{k=1}^{w_j} r^{k-1}\right)(1 - p + pr^{w_i})}$$

is strictly negatively unimodal on $\mathbb{R}_{>0}$.

Now, note that we can equivalently write

$$f(r) = \frac{(1 - r^{w_i})(1 - p + pr^{w_j})}{(1 - r^{w_j})(1 - p + pr^{w_i})}$$

with $f(1) = w_i/w_j$ being a smooth discontinuity. In this notation, it becomes evident that we can use the composition of functions trick one more time: Using $g(r) = r^{1/w_i}$, with $g : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ and image $\mathbb{R}_{>0}$, we get that it suffices to show that f is strictly negatively unimodal on $\mathbb{R}_{>0}$ when $w_i = 1$ (for all p and $w_j = w > 1$).

To prove this, we look at its derivative

$$f'(r) = \frac{(p - 1)\sum_{j=1}^{w-1} jr^{j-1} + p\sum_{j=1}^{w-1}(w - j)r^{w+j-1}}{\left((1 - p + pr)\sum_{j=0}^{w-1} r^j\right)^2}.$$

For $r \in \mathbb{R}_{>0}$, the denominator is strictly positive, so the sign of $f'(r)$ is equal to the sign of the numerator. The numerator is a polynomial with a single sign switch in the coefficient series, thus Descartes' rule of signs implies that f' has exactly one positive root, we'll denote it r^*. One can quickly check that $\lim_{r\to 0} f'(r) < 0$ to get that $f'(r) < 0$ when $r \in (0, r^*)$ and $f'(r) > 0$ when $r \in (r^*, \infty)$. This implies that f is strictly negatively unimodal on $\mathbb{R}_{\geq 0}$.

Finally, when $q = p$ it holds that $r = 1$. It is easy to confirm that $f'(1) = (2p - 1)\sum_{j=1}^{w-1} j$. Thus, $f'(1) > 0$ when $p > 1/2$, $f'(1) < 0$ when $p < 1/2$, and $f'(1) = 0$ when $p = 1/2$. Since $r(q)$ is strictly increasing in q, this implies the stated sign of the derivative at $q = p$ from the corollary. □

Proof of Corollary 4.4. As pointed out in the paper, any semivalue in which players are assumed to vote independently is a p-propensity Banzhaf power measure: If players vote independently, they each have a probability p_i of voting yes. Since semivalues are anonymous, all p_i are the same. Thus, any normalized semivalue where players vote independently is a p-propensity Banzhaf power index.

By Corollary 4.3, we know that for any fixed propensity p and primitive, regular weight sequence $\boldsymbol{w}$, the PLT holds for quota $q = p$ and for at most one other quota which we'll denote $\tilde{q}(p, \boldsymbol{w})$ as a function of p and $\boldsymbol{w}$. If we can show that for any p, there exist two weight sequences $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$

(that are regular and primitive) so that $\tilde{q}(p, \boldsymbol{w_1}) \neq \tilde{q}(p, \boldsymbol{w_2})$, it follows that for no q there exists a value of p other than $p = q$ such that the PLT hold with respect to the p-Propensity Banzhaf power index for all $\boldsymbol{w}$ (fulfilling our assumptions).

It remains to show this claim. We consider the two weight sequences $\boldsymbol{w_1} = 1, 2, 1, 1, 2, 1, 1, 2, 1, \ldots$ and $\boldsymbol{w_2} = 1, 2, 2, 1, 2, 2, 1, 2, 2, \ldots$. Assume towards a contradiction that $\tilde{q}(p, \boldsymbol{w_1}) = \tilde{q}(p, \boldsymbol{w_2})$ and denote this value as q. Let r_1 be the unique positive solution to $\frac{1}{3}\frac{pr_1}{1-p+pr_1} + \frac{2}{3}\frac{2pr_1^2}{1-p+pr_1^2} = q(\frac{1}{3} + 2 \cdot \frac{2}{3})$ and let r_2 be the unique positive solution to $\frac{2}{3}\frac{pr_1}{1-p+pr_1} + \frac{1}{3}\frac{2pr_1^2}{1-p+pr_1^2} = q(\frac{2}{3} + 2 \cdot \frac{1}{3})$.

Let's first consider the case $r_1 = r_2$ and denote this value as r. It holds that

$$\frac{\frac{1}{3}\frac{pr}{1-p+pr} + \frac{2}{3}\frac{2pr^2}{1-p+pr^2}}{\frac{5}{3}} = \frac{\frac{2}{3}\frac{pr}{1-p+pr} + \frac{1}{3}\frac{2pr^2}{1-p+pr^2}}{\frac{4}{3}},$$

which implies that

$$2\frac{pr}{1-p+pr} = \frac{2pr^2}{1-p+pr^2}$$

or equivalently

$$(1-p)r(r-1) = 0.$$

We know that $r \neq 0$, so it follows that $r = 1$. However, we know from the proof of Theorem 4.2 that $r = 1$ if and only if $q = p$, a contradiction.

Thus, there remains the case $r_1 \neq r_2$. However, since $q \neq p$, we know that $r_1, r_2 \neq 1$. This implies that there are 3 values $r_1, r_2, 1 > 0$ at which $f(r) = \lim_{n\to\infty} \frac{\beta_i^p(\mathcal{G}_n)}{\beta_j^p(\mathcal{G}_n)}$ takes the value $1/2$. This is a contradiction to f being strictly negatively unimodal on $\mathbb{R}_{\geq 0}$ as shown in the proof of Corollary 4.3. This finishes the proof. □

B.2 Oceanic games

Dubey and Shapley [1979] proved Theorem 4.5 for $p = 1/2$. They also proved a version of Theorem 4.5 for general p for unnormalized Banzhaf powers but do not consider normalized Banzhaf powers for $p \neq 1/2$. Conceptually, our proof of Theorem 4.5 is identical to the proof by Dubey and Shapley [1979] for normalized Banzhaf at $p = 1/2$, with modifications to work for any p.

For the proof, we will rely on a handful of well-known facts about the asymptotic behavior of binomial coefficients, which we restate here for the reader's convenience.

Lemma B.2 (Facts about Binomial Distributions). *Let $p \in (0, 1)$. Define $b(n, p, k) = p^k(1-p)^{n-k}\binom{n}{k}$ to be the probability that a binomial random variable with n trials and success probability p takes on value k. All $O(1)$ and $o(1)$ are in terms of n.*

(1) *It holds that*

$$b(n, p, np + O(1)) = \frac{1 \pm o(1)}{\sqrt{2\pi np(1-p)}}.$$

(2) *For any $s \neq p$ and sequence $(s_n)_{n\in N}$ such that $s_n = sn \pm O(1)$, it holds that*

$$\sum_{k=0}^{s_n} b(n, p, k) = \Theta(b(n, p, s_n)) \text{ if } s < p \text{ and } \quad \sum_{k=s_n}^{n} b(n, p, k) = \Theta(b(n, p, s_n)) \text{ if } s > p.$$

(3) *For any $r < s \leq p$ or $r > s \geq p$ and sequences $(s_n)_{n\in N}$ and $(r_n)_{n\in N}$ such that $s_n = sn \pm O(1)$ $(s_n)_{n\in N}$ and $r_n = rn \pm O(1)$, it holds that*

$$\frac{b(n, p, s_n)}{b(n, p, r_n)} = e^{\Theta(n)}.$$

(4) *For any $r < p < s$ and sequences $(s_n)_{n\in N}$ and $(r_n)_{n\in N}$ such that $s_n = sn \pm O(1)$ $(s_n)_{n\in N}$ and $r_n = rn \pm O(1)$, it holds that*

$$\lim_{n\to\infty} \sum_{k=r_n}^{s_n} b(n,p,k) = 1.$$

Proof. (1) This follows directly from the Central Limit Theorem for binomial distributions, see for example Grinstead and Snell [1998].

