Chapter 5 Empirical Analysis of Voting Rules and Election Paradoxes

This chapter details work on empirically testing the rates of occurrence of many of the paradoxes and issues that arise when using different voting rules. This work includes surveying existing datasets and empirical verification studies; the identification and mining of a novel dataset for theory testing; and extensive statistical analysis of the new dataset. The study of voting systems often takes place in the theoretical domain due to a lack of large samples of sincere, strictly ordered voting data. We derive several million elections (more than all the existing studies combined) from publicly available data: the Netflix Prize dataset [10]. The Netflix data is derived from millions of Netflix users who have an incentive to report sincere preferences, unlike random survey takers. We evaluate each of these elections under the Plurality, Borda, k-Approval, and Repeated Alternative Vote (RAV) voting rules. We examine the Condorcet Efficiency of each of the rules and the probability of occurrence of Condorcet's Paradox. We compare our votes to existing theories of domain restriction (e.g., single-peakedness) and statistical models used to generate election data for testing (e.g., Impartial Culture). We find a high consensus among the different voting rules; almost no instances of Condorcet's Paradox; almost no support for restricted preference profiles, and very little support for many of the statistical models currently used to generate election data for testing. Portions of this work have been previously published in refereed conference proceedings [86]. However, while the overall analysis is the same, this chapter details a greatly extended version of the work including several hundred million more elections than reported in the initial publication.

5.1 Motivation

As we have seen, voting rules and social choice methods have been used for centuries in order to make group decisions. Increasingly, in computer science, data collection and reasoning systems are moving towards distributed and multi-agent design paradigms [96]. With this design shift comes the need to aggregate the (possibly disjoint) observations and preferences of individual agents into a total ordering in order to synthesize knowledge and data.

One of the most common methods of preference aggregation and group decision making in human systems is voting. Many societies, both throughout history and across the planet, use voting to arrive at group decisions on a range of topics from deciding what to have for dinner to declaring war. Unfortunately, results in the field of social choice prove that there is no perfect voting system and, in fact, voting systems can succumb to a host of problems. Arrow's Theorem demonstrates that any preference aggregation scheme for three or more alternatives will fail to meet a set of simple fairness conditions [2]. Each voting method violates one or more properties that most would consider important for a voting rule (such as non-dictatorship) [61]. Questions about voting and preference aggregation have circulated in the math and social choice communities for centuries [3,27,97].

Many scholars wish to empirically study how often and under what conditions individual voting rules fall victim to various voting irregularities [22,61]. Due to a lack of large, accurate datasets, many computer scientists and political scientists are turning towards statistical distributions to generate election scenarios in order to verify and test voting rules and other decision procedures [111, 130]. These statistical models may or may not be grounded in reality and it is an open problem in both the political science and social choice fields as to what, exactly, election data looks like [124].

A fundamental problem in research into properties of voting rules is the lack of large data sets to run empirical experiments [108,124]. There have been studies of several distinct datasets but these are limited in both number of elections analyzed [22] and size of individual elections within the datasets analyzed [61,124]. While there is little agreement about the frequency with which different voting paradoxes occur or the consensus between voting methods, all the studies so far have found little evidence of **Condorcet's Voting**

Paradox [66] (a cyclical majority ordering) or **preference domain restrictions** such as **single peakedness** [13] (where one candidate out of a set of three is never ranked last). Additionally, most of the studies find a strong consensus between most voting rules except Plurality [22, 61, 108].

As the computational social choice community continues to grow there is increasing attention on empirical results (see, e.g., [130]). The empirical data will support and justify the theoretical concerns prevalent in the literature and that we have discussed in previous chapters [33,57]. Walsh explicitly called for the establishment of a repository of voting data in his COMSOC 2010 talk [131]. We begin to respond to this call through the identification, analysis, and posting of a new repository of voting data.

In this Chapter we detail the discovery and evaluation of an extremely large number of distinct 3 and 4 candidate elections derived from a novel dataset. We begin in Section 5.2 with a survey of the datasets that are commonly used in the literature. We then detail in Section 5.3 our new dataset, including summary statistics and a basic overview of the data. We then move into Section 5.4 which is broken into multiple subsections where we attempt to answer many of the questions about voting. Section 5.4.1 details an analysis that attempts to answer the questions "How often does Concert's Paradox occur?" and "How often does any voting cycle occur?" We continue with Section 5.4.2 which looks at the prevalence of single peaked preferences and other domain restricted election profiles [13, 117]. Section 5.4.3 investigates the consensus between multiple voting rules. We evaluate our millions of elections under the voting rules: Plurality, Copeland, Borda, Repeated Alternative Vote, and *k*-Approval. In Section 5.4.4 we evaluate our new dataset against many of the statistical models that are in use in the ComSoc and social choice communities to generate synthetic election data. We conclude in Section 5.5 with observations about our data in the context election systems and the current trends in computational social choice.

5.2 Survey of Existing Datasets

The literature on the empirical analysis of large voting datasets is somewhat sparse, and many studies use the same datasets [61, 124]. These problems can be attributed to the lack of large amounts of data from real elections [108]. Chamberlin et al. [22] provided empirical analysis of five elections of the American Psychological Association (APA). These elections range in size from 11,000 to 15,000 ballots (some of the largest elections studied). Within these elections there are no cyclical majority orderings and, of the six voting rules under study, only Plurality fails to coincide with the others on a regular basis. Similarly, Regenwetter et al. analyzed APA data from later years [109] and observed the same phenomena: a high degree of stability between elections rules. Felsenthal et al. [61] analyzed a dataset of 36 unique voting instances from unions and other professional organizations in Europe. Under a variety of voting rules Felsenthal et al. also found a high degree of consensus between voting rules (with the notable exception of Plurality).