(2) Let's first assume $s < p$. For any $k \in \{0, ..., s_n\}$ and n large enough, it holds that

$$b(n,p,k-1) = b(n,p,k)\frac{1-p}{p}\frac{k}{n-k+1} \leq b(n,p,k)\frac{1-p}{p}\frac{s_n}{n-s_n+1}.$$

We denote $c_n = \frac{1-p}{p}\frac{s_n}{n-s_n-1}$ and note that $\lim_{n\to\infty} c_n = \frac{1-p}{p}\frac{s}{1-s} \in (0,1)$. Thus, we get that for n large enough

$$\sum_{k=0}^{s_n} b(n,p,k) \leq b(n,p,s_n)\sum_{k=0}^{s_n}(c_n)^k \leq b(n,p,s_n)\frac{1}{1-c_n} = O(b(n,p,s_n)).$$

The case $s > p$ can be proved analogously.

(3) By Stirling's formula, we get that

$$\frac{b(n,p,s_n)}{b(n,p,r_n)} = (1+o(1))\sqrt{\frac{r(1-r)}{s(1-s)}}e^{n(D(r\|p)-D(s\|p))} = e^{\Theta(n)},$$

where $D(r\|p) - D(s\|p) > 0$ since $r < s \leq p$ or $r > s \geq p$.

(4) From part (2), we know that

$$\lim_{n\to\infty} \sum_{k=0}^{r_n-1} b(n,p,k) = \lim_{n\to\infty} O(b(n,p,r_n) = 0,$$

$$\lim_{n\to\infty} \sum_{k=s_n+1}^{n} b(n,p,k) = \lim_{n\to\infty} O(b(n,p,s_n)) = 0.$$

Since $\sum_{k=0}^{n} b(n,p,k) = 1$, the claim follows.

□

Proof of Theorem 4.5. We'll first calculate the unnormalized propensity Banzhaf powers of the players. We then examine whether the total power of the large players, $\sum_{i\in[\ell]} \hat{\beta}_i^p(\mathcal{G}_n)$, or the total power of the small players, $\sum_{i=\ell+n}^{\ell+n} \hat{\beta}_i^p(\mathcal{G}_n)$, dominates as n goes to infinity. In the cases where the large player power dominates, we'll then solve for the relative powers of the players.

Let $S \subseteq [\ell]$ be any subset of large players. We let $r_S = \left\lfloor \frac{(q-\sum_{i\in S} w_i)n}{\alpha} \right\rfloor$, so that a coalition of S and r_S small players is losing but a coalition of S and $r_S + 1$ small players is winning. Thus, the unnormalized Banzhaf powers of the large players are

$$\begin{aligned}\hat{\beta}_i^p(\mathcal{G}_n) &= \sum_{S\subseteq[\ell]\setminus\{i\}} \sum_{k=r_{S\cup\{i\}}+1}^{r_S} \binom{n}{k} p^{k+|S|}(1-p)^{(n+\ell-1)-(k+|S|)} \\ &= \sum_{S\subseteq[\ell]\setminus\{i\}} \sum_{k=r_{S\cup\{i\}}+1}^{r_S} b(n,p,k)p^{|S|}(1-p)^{(\ell-1)-|S|},\end{aligned}$$

where $b(n,p,k) = p^k(1-p)^{n-k}\binom{n}{k}$ and we define binomial coefficients with negative bottom entries as 0. We let $\hat{\beta}_{\text{large}}(\mathcal{G}_n) = \sum_{i\in[\ell]} \hat{\beta}_i^p(\mathcal{G}_n)$ be the total power of large players. Similarly, we can get the total power of the small players as

$$\begin{aligned}\hat{\beta}_{\text{small}}(\mathcal{G}_n) = \sum_{i=\ell+1}^{\ell+n} \hat{\beta}_i^p(\mathcal{G}_n) &= n \sum_{S\subseteq[\ell]} \binom{n-1}{r_S} p^{r_S+|S|}(1-p)^{(n+\ell-1)-(r_S+|S|)} \\ &= \sum_{S\subseteq[\ell]} (n-r_S)\binom{n}{r_S} p^{r_S+|S|}(1-p)^{(n+\ell-1)-(r_S+|S|)} \\ &= \sum_{S\subseteq[\ell]} (n-r_S) b(n,p,r_S) p^{|S|}(1-p)^{(\ell-1)-|S|}.\end{aligned}$$

Let's now examine how the ratio of the total power of large players to the total power of small players, $\hat{\beta}_{\text{large}}(\mathcal{G}_n)/\hat{\beta}_{\text{small}}(\mathcal{G}_n)$, is going to behave in the limit:

- First, assume there exists a coalition S^* such that $\lim_{n\to\infty} \frac{r_{S^*}}{n} = p$. This case applies precisely if there exists a coalition $S^* \subseteq [\ell]$ such that $p = \lim_{n\to\infty} r_{S^*}/n = (q - \sum_{i\in S^*})/\alpha$, i.e., $q = \alpha p + \sum_{i\in S^*} w_i$, or equivalently, $q \in P$. By Lemma B.2, part (1), there thus exists a term of the form $b(n,p,np+O(1))$ in the sum of the total powers of the small powers. We thus know that $\hat{\beta}_{\text{small}}(\mathcal{G}_n) \geq (n-r_s)\Theta(\frac{1}{\sqrt{n}}) = \Theta(\sqrt{n})$. However, $\hat{\beta}_i(\mathcal{G}_n) \leq 1$ for all i, so $\hat{\beta}_{\text{large}}(\mathcal{G}_n) \leq \ell$. It follows that $\hat{\beta}_{\text{large}}(\mathcal{G}_n)/\hat{\beta}_{\text{small}}(\mathcal{G}_n) \to 0$ and therefore $\lim_{n\to\infty} \beta_i(\mathcal{G}_n) \to 0$ for all $i \in [\ell]$.
- Next, we consider the case of the quota q being such that $\lim_{n\to\infty} r_\emptyset/n < p$ (thus $\lim_{n\to\infty} r_S/n < p$ for all S) or $\lim_{n\to\infty} r_{[\ell]}/n > p$ (thus $\lim_{n\to\infty} r_S/n > p$ for all S). This case applies precisely when $q < p\alpha$ or $q > \alpha p + W = 1-\alpha(1-p)$. Let's focus on $\lim_{n\to\infty} r_\emptyset/n < p$ first. By Lemma B.2, part (2), we know that $\hat{\beta}_{\text{large}}(\mathcal{G}_n) = \Theta(b(n,p,r_\emptyset))$ and $\hat{\beta}_{\text{small}}(\mathcal{G}_n) = \Theta(nb(n,p,r_\emptyset))$. It follows that $\hat{\beta}_{\text{large}}(\mathcal{G}_n)/\hat{\beta}_{\text{small}}(\mathcal{G}_n) \to 0$ and therefore $\lim_{n\to\infty} \beta_i(\mathcal{G}_n) \to 0$ for all $i \in [\ell]$. The case $\lim_{n\to\infty} r_{[\ell]}/n > p$ is analogous.
- There remains the case $q \in [\alpha p, 1-\alpha(1-p)] \setminus P$. Note that due to $S = \emptyset$ and $S = [\ell]$, αp and $1-\alpha(1-p)]$ are always in P, so we may exclude these values from the interval. By definition of P, we know that in this case no binomial coefficient $\binom{n}{r_S}$ with $\lim_{n\to\infty} r_S/n = p$ appears in $\hat{\beta}_{\text{small}}(\mathcal{G}_n)$. Thus, we know by Lemma B.2, part (2), that $\hat{\beta}_{\text{small}}(\mathcal{G}_n) = O(nb(n,p,s_n) + nb(n,p,s'_n))$ for sequences $(s_n)_{n\in\mathbb{N}}$ and $(s'_n)_{n\in\mathbb{N}}$ with $s_n = sn + O(1)$ and $s'_n = s'n + O(1)$ so that $s < p < s'$.
  In contrast, we know that $\binom{n}{\lceil np\rceil}$ has to appear in some $\hat{\beta}_{\text{large}}(\mathcal{G}_n)$. In particular, assume WLOG that $w_\ell \geq w_1, \ldots, w_{\ell-1}$. For some $j \in [\ell-1]$, it holds that $\sum_{i\in[j]} w_i + \alpha p < q$ but $\sum_{i\in[j+1]} w_i + \alpha p > q$ (equaling q is not possible since we know $q \notin P$). Since w_ℓ is the largest weight, it follows that $\sum_{i\in[j]} w_i + \alpha p \leq q$ but $\sum_{i\in[j]\cup\{\ell\}} w_i + \alpha p > q$. Therefore, $p \in [\lim_{n\to\infty} r_{[j]\cup\{\ell\}}/n, \lim_{n\to\infty} r_{[j]}/n]$. This implies that $\hat{\beta}_{\text{large}}(\mathcal{G}_n) = \Omega(b(n,p,\lceil np\rceil))$.
  By Lemma B.2, part (3), it follows that $\hat{\beta}_{\text{large}}(\mathcal{G}_n)/\hat{\beta}_{\text{small}}(\mathcal{G}_n) \geq e^{\Theta(n)}/n \to \infty$ as $n \to \infty$.

What remains is to determine $\beta_i(\mathcal{G}_n)$ in the third case. By Lemma B.2, part (4), we know that

$$\lim_{n\to\infty} \sum_{k=r_{S\cup\{i\}}+1}^{r_S} b(n,p,k) = \begin{cases} 1 & \text{if } \lim_{n\to\infty} r_{S\cup\{i\}}/n < p < \lim_{n\to\infty} r_S/n \\ 0 & \text{if } \lim_{n\to\infty} r_{S\cup\{i\}}/n > p \text{ or } p > \lim_{n\to\infty} r_S/n \end{cases}.$$

Since we know that in this case $\lim_{n\to\infty} r_S/n \neq p$ for all S, we don't need to be concerned about the other cases. Now, observe that for any coalition S, $r_S/n < p$ implies $\sum_{j\in S} w_i > q - \alpha p$ while $\lim_{n\to\infty} r_S/n > p$ implies $\sum_{j\in S} w_i < q - \alpha p$. Thus, the above expression converges to 1 exactly if

coalition S is less than a threshold but coalition $S \cup \{i\}$ is exceeding this threshold—this is exactly definition of i being pivotal for a coalition S. We can write

$$\hat{\beta}_i^p(\mathcal{G}_n) = \sum_{S \subseteq [\ell] \setminus \{i\}} \mathrm{piv}_{\mathcal{G}_p^{(0)}}(i, S)\, p^{|S|}(1-p)^{(\ell-1)-|S|} = \hat{\beta}_i^p(\mathcal{G}_p^{(0)}),$$

where $\mathcal{G}_p^{(0)} = (\frac{w_1}{W}, ..., \frac{w_\ell}{W}; \frac{q-\alpha p}{W})$ as defined in the theorem statement. □

B.3 The Emergence of Veto Players

Proof of Theorem 5.1. Following the proof outline from Section 5, we first show how to construct, for any $q \in (0,1), p \in (0,1)$, a family of WVGs $\{\mathcal{G}_n\}$ where $\mathcal{G}_n = (\boldsymbol{w^n}; q)$ so that $w_1^n > 1-q$ for all n (i.e., player 1 is a veto player in $\mathcal{G}_n$) but $\lim_{n\to\infty} \beta_1^p(\mathcal{G}_n) = 0$.

We start by considering the cases where $q < \frac{1+p}{2}$. We let

$$\mathcal{G}_n = (w_1, w_2, \underbrace{\tfrac{\alpha}{n}, \ldots, \tfrac{\alpha}{n}}_{n}; q),$$

where $w_1 = 1/2$, $\alpha = \frac{q-1/2}{p}$, and $w_2 = 1 - w_1 - \alpha$. Note that the WVG is well-defined: By definition $w_1 + w_2 + \alpha = 1$; furthermore, $\alpha \in (0,1)$ since $\frac{1}{2} < q < \frac{1+p}{2}$ and $w_2 \in (0,1)$ since $w_1 + \alpha = \frac{p/2+q-1/2}{p} < \frac{p/2+p/2}{p} = 1$. As desired, player 1 is a veto player since $w_1 > 1-q$. Since $q = \alpha p + w_1$, we get by Theorem 4.5 that $\lim_{n\to\infty} \beta_1^p(\mathcal{G}_n) = 0$.

Next, consider the remaining cases $q \geq \frac{1+p}{2}$. Note that for $p \in (0,1)$, $q \in (1/2, 1)$, this implies $q > p$. We let

$$\mathcal{G}_n = (w_1, \underbrace{\tfrac{\alpha}{n}, \ldots, \tfrac{\alpha}{n}}_{n}; q),$$

where $w_1 = \frac{q-p}{1-p}$ and $\alpha = \frac{1-q}{1-p}$. Note that the WVG is well-defined: $w_1 + \alpha = 1$, and $w_1 > 0$ since $q > p$ and $\alpha > 0$ since $q < 1$. Furthermore, we get that $w_1 > 1-q$, since $\frac{q-p}{1-p} > 1-q \Leftrightarrow \frac{2q-1}{q} > p$, which follows from $q < 1$ and $p < 2q-1$, by the case assumption. Since $q = \alpha p + w_1$, we get by Theorem 4.5 that $\lim_{n\to\infty} \beta_1^p(\mathcal{G}_n) = 0$.

Now, for $\{\mathcal{G}_n\}$ chosen according to which case of p and q we are in, let $\boldsymbol{m^n} = \boldsymbol{\beta}^p(\mathcal{G}_n)$ and let $\boldsymbol{w^n}$ be the weights in $\mathcal{G}_n$. We know, by construction, that

$$\mathrm{discr}_{\boldsymbol{m^n}, q, \boldsymbol{\beta}^p}(\boldsymbol{w^n}) = \left\| \boldsymbol{m} - \boldsymbol{\beta}^p((\boldsymbol{w^n}; q)) \right\|_1 = \|\boldsymbol{m} - \boldsymbol{m}\|_1 = 0.$$

Thus, $\boldsymbol{w^n} \in W^*(\boldsymbol{m^n}, q, \boldsymbol{\beta}^p)$. We get that

$$\text{veto-dist}_q(\boldsymbol{\beta}^p) \geq \text{veto-dist}(\boldsymbol{m}^n, q, \boldsymbol{\beta}^p) \geq \frac{1-q}{m_1^n}.$$

Since $\lim_{n\to\infty} m_i^n = \lim_{n\to\infty} \beta_1^p(\mathcal{G}_n) = 0$, it follows that $\text{veto-dist}_q(\boldsymbol{\beta}^p)$ is unbounded. □

Proof of Theorem 5.2. A player is called a *dictator* if any coalition they are in is winning, that is if $w_i \geq q$ (for a non-strict quota). Since $w_i \geq 1 - q = 1/2$ implies $w_i \geq q$, any veto player is also a dictator. If a player is a dictator and a veto player, every coalition they are in is winning and any coalition they are not a part of is losing. Thus, they are pivotal for any coalition without them, while no other player is pivotal for any coalition. Thus, in this setting, if there exists a veto player, this player has power 1 while all other players have power 0.