All of the empirical studies surveyed [22, 61, 95, 108, 109, 124] came to a similar conclusion: there is scant evidence for occurrences of Condorcet's Paradox [97]. Many of these studies find no occurrence of majority cycles (and those that find cycles find them in rates of much less than 1% of elections). Additionally, each of these (with the exception of Niemi and his study of university elections, which he observes is a highly homogeneous population [95]) find almost no occurrences of either single-peaked preferences [13] or the more general value-restricted preferences [117].

Given this lack of data and the somewhat surprising results regarding voting irregularities, some authors have taken a more statistical approach. Over the years multiple statistical models have been proposed to generate election pseudo-data to analyze (e.g., [108, 124]). Gehrlein [66] provides an analysis of the probability of occurrence of Condorcet's Paradox in a variety of election cultures. Gehrlein exactly quantifies these probabilities and concludes that Condorcet's Paradox probably will only occur with very small electorates. Gehrlein states that some of the statistical cultures used to generate election pseudo-data, specifically the Impartial Culture, may actually represent a worst-case scenario when analyzing voting rules for single-peaked preferences and the likelihood of observing Condorcet's Paradox [66]

Tideman and Plassmann have undertaken the task of verifying the statistical cultures used to generate pseudo-election data [124]. Using one of the largest datasets available, Tideman and Plassmann find little evidence supporting the models currently in use to generate election data. Additionally, Tideman and Plassmann propose several novel statistical models which better fit their empirical data.

5.3 The New Data

We have mined strict preference orders from the Netflix Prize Dataset [10]. The Netflix dataset offers a vast amount of preference data; compiled and publicly released by Netflix for its Netflix Prize [10]. There are 100,480,507 distinct ratings in the database. These ratings cover a total of 17,770 movies and 480,189 distinct users. Each user provides a numerical ranking between 1 and 5 (inclusive) of some subset of the movies. While all movies have at least one ranking it is not that case that all users have rated all movies. The dataset contains every movie rating received by Netflix, from its users, between when Netflix started tracking the data (early 2002) up to when the competition was announced (late 2005). This data has been perturbed to protect privacy and is conveniently coded for use by researchers.

The Netflix data is rare in preference studies: it is more sincere than most other preference data sets. Since users of the Netflix service will receive better recommendations from Netflix if they respond truthfully to the rating prompt, there is an incentive for each user to express sincere preference. This is in contrast to many other datasets which are compiled through surveys or other methods where the individuals questioned about their preferences have no stake in providing truthful responses.

We define an election as $E(m,n)$, where m is a set of candidates, $\{c_1,\ldots,c_m\}$, and n is a set of votes. A vote is a strict preference ordering over all the candidates $c_1 > c_2 > \cdots > c_m$. For convenience and ease of exposition we will often speak in the terms of a three candidate election and label the candidates as A, B, C and preference profiles as $A > B > C$. All results and discussion can be extended to the case of more than three candidates. A voting rule takes, as input, a set of candidates and a set of votes and returns a set of winners which may be empty or contain one or more candidates. In our discussion, elections return a complete ordering over all the candidates in the election with no ties between candidates (after a tiebreaking rule has been applied). The candidates in our data set correspond to movies from the Netflix dataset and the votes correspond to strict preference orderings over these movies. We break ties according to the lowest numbered movie identifier in the Netflix set; these are random, sequential numbers assigned to every movie.

We construct vote instances from this dataset by looking at combinations of three movies. If we find a user with a strict preference ordering over the three moves, we tally that as a vote. For example, given movies A,B, and C: if a user rates movie $A = 1$, $B = 3$, and $C = 5$, then the user has a strict preference profile over the three movies we are considering and hence a vote. If we can find 350 or more votes for a particular movie triple then we regard that movie triple as an election and we record it. We use 350 as a cutoff for an election as it is the number of votes used by Tideman and Plassmann [124] in their study of voting data. While this is a somewhat arbitrary cutoff, Tideman and Plassmann claim it is a sufficient number to eliminate random noise in the elections [124]. We use the 350 number so that our results are directly comparable to the results reported by Tideman and Plassmann.

The dataset is too large to use completely ($\binom{17770}{3} \approx 1 \times 10^{12}$) and we have therefore subdivide. We have divided the movies into 10 independent (non-overlapping with respect to movies), randomly drawn samples of 1777 movies. This completely partitions the set of movies. For each sample we search all the $\binom{17770}{3} \approx 9.33 \times 10^8$ possible elections for those with more than 350 votes. For 3 candidate elections, this search generated 14,003,522

Figure 5.1: Empirical CDF of Set 1 for 3 candidate elections.