WLOG, assume the population target $\boldsymbol{m}$ is sorted from largest to smallest. A weight vector $\boldsymbol{w}$ with player i as a veto player will lead to discrepancy $2(1 - m_i)$, so any weight vector with a veto player has discrepancy at least $2(1 - m_1)$. In contrast, the weight vector $\boldsymbol{w} = (1/2, 1/2, 0, ...)$ does

not induce a veto player and leads to powers $\boldsymbol{\beta}^p = (1/2, 1/2, 0, \ldots)$. If $m_1 < \frac{1}{2}$, i.e., player 1 is not a deserving veto player, the discrepancy of $\boldsymbol{w} = (1/2, 1/2, 0, \ldots)$ is exactly $2(1 - m_1 - m_2)$, so strictly less than the discrepancy achievable with a veto player (since $m_2 > 0$ if $m_1 < 1$). If $m_1 \geq \frac{1}{2}$, player 1 is a deserving veto player, so the discrepancy of any weight vector with a veto player is at least $2(1 - m_2)$. The discrepancy of $\boldsymbol{w} = (1/2, 1/2, 0, \ldots)$ is exactly $2(1/2 - m_2)$, so strictly less than the discrepancy achievable with a veto player.

It follows that in no case a weight vector inducing an undeserving veto player will be optimal. Thus, the veto distortion is 1. □

C Full Empirical Results

C.1 Ontario weight table

Table 1. **Weighted voting under the Banzhaf power index.** Towns, populations, optimized weights, and powers for ordinary Banzhaf (propensity 1/2).

Fixed P	Town name	Pop. share	$T = 1/2$		$T = 3/5$		$T = 2/3$		$T = 3/4$	
			weight	power	**weight**	power	**weight**	power	**weight**	power
Town 1	Town of South Bristol	0.01459	**1487**	0.01459	**1484**	0.01460	**1438**	0.01459	**1229**	0.01457
Town 2	Town of Canadice	0.01483	**1511**	0.01484	**1510**	0.01485	**1460**	0.01483	**1247**	0.01482
Town 3	Town of Bristol	0.02031	**2105**	0.02031	**2063**	0.02031	**1999**	0.02031	**1711**	0.02032
Town 4	Town of Naples	0.02137	**2210**	0.02136	**2172**	0.02137	**2104**	0.02137	**1802**	0.02137
Town 5	Town of Seneca	0.02351	**2421**	0.02351	**2388**	0.02351	**2313**	0.02351	**1987**	0.02352
Town 6	Town of West Bloomfield	0.02437	**2505**	0.02436	**2476**	0.02436	**2397**	0.02436	**2058**	0.02437
Town 7	Town of Richmond	0.02988	**3123**	0.02988	**3029**	0.02988	**2939**	0.02988	**2522**	0.02987
Town 8	Town of Geneva	0.03088	**3203**	0.03087	**3132**	0.03088	**3042**	0.03088	**2607**	0.03088
Town 9	Town of East Bloomfield	0.03237	**3340**	0.03237	**3282**	0.03237	**3192**	0.03237	**2732**	0.03236
Town 10	City of Geneva (5,6)	0.03272	**3372**	0.03272	**3320**	0.03272	**3223**	0.03271	**2760**	0.03273
Town 11	City of Geneva (3,4)	0.03487	**3582**	0.03487	**3534**	0.03487	**3435**	0.03487	**2939**	0.03487
Town 12	Town of Hopewell	0.03496	**3588**	0.03495	**3544**	0.03496	**3442**	0.03495	**2948**	0.03496
Town 13	Town of Gorham	0.03651	**3741**	0.03651	**3698**	0.03651	**3599**	0.03652	**3080**	0.03652
Town 14	City of Canandaigua (2,3)	0.04571	**4684**	0.04572	**4625**	0.04571	**4505**	0.04571	**3855**	0.04572
Town 15	City of Geneva (1,2)	0.04633	**4741**	0.04634	**4686**	0.04632	**4565**	0.04634	**3903**	0.04634
Town 16	City of Canandaigua (1,4)	0.04834	**4932**	0.04834	**4892**	0.04833	**4767**	0.04834	**4076**	0.04835
Town 17	Town of Phelps	0.05902	**6001**	0.05901	**5948**	0.05901	**5823**	0.05902	**4985**	0.05902
Town 18	Town of Manchester	0.08362	**8381**	0.08363	**8356**	0.08362	**8318**	0.08362	**7045**	0.08362
Town 19	Town of Canandaigua	0.09879	**9750**	0.09879	**9829**	0.09878	**9808**	0.09878	**8339**	0.09878
Town 20	Town of Farmington	0.12600	**12093**	0.12600	**12320**	0.12599	**12710**	0.12600	**11576**	0.12600
Town 21	Town of Victor	0.14103	**13230**	0.14103	**13712**	0.14102	**14921**	0.14104	**26599**	0.14103
SUM			**100,000**	1	**100,000**	1	**100,000**	1	**100,000**	1
L^1 error			0.00013		0.00011		0.00008		0.00012	

C.2 Convergence of heuristic algorithm

See Figure 8 for the convergence plot as discussed in Section 6.2.

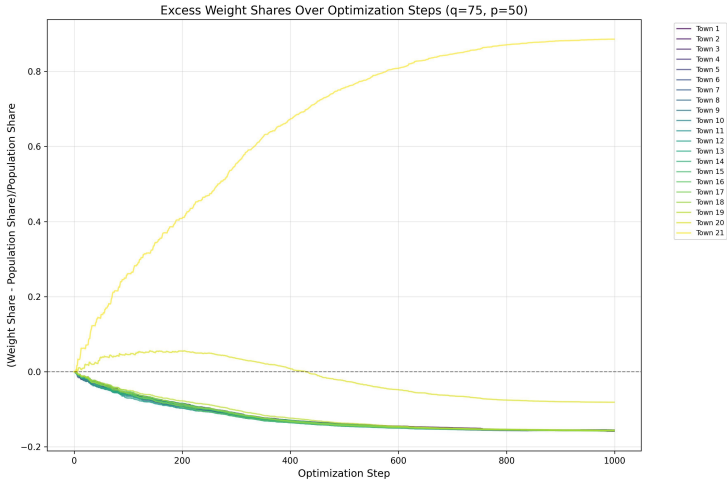


Fig. 8. Weights of each of the 21 towns/wards in Ontario County over the course of a heuristic optimization run. Only the steps where an improvement was made are shown. The largest town, Victor (Town 21), ended up with a weight that was inflated beyond its population share by an extra 88.6%.

C.3 Optimal Weights for Hamilton County

All the optimized weights used in Section 6 were obtained using the random local search algorithm described. To ensure that the observed behavior is not due to this algorithm getting stuck in a local optimum, we use an algorithm due to Kurz [2012] to solve for the exact (globally) optimal weights.

The algorithm is based on modeling the inverse power problem as a mixed integer linear program. For the Banzhaf power index, we can check for any fixed $a \in \mathbb{Q}$ whether weights with discrepancy less than a to a given target power distribution exist. Thus, with binary search on a, we can identify weights giving near-optimal discrepancy up to an arbitrarily small margin. It is not difficult to adapt the MILPs to correspond to the p-propensity Banzhaf power index or the adaptive Banzhaf power index. It is furthermore straightforward to impose as a constraint that no undeserving veto player may be induced by the optimal weights.

Unfortunately, the number of binary optimization variables used in the MILP increases exponentially with the number of players. Thus, this approach only works well for small n. In particular, we find that within reasonable time, $n \approx 12$ are the largest instances for which we can determine exactly optimal weights, while $n \approx 14$ are the largest instances for which we can solve a single MILP to check for a single discrepancy whether it is achievable.