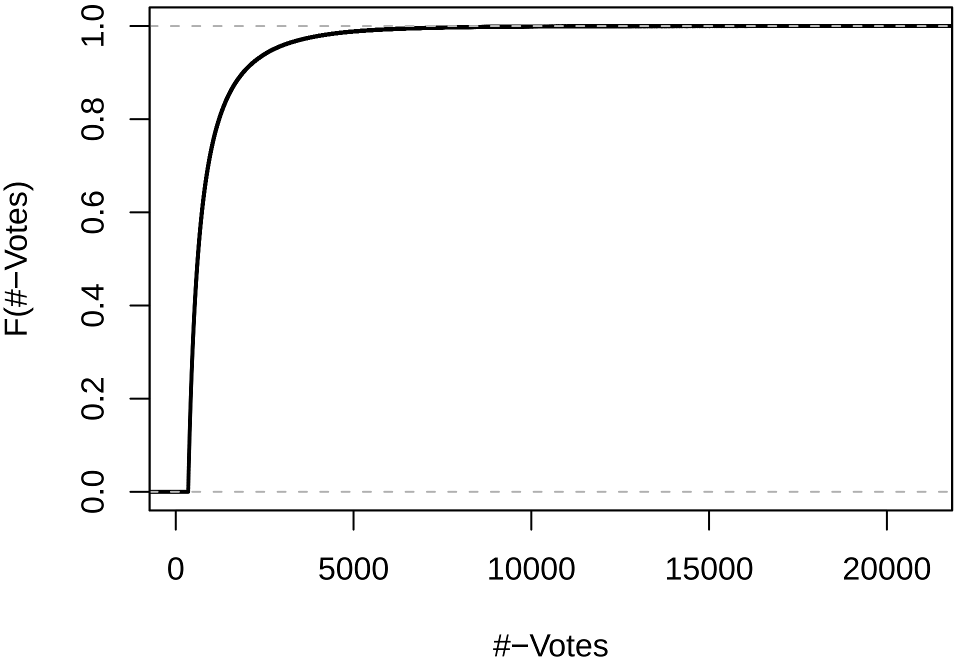
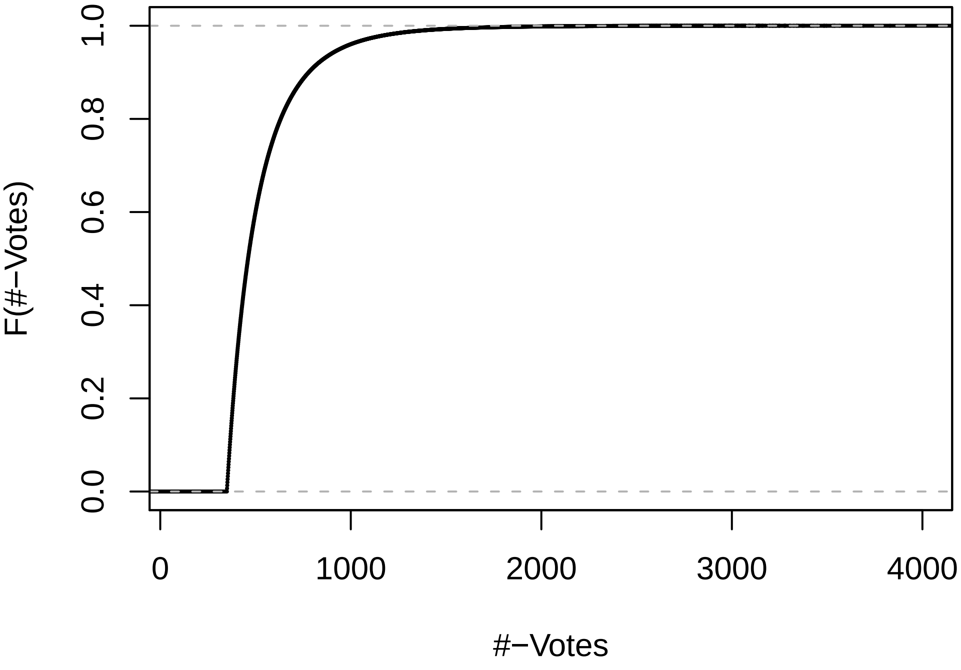


Figure 5.2: Empirical CDF of Set 1 for 4 candidate elections.



distinct movie triples in total over all the subdivisions. Not all users have rated all movies so the actual number of elections for each set is not consistent. The maximum election size found in the dataset is 24,670 votes; metrics of central tendency are presented in Tables 5.3 and 5.3. Figures 5.1 and 5.2 show the empirical cumulative distribution functions (ECFD) for Set 1 with 3 and 4 candidates respectively. The other set CDF's are similar.

Table 5.1: Summary statistics for 3 candidate elections.

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10
Min.	350.0	350.0	350.0	350.0	350.0	350.0	350.0	350.0	350.0	350.0
1st Qu.	443.0	438.0	440.0	435.0	435.0	435.0	435.0	435.0	441.0	433.0
Median	610.0	592.0	597.0	583.0	581.0	584.0	585.0	580.0	600.0	573.0
Mean	964.8	880.6	893.3	843.3	829.9	853.2	868.4	841.3	862.7	779.2
3rd Qu.	1,011.0	958.0	960.0	921.0	915.0	923.0	935.0	911.0	963.0	876.0
Max.	18,270.0	19,480.0	19,040.0	17,930.0	12,630.0	20,250.0	24,670.0	21,260.0	17,750.0	13,230.0
Elements	1,453,012.0	1,640,584.0	1,737,858.0	1,495,316.0	1,388,892.0	1,344,775.0	931,403	1,251,478	1,500,040	1,260,164

Using the notion of item-item extension [70] we attempted to extend every triple found in the initial search. Item-item extension allows us to trim our search space by only searching for 4 movie combinations which contain a combination of 3 movies that was a valid voting instance. For each set we only searched for extensions within the same draw of 1777 movies, making sure to remove any duplicate extensions. The results of this search are summarized in Table 5.3. For 4 candidate elections, this search generated 11,362,358 distinct movie triples over all subdivisions. Our constructed datasets contains more than 5 orders of magnitude more distinct elections than all the previous studies *combined* and the largest single election contains slightly more votes than the largest previously studied election from data.

Table 5.2: Summary statistics for 4 candidate elections.

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10
Min.	350.0	350.0	350.0	350.0	350.0	350.0	350.0	350.0	350.0	350.0
1st Qu.	397.0	390.0	392.0	388.0	386.0	389.0	390.0	388.0	383.0	380.0
Median	471.0	450.0	458.0	446.0	440.0	449.0	454.0	447.0	432.0	424.0
Mean	555.6	512.2	532.7	508.0	490.2	512.2	521.3	513.0	475.8	468.2
3rd Qu.	623.0	566.0	588.0	558.0	514.0	563.0	579.0	561.0	521.0	507.0
Max.	3,519.0	2,965.0	4,032.0	2,975.0	2,192.0	3,400.0	3,511.0	3,874.0	2,574.0	2,143.0
Elements	1,881,695.0	1,489,814.0	1,753,990	1,122,227.0	1,032,874	1,082,377.0	642,537	811,130	1,117,798	427,916

The data mining and experiments were performed on a pair of dedicated machines with dual-core Athlon 64x2 5000+ processors and 4 gigabytes of RAM. All the programs for searching the dataset and performing the experiments were written in C++. All of the statistical analysis was performed in R using RStudio.

The initial search of three movie combinations took approximately 36 hours (parallelized over the two cores) for each of the ten independently drawn sets. The four movie extension searches took approximately 250 hours per set. Computing the results of the various voting rules, checking for domain restrictions, and checking for cycles took approximately 20 hours per dataset. Calibrating and verifying the statistical distributions took approximately 20 hours per dataset. All the computations for this project are straightforward, the benefit of modern computational power allows our parallelized code to more quickly search the billions of possible movie combinations.

5.4 Analysis and Discussion

We have found a large correlation between each pair of voting rules under study with the exception of Plurality (when $m = 3,4$) and 2-Approval (when $m = 3$). A **Condorcet Winner** is a candidate who is preferred by a majority of the voters to each of the other candidates in an election [61]. The voting rules under study, with the exception of Copeland, are not **Condorcet Consistent**: they do not necessarily select a Condorcet Winner if one exists [97]. Therefore, we also analyze the voting rules in terms of their **Condorcet Efficiency**, the rate at which the rule selects a Condorcet Winner if one exists [93]. In Section 5.4.3 we see that the voting rules exhibit a high degree of Condorcet Efficiency in our dataset. The results in Section 5.4.1 show extremely small evidence for cases of single peaked preferences and very low rates of occurrence of preference cycles. Finally, the experiments in Section 5.4.4 indicate that several statistical models currently in use for testing new voting rules [111] do not reflect the reality of our dataset. All of these results are in keeping with the analysis of other, distinct, datasets [22,61,95,108,109,124] and provide support for their conclusions.

5.4.1 Preference Cycles

Condorcet's Paradox of Voting is the observation that rational group preferences can be aggregated, through a voting rule, into an irrational total preference [97]. It is an important theoretical and practical concern to evaluate how often the scenario arises in empirical data. In addition to analyzing instances of **total cycles** (Condorcet's Paradox) involving all candidates in an election, we check for two other types of cyclic preferences. We also search our results for both **partial cycles**, a cyclic ordering that does not include the top candidate (Condorcet Winner), and **partial top cycles**, a cycle that includes the top candidate but excludes one or more other candidates [61].

Tables 5.3 and 5.4 summarize the rates of occurrence of the different types of voting cycles found in our data sets. The cycle counts for $m = 3$ are all equivalent due to the fact that there is only one type of possible cycle when $m = 3$. There is an extremely low instance of total cycles for all our data ($< 0.11\%$ of all elections). This corresponds to findings in the empirical literature that support the conclusion that Condorcet's Paradox has a low incidence of occurrence. Likewise, cycles of any type occur in rates $< 0.4\%$ and therefore seem of little practical importance in our dataset as well. Our results for cycles that do not include the winner mirror the results of Felsenthal et al. [61]: many cycles occur in the lower ranks of voters' preference orders in the election due to the voters' inability to distinguish between, or indifference towards, candidates the voter has a low ranking for or considers irrelevant.

Table 5.3: Number of elections demonstrating various types of voting cycles for 3 candidate elections.

	Set 1	Set 2	Set 3	Set 4	Set 5
Partial Cycle	667 (0.05%)	775 (0.05%)	995 (0.06%)	576 (0.04%)	638 (0.05%)
Partial Top	667 (0.05%)	775 (0.05%)	995 (0.06%)	576 (0.04%)	638 (0.05%)
Total	667 (0.05%)	775 (0.05%)	995 (0.06%)	576 (0.04%)	638 (0.05%)

	Set 6	Set 7	Set 8	Set 9	Set 10
Partial Cycle	560 (0.04%)	446 (0.05%)	512 (0.04%)	556 (0.04%)	896 (0.07%)
Partial Top	560 (0.04%)	446 (0.05%)	512 (0.04%)	556 (0.04%)	896 (0.07%)
Total	560 (0.04%)	446 (0.05%)	512 (0.04%)	556 (0.04%)	896 (0.07%)

Table 5.4: Number of elections demonstrating various types of voting cycles for 4 candidate elections.

	Set 1	Set 2	Set 3	Set 4	Set 5
Partial Cycle	4,088 (0.22%)	4,360 (0.29%)	3,879 (0.22%)	1,599 (0.14%)	1,316 (0.13%)
Partial Top	2,847 (0.15%)	3,042 (0.20%)	2,951 (0.17%)	1,165 (0.10%)	974 (0.09%)
Total	892 (0.05%)	1,110 (0.07%)	937 (0.05%)	427 (0.04%)	293 (0.03%)

	Set 6	Set 7	Set 8	Set 9	Set 10
Partial Cycle	1,597 (0.15%)	1,472 (0.23%)	1,407 (0.17%)	1,274 (0.11%)	1,646 (0.38%)
Partial Top	1,189 (0.11%)	1,222 (0.19%)	1,018 (0.13%)	870 (0.08%)	1,123 (0.26%)
Total	325 (0.03%)	438 (0.07%)	331 (0.04%)	198 (0.02%)	451 (0.11%)

5.4.2 Domain Restrictions

Black first introduced the notion of single-peaked preferences [13], a domain restriction that states that the candidates can be ordered along one axis of preference and there is a single peak to the graph of all votes by all voters if the candidates are ordered along this axis. Informally, the idea is that every member of the society has an (not necessarily identical) ideal point along a single axis and that, the farther an alternative is from the bliss point, the lower that candidate will be ranked. A conical example is that everyone has a preference for the volume of music in a room, the farther away (either louder or softer) the music is set, the less prefered that volume is.

This is expressed in an election as the scernio when some candidate, in a three candidate election, is never ranked last. The notion of restricted preference profiles was extended by Sen [117] to include the idea of candidates who are never ranked first (single-bottom) and candidates who are always ranked in the middle (single-mid). Domain restrictions can be expanded to the case where elections contain more than three candidates [3]. Preference restrictions have important theoretical applications and are widely studied in the area of election manipulation. Many election rules become trivially easy to affect through bribery or manipulation when electorates preferences are single-peaked [19].