Luckily, there is one county in New York State with only 9 Towns: Hamilton County. The populations of the towns are 92, 221, 292, 355, 413, 683, 791, 897, and 1363. For quota 75%, we plot the heuristically optimized weights and the globally optimal weights in Figure 9. The heuristically optimized weights give a discrepancy of 0.0287, while truly optimal weights achieve a discrepancy of 0.0237. Most importantly, not only are the discrepancies close, but the weights themselves are close. In particular, optimal weights over-weight Indian Lake as predicted, with the algorithm outputs actually underestimating the magnitude.

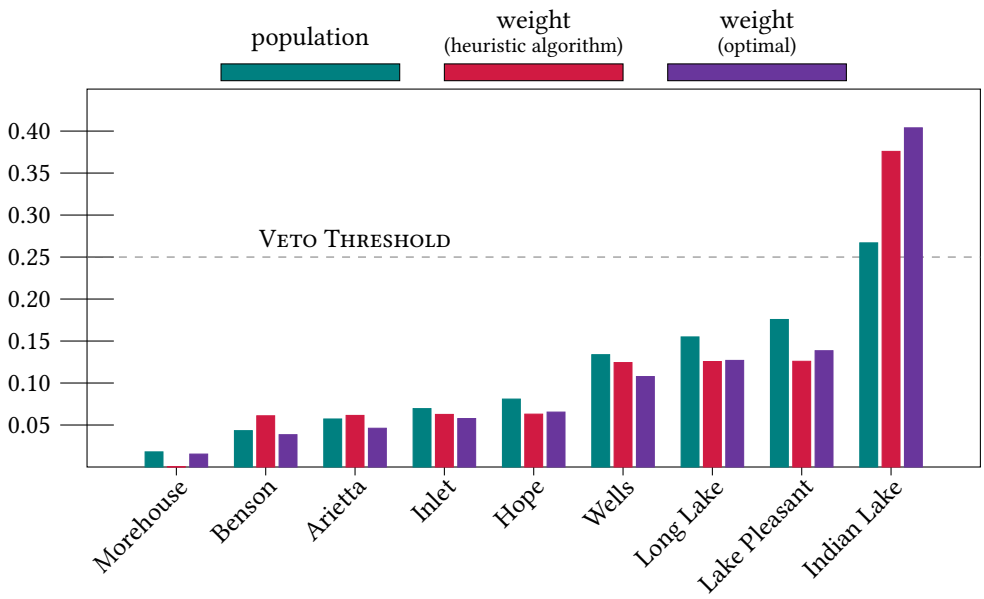


Fig. 9. The weights found by the randomized optimization algorithm and the optimal MILP algorithm under the Banzhaf power index for Hamilton County at quota 75%.

We run the same experiment with adaptive Banzhaf power. For quota 75%, we plot the weights the optimization algorithm found and the optimal weights in Figure 10. The heuristic achieved population-power discrepancy of 0.06427, while the global minimum is 0.06184. And again, the heuristic correctly tracks the key feature: $\boldsymbol{m} \approx \boldsymbol{w} \approx \boldsymbol{\rho}$.

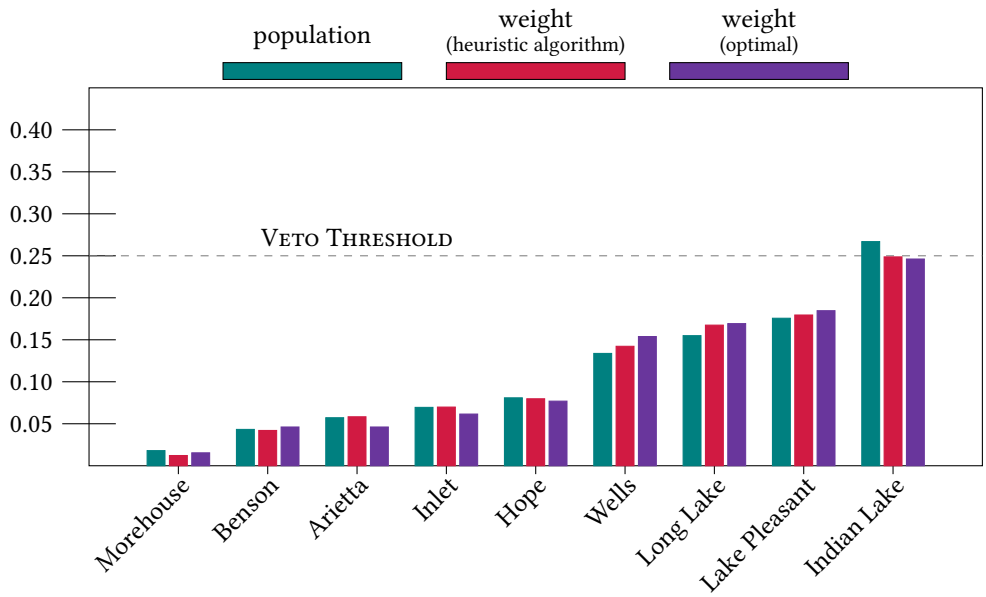


Fig. 10. The weights found by the randomized optimization algorithm and the optimal MILP algorithm under the adaptive Banzhaf power index for Hamilton County at quota 75%.

C.4 Additional Heatmaps

Here we show all plots of the excess weight of the largest player and L^1 distortions for Ontario County, Livingston County, and four synthetic instances:

Linear. 21 towns with evenly-spaced populations.

Big-Small-1. 2 large towns with equal populations, and 19 small towns with equal populations.

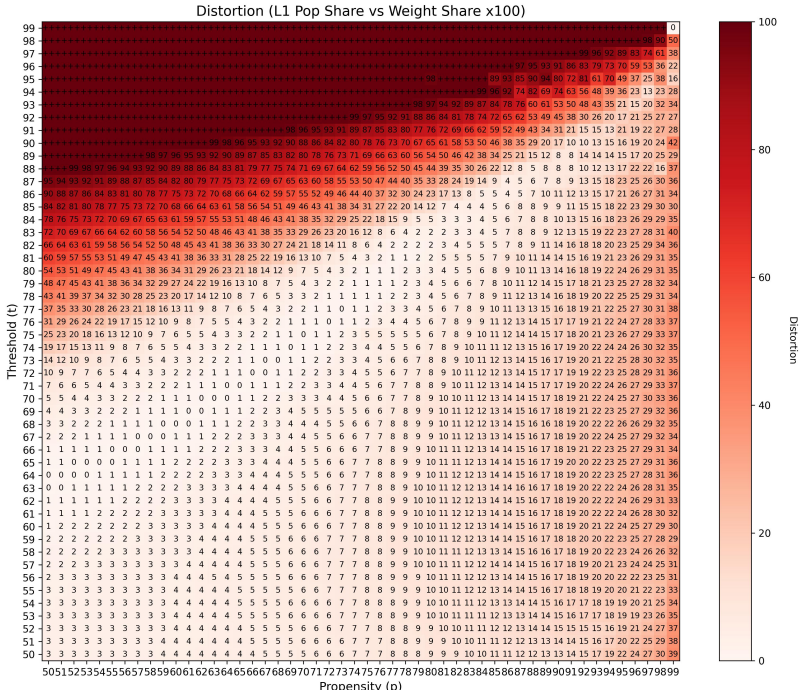
Big-Small-2. 2 large towns with slightly different populations, and 19 small towns with slightly different populations.