Table 5.5 and Table 5.6 summarizes our results for the analysis of different restricted preference profiles. There is (nearly) a complete lack (10 total instances over all sets) of preference profile restrictions when $m = 4$ and near lack ($< 0.05\%$) when $m = 3$. It is important to remember that the underlying objects in this dataset are movies, and individuals, most likely, evaluate movies for many different reasons. Therefore, as the results of our analysis confirm, there are very few items that users rate with respect to a single dimension.

Table 5.5: Number of 3 candidate elections demonstrating preference profile restrictions.

	Set 1	Set 2	Set 3	Set 4	Set 5
Single Peaked	29 (0.002%)	92 (0.006%)	624 (0.036%)	54 (0.004%)	11 (0.001%)
Single Mid	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)
Single Bottom	44 (0.003%)	215 (0.013%)	412 (0.024%)	176 (0.012%)	24 (0.002%)

	Set 6	Set 7	Set 8	Set 9	Set 10
Single Peaked	162 (0.012%)	148 (0.016%)	122 (0.010%)	168 (0.011%)	43 (0.003%)
Single Mid	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)
Single Bottom	590 (0.044%)	147 (0.016%)	152 (0.012%)	434 (0.029%)	189 (0.015%)

Table 5.6: Number of 4 candidate elections demonstrating preference profile restrictions.

	Set 1	Set 2	Set 3	Set 4	Set 5
Single Peaked	0 (0.000%)	0 (0.000%)	1 (0.000%)	0 (0.000%)	0 (0.000%)
Single Mid	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)
Single Bottom	0 (0.000%)	0 (0.000%)	2 (0.000%)	0 (0.000%)	0 (0.000%)

	Set 6	Set 7	Set 8	Set 9	Set 10
Single Peaked	0 (0.000%)	3 (0.000%)	0 (0.000%)	1 (0.000%)	0 (0.000%)
Single Mid	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)
Single Bottom	1 (0.000%)	3 (0.000%)	0 (0.000%)	2 (0.000%)	0 (0.000%)

5.4.3 Voting Rules

The variety of voting rules and election models that have been implemented or "improved" over time is astounding. Arrow shows that any preference aggregation scheme for three or more alternatives cannot meet some simple fairness conditions [2]. This leads most scholars to question "which voting rule is the best?" We analyze our dataset under the voting rules Plurality, Borda, 2-Approval, and Repeated Alternative Vote (RAV). We briefly describe the voting rules under analysis. A more complete treatment of voting rules and their properties can be found in Nurmi [97], Arrow, Sen, and Suzumura [3], or Section 2.2.1.

Plurality: Plurality is the most widely used voting rule [97] (and, to many Americans, synonymous with the term "voting"). The Plurality score of a candidate is the sum of all the first place votes for that candidate. No other candidates in the vote are considered besides the first place vote. The winner is the candidate with the highest score.

k-Approval: Under k-Approval voting, when a voter casts a vote, the first k candidates each receive the same number of points. In a 2-Approval scheme, the first 2 candidates of every voter's preference order would receive the same number of points. The winner of a k-Approval election is the candidate with the highest total score.

Copeland: In a Copeland election each pairwise contest between candidates is considered. If candidate a defeats candidate b in a head-to-head comparison of first place votes then candidate a receives 1 point; a loss is -1 and a tie is worth 0 points. After all head-to-head comparisons are considered, the candidate with the highest total score is the winner of the election.

Borda: Borda's System of Marks involves assigning a numerical score to each position. In most implementations [97] the first place candidate receives $c-1$ points, with each candidate later in the ranking receiving one less points down to 0 points for the last ranked candidate. The winner is the candidate with the highest total score.

Repeated Alternative Vote: Repeated Alternative Vote (RAV) is an extension of the Alternative Vote (AV) into a rule which returns a complete order over all the candidates

[61]. For the selection of a single candidate there is no difference between RAV and AV. Scores are computed for each candidate as in Plurality. If no candidate has a strict majority of the votes the candidate receiving the fewest first place votes is dropped from all ballots and the votes are re-counted. If any candidate now has a strict majority, they are the winner. This process, for c candidates, is repeated up to $c-1$ times [61]. In RAV this procedure is repeated, removing the winning candidate from all votes in the election after they have won, until no candidates remain. The order in which the winning candidates were removed is the total ordering of all the candidates.

We follow the analysis outlined by Felsenthal et al. [61]. We establish the Copeland order as "ground truth" in each election; Copeland always selects the Condorcet Winner if one exists and many feel the ordering generated by the Copeland rule is the "most fair" when no Condorcet Winner exists [61, 97]. After determining the results of each election, for each voting rule, we compare the order produced by each rule to the Copeland order and compute the Spearman's Rank Order Correlation Coefficient (Spearman's ρ) to measure similarity [61].

Table 5.7 and Table 5.8 lists the mean and standard deviation for Spearman's Rho between the various voting rules and Copeland. All sets had a median value of 1.0. Our analysis supports other empirical studies in the field that find a high consensus between the various voting rules [22, 61, 109]. Plurality performs the worst as compared to Copeland across all the datasets. 2-Approval does fairly poorly when $m = 3$ but does surprisingly well when $m = 4$. We suspect this discrepancy is due to the fact that when $m = 3$, individual voters are able to select a full $^2/_3$ of the available candidates.

Table 5.7: Voting results (Spearman's ρ) for 3 candidate elections.

		Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10
Plurality	Mean	0.9277	0.9275	0.9192	0.9375	0.9279	0.9308	0.9226	0.9280	0.9358	0.9130
	SD	0.2013	0.1998	0.2172	0.1849	0.1983	0.1966	0.2110	0.1998	0.1873	0.2263
2-Approval	Mean	0.9171	0.9157	0.9042	0.9191	0.9219	0.9222	0.9095	0.9152	0.9318	0.9065
	SD	0.2132	0.2149	0.2299	0.2058	0.2048	0.2112	0.2252	0.2130	0.1920	0.2317
Borda	Mean	0.9789	0.9792	0.9761	0.9801	0.9792	0.9803	0.9778	0.9782	0.9819	0.9756
	SD	0.1024	0.1021	0.1090	0.0995	0.1019	0.0993	0.1208	0.1039	0.0953	0.1105
RAV	Mean	0.9982	0.9982	0.9978	0.9987	0.9984	0.9985	0.9980	0.9984	0.9986	0.9975
	SD	0.0367	0.0372	0.0414	0.0318	0.0353	0.0333	0.0373	0.0337	0.0331	0.0443

Table 5.8: Voting results (Spearman's ρ) for 4 candidate election.

		Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10
Plurality	Mean	0.8757	0.8851	0.8551	0.9178	0.9037	0.8983	0.8835	0.8922	0.9047	0.7979
	SD	0.2003	0.1853	0.2192	0.1487	0.1590	0.1765	0.1940	0.1802	0.1675	0.2738
2-Approval	Mean	0.9504	0.9540	0.9511	0.9562	0.9554	0.9575	0.9554	0.9536	0.9677	0.9370
	SD	0.1028	0.1000	0.1032	0.0950	0.0943	0.0975	0.0984	0.0993	0.0838	0.1168
Borda	Mean	0.9747	0.9762	0.9739	0.9779	0.9792	0.9797	0.9765	0.9755	0.9847	0.9649
	SD	0.0734	0.0717	0.0745	0.0679	0.0652	0.0656	0.0710	0.0722	0.0567	0.0865
RAV	Mean	0.9962	0.9958	0.9956	0.9980	0.9982	0.9976	0.9962	0.9973	0.9982	0.9929
	SD	0.0365	0.0395	0.0386	0.0258	0.0246	0.0278	0.0357	0.0293	0.0253	0.0496

There are many considerations one must make when selecting a voting rule for use within a given system. Merrill suggests that one of the most powerful metrics is Condorcet Efficiency [93]. Table 5.9 and Table 5.10 shows the proportion of Condorcet Winners selected by the various voting rules under study. We eliminated all elections that did not have a Condorcet Winner in this analysis. All voting rules select the Condorcet Winner a surprising majority of the time. The worst case is 2-Approval, when $m = 3$, as it results in the lowest Condorcet Efficiency in our dataset. The high rate of elections that have a Condorcet Winner ($> 80\%$) could be an artifact of how we select elections. By virtue of enforcing strict orders we are causing a selection bias in our set: we are only checking elections where many voters have a preference between any two items in the dataset.

Table 5.9: Condorcet Efficiency of the various voting rules for 3 candidate elections.

	Total Elections	Condorcet Winners	Plurality	2-Approval	Borda	RAV
Set 1	1,453,012	1,447,881	0.9661	0.8789	0.9785	0.9977
Set 2	1,640,584	1,634,829	0.9664	0.8775	0.9785	0.9975
Set 3	1,737,858	1,731,039	0.9583	0.8615	0.9744	0.9968
Set 4	1,495,316	1,490,265	0.9703	0.8752	0.9782	0.9982
Set 5	1,388,892	1,383,674	0.9676	0.8860	0.9795	0.9977
Set 6	1,344,775	1,340,321	0.9694	0.8882	0.9799	0.9979
Set 7	931,403	927,878	0.9612	0.8709	0.9771	0.9970
Set 8	1,251,478	1,247,199	0.9667	0.8729	0.9773	0.9977
Set 9	1,500,040	1,495,133	0.9725	0.9003	0.9825	0.9981
Set 10	1,260,164	1,254,712	0.9562	0.8711	0.9762	0.9964

Table 5.10: Condorcet Efficiency of the various voting rules for 4 candidate elections.

	Total Elections	Condorcet Winners	Plurality	2-Approval	Borda	RAV
Set 1	1,881,695	1,864,788	0.9450	0.9171	0.9610	0.9950
Set 2	1,489,814	1,474,697	0.9319	0.9100	0.9582	0.9926
Set 3	1,753,990	1,737,939	0.9382	0.9263	0.9629	0.9924
Set 4	1,122,227	1,113,378	0.9506	0.9010	0.9529	0.9961
Set 5	1,032,874	1,024,918	0.9583	0.9181	0.9632	0.9975
Set 6	1,082,377	1,074,434	0.9667	0.9380	0.9716	0.9970
Set 7	642,537	636,469	0.9294	0.9212	0.9601	0.9923
Set 8	811,130	804,496	0.9394	0.9058	0.9555	0.9946
Set 9	1,117,798	1,110,069	0.9708	0.9564	0.9816	0.9973
Set 10	427,916	422,489	0.9054	0.9205	0.9573	0.9891

Overall, we find a consensus between the various voting rules in our tests. This supports the findings of other empirical studies in the field [22, 61, 109]. Merrill finds much lower rates for Condorcet Efficiency than we do in our study [93]. However, Merrill uses statistical models to generate elections rather than empirical data to compute his numbers and this is likely the cause of the discrepancy [66].

5.4.4 Statistical Models of Elections

We evaluate our dataset to see how it matches up to different probability distributions found in the literature. We briefly detail several probability distributions (or "cultures") here that we test. Tideman and Plassmann provide a more complete discussion of the variety of statistical cultures in the literature [124]. There are other election generating cultures, such as weighted Independent Anonymous Culture, which generate preference profiles that are skewed towards single-peakedness or single-bottomness. As we have found no support in our analysis for restricted preference profiles we do not analyze these cultures (a further discussion and additional election generating statistical models can be found in [124]).

We follow the general outline in Tideman and Plassmann to guide us in this study [124]. For ease of discussion we divide the models into two groups: probability models (IC, DC, UC, UUP) and generative models (IAC, Urn, IAC-Fit). Probability models define a probability vector over each of the $m!$ possible strict preference rankings. We note these probabilities as $pr(ABC)$, which is the probability of observing a vote $A > B > C$ for each of the possible orderings. In order to compare how the statistical models describe the empirical data, we compute the mean Euclidean distance between the empirical probability distribution and the one predicted by the model.