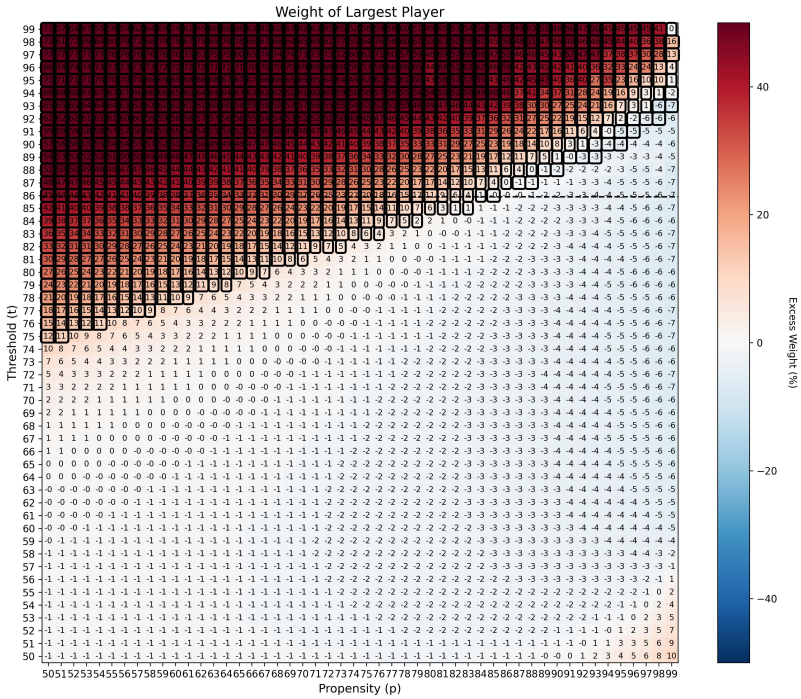
Anti-Ontario. The largest and smallest population in Ontario sum to 17,501. Anti-Ontario is constructed by subtracting all their populations from that total, so that there are two smallest towns and a large number of nearly-equal large ones.

Note that for the distortion plots, a "++" indicates a distortion of 100% or greater.

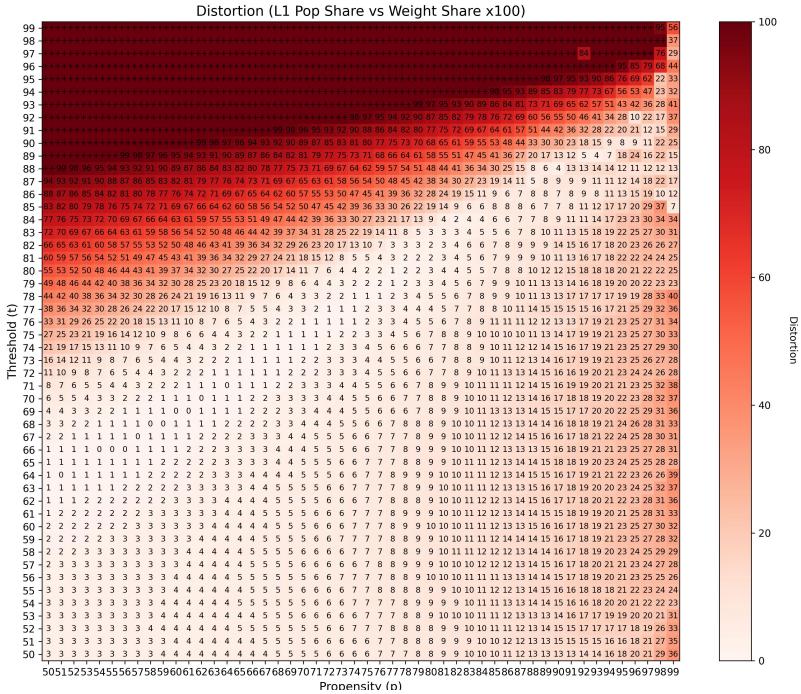
We believe that the two "lines" of low distortion, that are very clear in Big-Small-1 and slightly diluted but still clearly visible in Big-Small-2, are especially noteworthy: They align very well with the pitfall points predicted by Theorem 4.5.

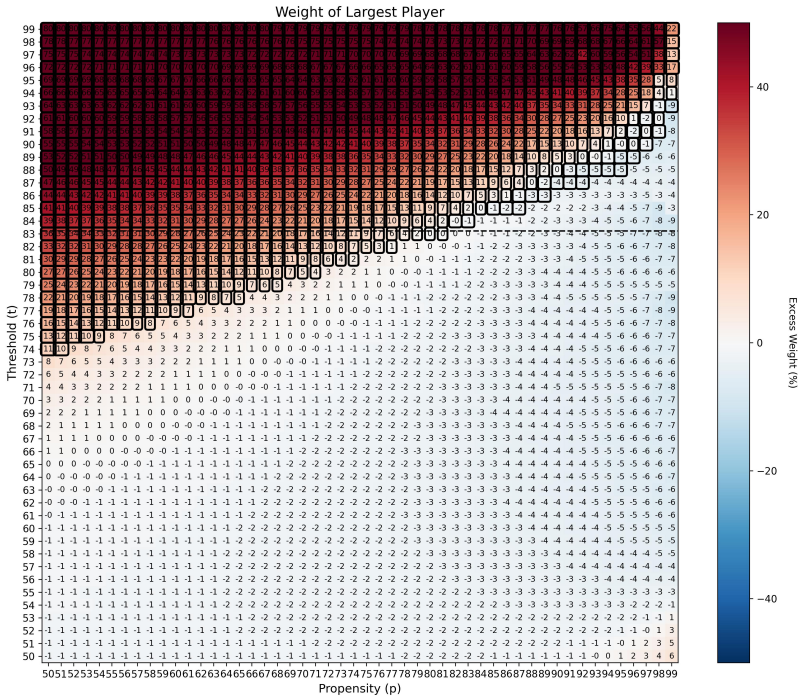
C.4.1 Ontario County. Populations: [1641, 1668, 2284, 2403, 2644, 2740, 3360, 3473, 3640, 3679, 3921, 3931, 4106, 5140, 5210, 5436, 6637, 9404, 11109, 14170, 15860]