Impartial Culture (IC): An even distribution over every vote exists. That is, for the $m!$ possible votes, each vote has probability $1/m!$ (a uniform distribution).

Dual Culture (DC): The dual culture assumes that the probability of opposite preference orders is equal. So, $pr(ABC) = pr(CBA)$, $pr(ACB) = pr(BCA)$ etc. This culture is based on the idea that some groups are polarized over certain issues.

Uniform Culture (UC): The uniform culture assumes that the probability of distinct pairs of lexicographically neighboring orders (that share the same top candidate) are equal. For example, $pr(ABC) = pr(ACB)$ and $pr(BAC) = pr(BCA)$ but not $pr(ACB) = pr(CAB)$ (as, for three candidates, we pair them by the same winner). This culture corresponds to situations where voters have strong preferences over the top candidates but may be indifferent over candidates lower in the list.

Unequal Unique Probabilities (UUP): The unequal unique probabilities culture defines the voting probabilities as the maximum likelihood estimator over the entire dataset. We determine, for each of the data sets, the UUP distribution as described below.

For DC and UC each election generates its own statistical model according to the definition of the given culture. In order to calibrate the UUP we need to determine a multinomial probability distribution over the vote vectors. We follow a similar method described in Tideman and Plassmann [124]. To simplify discussion assume we have 3 candidates ($m = 3$) and therefore $m! = 6$ possible vote vectors. Call these $p_1 = ABC$, $p_2 = ACB$, $p_3 = BAC$, $p_4 = BCA$, $p_5 = CAB$, and $p_6 = CBA$.

We are calibrating UUP to some empirical (or observed) set of vote vectors. For each observation we re-label the voters so that, in the most common order, A is first, B is second, and C is third. Once this relabeling has occurred we want to find, for each election, the probability vector that maximizes the log likelihood of Equation 5.1.

$$f(N_1,\dots,N_6;N,p_1,\dots,p_6) = \frac{N!}{\prod_{r=1}^{6} N_r!} \prod_{r=1}^{6} p_r^{N_r} \tag{5.1}$$

Where N_r is the number of votes for vector r; N is the total number of votes in an election; and p_r is the probability (proportional share) of votes received by the particular preference ordering r. Note that $0 \leq p_r \leq 1$, $p_r = {}^{N_r}\!/_{N}$, and $\sum_{r=1}^{6} p_r = 1$. When we take the log of this equation we end up with Equation 5.2.

$$g(N_1,\ldots,N_6;N,p_1,\ldots,p_6)=\left\{\log N!-(\sum_{r=1}^{6}\log N_r!)\right\}+(\sum_{r=1}^{6}N_r\log p_r) \tag{5.2}$$

Since the term in the brackets is not dependent on the probability distribution we can drop it from our maximization problem. So, after rewriting we want to maximize Equation 5.3.

$$\max_{p_r}\left\{\sum_{r=1}^{6}N_r\log(p_r)\right\} \tag{5.3}$$

Using Theorem 1.2.3 from Roman [112], page 22, we know that the vector which maximizes this equation is exactly the vector p_r from the empirical observation. So the maximum log-likelihood estimator for each election, according to the link function in Equation 5.1, is the empirical vector of vote shares.

In order to find the UUP distribution for an entire group of elections we have to find a probability vector that maximizes the log-likelihood of predicting all the reordered vote vectors. Let our set of relabeled elections be E. Again, we can drop the first term from Equation 5.2 as it has no effect since it is a constant scalar. Let $N_{r,i}$ be the number of votes for order $r \in R$ for election $i \in E$. Rewriting Equation 5.3 for all elections gives us an equation for our UUP distribution.

$$UUP=\max_{p_r}\left\{\sum_{r=1}^{6}\left(\sum_{i\in E}N_{r,i}\right)\log(p_r)\right\} \tag{5.4}$$

With this equation we can again apply Theorem 1.2.3 from Roman [112]. Let

$$M=\sum_{r=0}^{6}\sum_{i\in E}N_{r,i}.$$

And let

$$S_r=\sum_{i\in E}N_{r,i}.$$

Then we can rewrite Equation 5.4 and are left with Equation 5.5.

$$\forall r \in \{1,\ldots,6\} : p_r = \frac{S_r}{M} \tag{5.5}$$

We can expand the above maximization problem for all 24 possible vectors when $m = 4$. To compute the error between the culture's distribution and the empirical observations, we re-label the culture distribution so that the preference order with the most votes in the empirical distribution matches the culture distribution and compute the error as the mean Euclidean distance between the discrete probability distributions.

Urn Model: The Polya Eggenberger urn model is a method designed to introduce some correlation between votes and does not assume a complete uniform random distribution [11]. We use a setup as described by Walsh [130]; we start with a jar containing one of each possible vote. We draw a vote at random and place it back into the jar with $a \in \mathbb{Z}_+$ additional votes of the same kind. We repeat this procedure until we have created a sufficient number of votes.

Impartial Anonymous Culture (IAC): Every distribution over orders has an equal likelihood. For each generated election we first randomly draw a distribution over all the $m!$ possible voting vectors and then use this model to generate votes in an election.

IAC-Fit: For this model we first determine the vote vector that maximizes the log-likelihood of Equation 5.1 without the reordering described for UUP. Using the probability vector obtained for $m = 3$ and $m = 4$ we randomly generate elections. This method generates a probability distribution or culture that represents our entire dataset.

For the generative models we must generate data in order to compare them to the culture distributions. To do this we average the total elections found for $m = 3$ and $m = 4$ and generate 1,400,352 and 1,132,636 elections, respectively. We then draw the individual election sizes randomly from the distribution represented in our dataset. After we generate these random elections we compare them to the probability distributions predicted by the various cultures.

Table 5.11: Mean Euclidean distance between the empirical data set and different statistical cultures (standard error in parentheses) for elections with 3 candidates.

	IC	DC	UC	UUP
Set 1	0.3064 (0.0137)	0.2742 (0.0113)	0.1652 (0.0087)	0.2817 (0.0307)
Set 2	0.3106 (0.0145)	0.2769 (0.0117)	0.1661 (0.0089)	0.2818 (0.0311)
Set 3	0.3005 (0.0157)	0.2675 (0.0130)	0.1639 (0.0091)	0.2860 (0.0307)
Set 4	0.3176 (0.0143)	0.2847 (0.0113)	0.1758 (0.0100)	0.2833 (0.0332)
Set 5	0.2974 (0.0125)	0.2677 (0.0104)	0.1610 (0.0082)	0.2774 (0.0300)
Set 6	0.3425 (0.0188)	0.3027 (0.0143)	0.1734 (0.0108)	0.3113 (0.0399)
Set 7	0.3043 (0.0154)	0.2704 (0.0125)	0.1660 (0.0095)	0.2665 (0.0289)
Set 8	0.3154 (0.0141)	0.2816 (0.0114)	0.1712 (0.0091)	0.2764 (0.0318)
Set 9	0.3248 (0.0171)	0.2906 (0.0130)	0.1686 (0.0100)	0.3005 (0.0377)
Set 10	0.2934 (0.0144)	0.2602 (0.0121)	0.1583 (0.0087)	0.2634 (0.0253)
Urn	0.6228 (0.0249)	0.4745 (0.0225)	0.4745 (0.0225)	0.4914 (0.1056)
IAC	0.2265 (0.0056)	0.1691 (0.0056)	0.1690 (0.0056)	0.2144 (0.0063)
IAC-Fit	0.0363 (0.0002)	0.0282 (0.0002)	0.0262 (0.0002)	0.0347 (0.0002)

Table 5.12: Mean Euclidean distance between the empirical data set and different statistical cultures (standard error in parentheses) for elections with 4 candidates.

	IC	DC	UC	UUP
Set 1	0.2394 (0.0046)	0.1967 (0.0031)	0.0991 (0.0020)	0.2533 (0.0120)
Set 2	0.2379 (0.0064)	0.1931 (0.0042)	0.0975 (0.0023)	0.2491 (0.0127)
Set 3	0.2633 (0.0079)	0.2129 (0.0051)	0.1153 (0.0032)	0.2902 (0.0159)
Set 4	0.2623 (0.0069)	0.2156 (0.0039)	0.1119 (0.0035)	0.2767 (0.0169)
Set 5	0.2458 (0.0044)	0.2040 (0.0028)	0.1059 (0.0027)	0.2633 (0.0138)
Set 6	0.3046 (0.0077)	0.2443 (0.0045)	0.1214 (0.0040)	0.3209 (0.0223)
Set 7	0.2583 (0.0088)	0.2094 (0.0053)	0.1060 (0.0038)	0.2710 (0.0161)
Set 8	0.2573 (0.0052)	0.2095 (0.0034)	0.1059 (0.0023)	0.2508 (0.0145)
Set 9	0.2981 (0.0090)	0.2414 (0.0049)	0.1202 (0.0045)	0.3258 (0.0241)
Set 10	0.2223 (0.0046)	0.1791 (0.0035)	0.1053 (0.0021)	0.2327 (0.0085)
Urn	0.6599 (0.0201)	0.4744 (0.0126)	0.4745 (0.0126)	0.6564 (0.1022)
IAC	0.1258 (0.0004)	0.0899 (0.0004)	0.0900 (0.0004)	0.1274 (0.0004)
IAC-Fit	0.0463 (0.0001)	0.0340 (0.0001)	0.0318 (0.0001)	0.0472 (0.0001)

Table 5.11 and Table 5.12 summarizes our results for the analysis of different statistical models used to generate elections. In general, none of the probability models captures our empirical data. Uniform Culture (UC) has the lowest error in predicting the distributions found in our empirical data. We conjecture that this is due to the process by which we select

movies and the fact that these are ratings on movies. Since we require strict orders and, generally, most people rate good movies better than bad movies, we obtain elections that look like UC scenarios. By this we mean that *The Godfather* is an objectively good movie while *Mega Shark vs. Crocosaurus* is pretty bad. While there are some people who may reverse these movies, most users will rate *The Godfather* higher. This gives the population something close to a UC when investigated in the way that we do here.

The data generated by our IAC-Fit model fits very closely to the various statistical models. This is most likely due to the fact that the distributions generated by the IAC-Fit procedure closely resemble an Impartial Culture (since our sample size is so large). We, like Tideman and Plassmann, find little support for the static cultures' ability to model real data [124]

5.5 Observations and Summary

In this chapter we have identified and thoroughly evaluated a novel dataset as a source of sincere election data. We find overwhelming support for many of the existing conclusions in the empirical literature. Namely, we find a high consensus among a variety of voting methods; low occurrences of Condorcet's Paradox and other voting cycles; low occurrences of preference domain restrictions such as single-peakedness; and a lack of support for existing statistical models which are used to generate election pseudo-data. Our study is significant as it adds more results to the current discussion of what an election is and how often voting irregularities occur. Voting is a common method by which agents make decisions both in computers and as a society. Understanding the statistical and mathematical properties of voting rules, as verified by empirical evidence across multiple domains, is an important step. We provide a new look at this question with a novel dataset that is several orders of magnitude larger than the sum of the data in previous studies.

This chapter represents an initial foray into empirically testing properties of elections. While we have not directly addressed the questions of manipulation and bribery with our empirical study, we have laid the groundwork. This chapter provides us with perspective on the overall discussion of voting rules. By testing some of the theoretical properties of voting rules, and coming to the conclusion that some of the theoretical results are of little practical importance, we establish that more needs to be done to develop the empirical side of ComSoc. This empirical work is very much in the spirit of the overall ComSoc approach: we are using computational tools (data mining and access to extremely large sets of preference data) to address concerns in the social choice community. It is our hope that, with this dataset, we inspire others to look for novel datasets and empirically test some of their theoretical results.