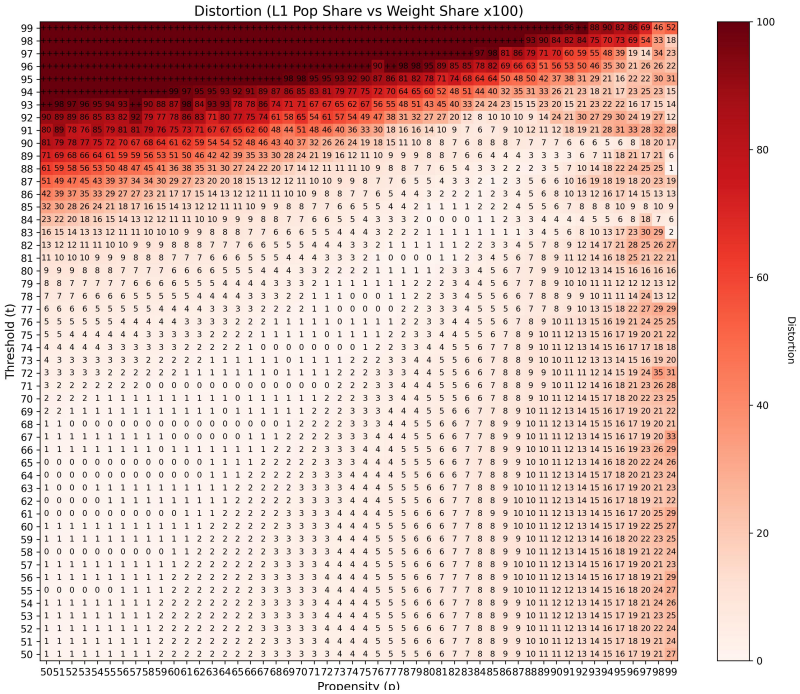


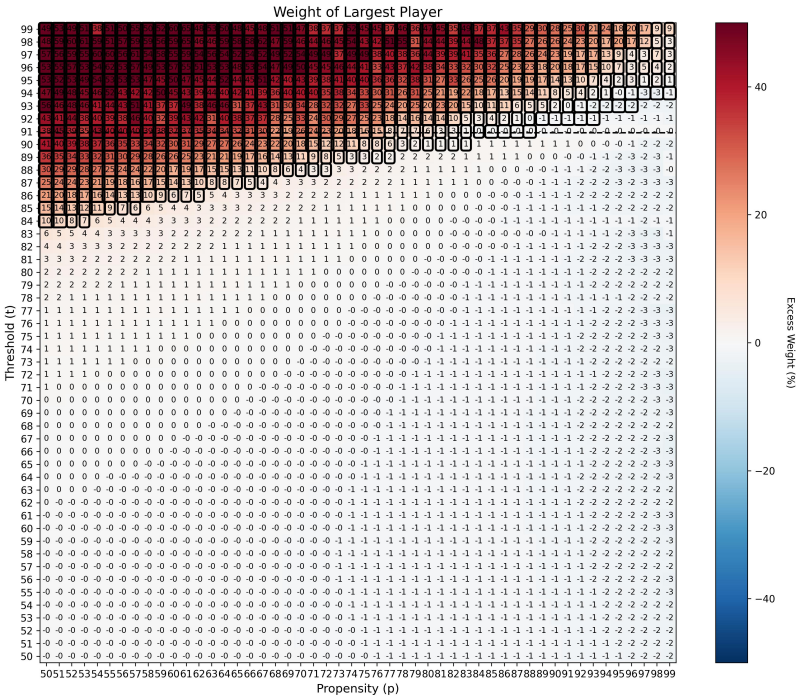
C.4.2 Livingston County. Populations: [725, 765, 1157, 1464, 1583, 2087, 2292, 2322, 2695, 3187, 4156, 4158, 4452, 5341, 6945, 7508, 10242]



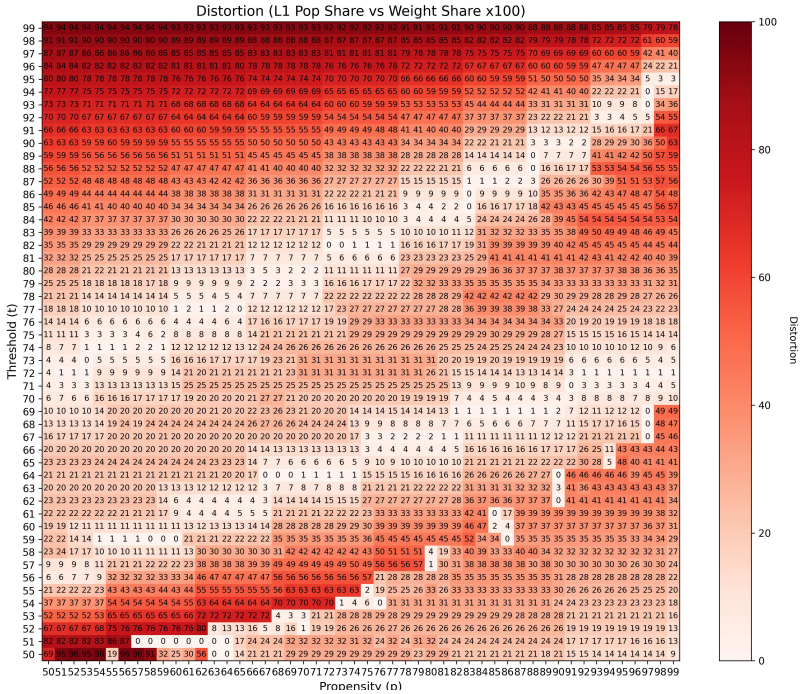


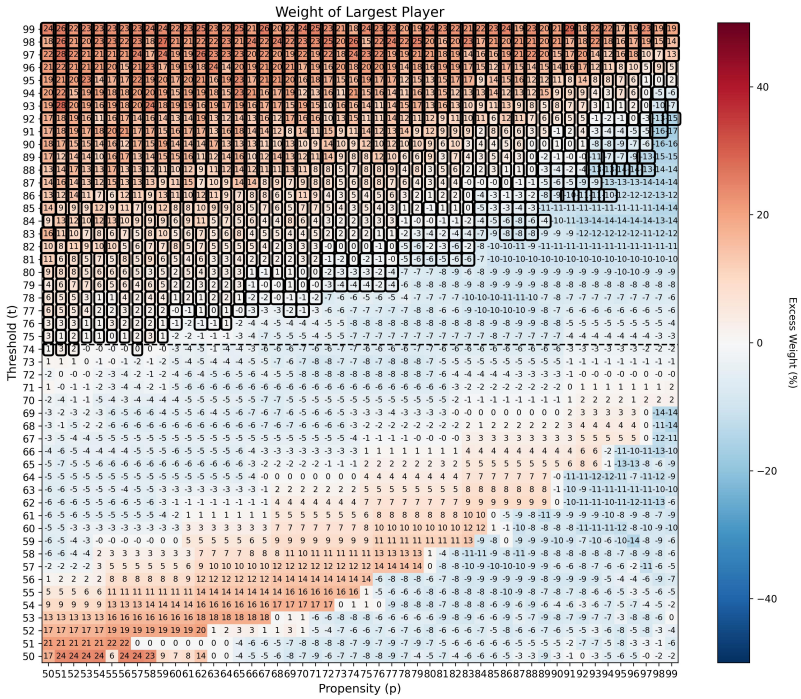
C.4.3 Linear. Populations: [1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 11000, 12000, 13000, 14000, 15000, 16000, 17000, 18000, 19000, 20000, 21000]



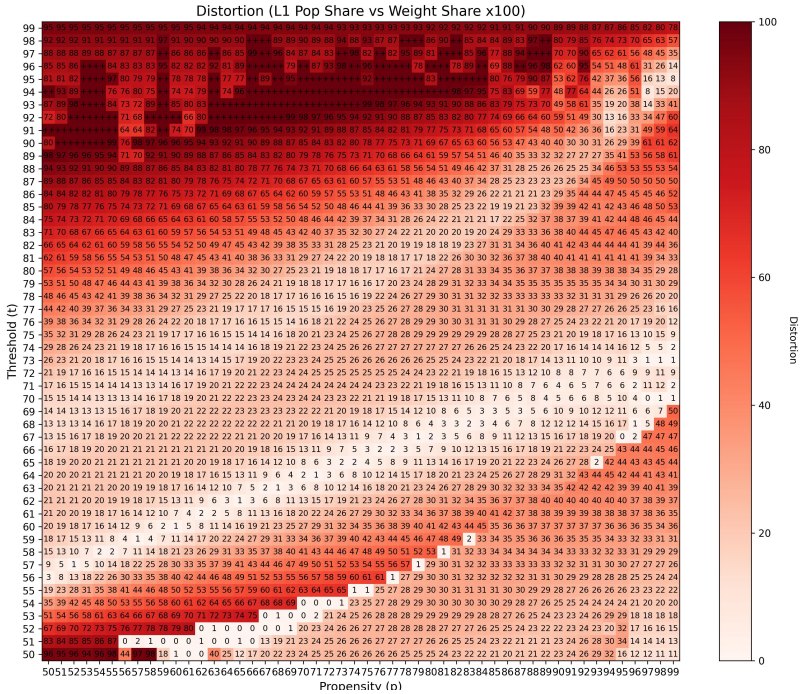


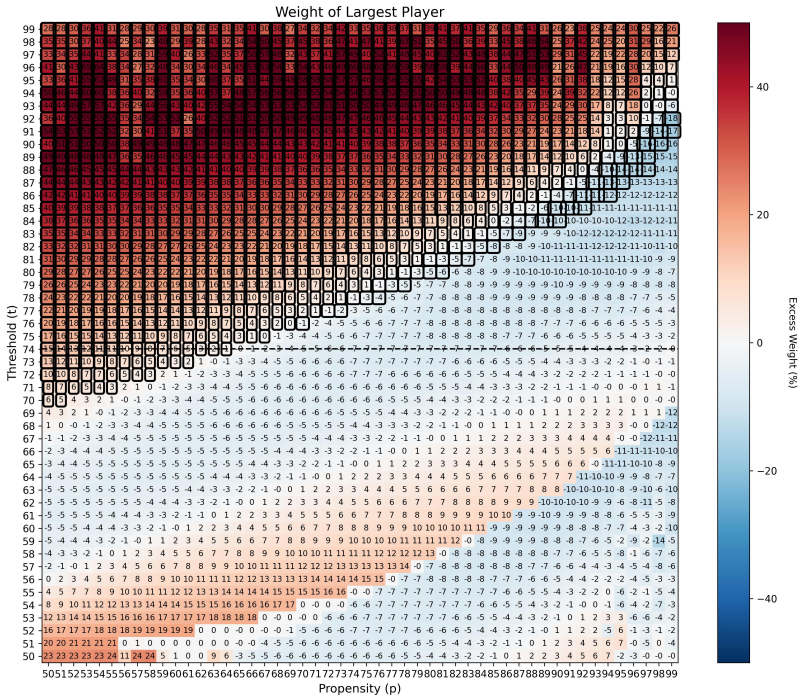
C.4.4 Big-Small-1. Populations: [2000, 20000]



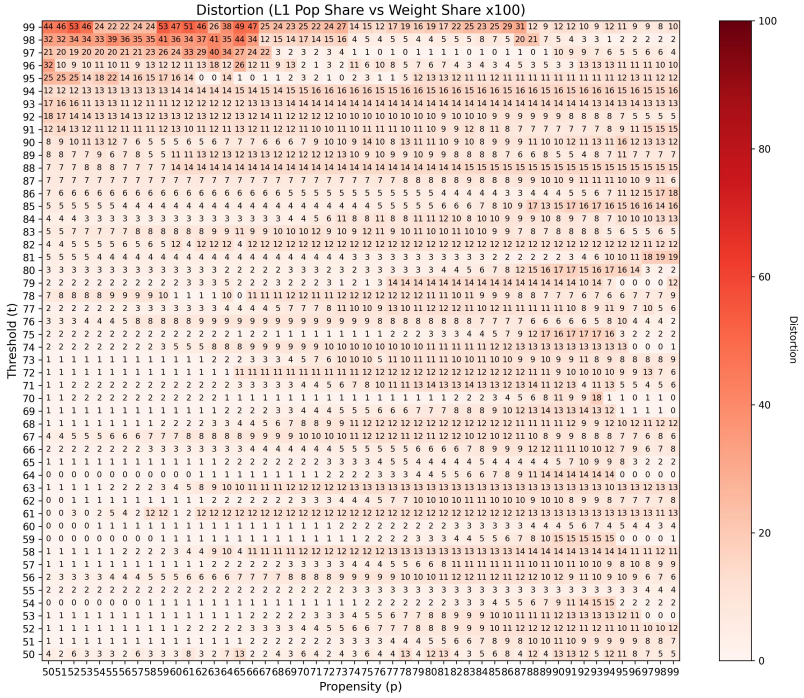


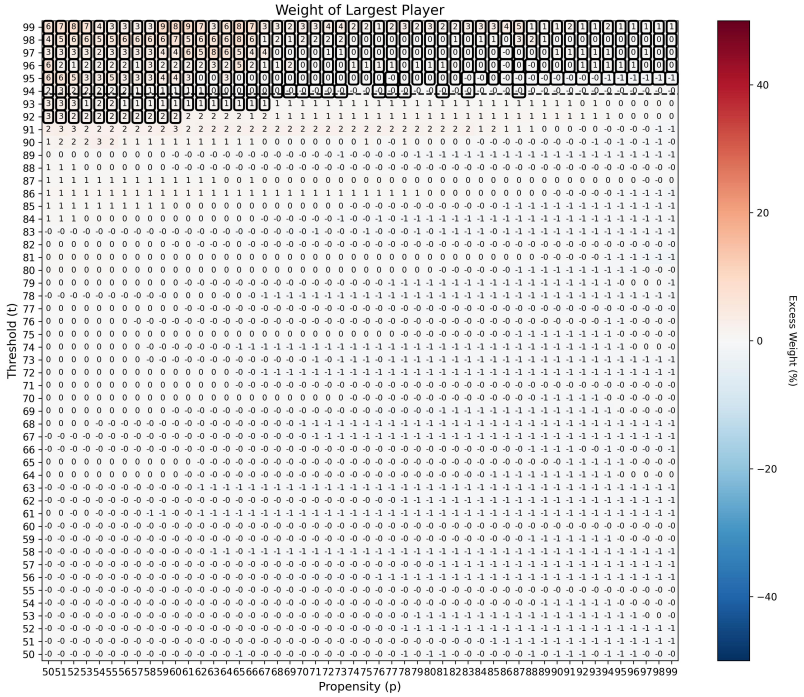
C.4.5 Big-Small-2. Populations: [1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2700, 2800, 2900, 19000, 20000]





C.4.6 Anti-Ontario. Populations: [1641, 3331, 6392, 8097, 10864, 12065, 12291, 12361, 13395, 13570, 13580, 13822, 13861, 14028, 14141, 14761, 14857, 15098, 15217, 15833, 15860]





C.5 Executive directors of the International Monetary Fund

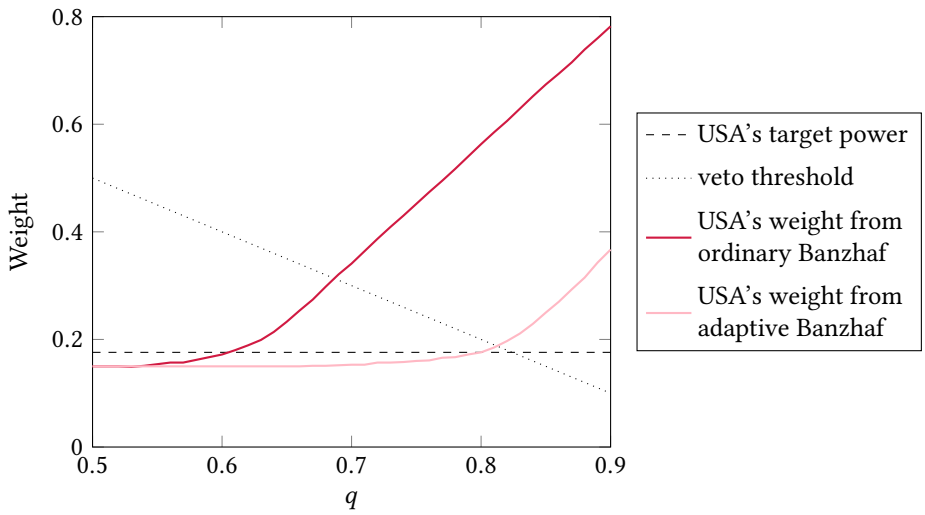


Fig. 11. Leech [2002c] observed that in the IMF weighted voting game explained above, the United States receives disproportionally much weight at high quotas. At the quota of 85%, which is actually being used by the IMF, the US would need to receive over 60% of the weight to get their fair share of . Here we see that the use of adaptive Banzhaf power mitigates the problem significantly.