An Empirical Study of Voting Rules and Manipulation with Large Datasets

Nicholas Mattei and James Forshee and Judy Goldsmith

Abstract

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 a publicly available data, the Netflix Prize dataset. 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). Additionally, we examine the relationship between coalition size and vote deficit for manipulations of elections under the Borda rule. We find a high consensus among the different voting rules; almost no instances of Condorcet's Paradox; almost no support for restricted preference profiles, very little support for many of the statistical models currently used to generate election data for testing, and very small coalitions needed to promote second-place candidates to the winning position in elections.

1 Introduction

One of the most common methods of preference aggregation and group decision making in human systems is voting. Many scholars wish to empirically study how often and under what conditions individual voting rules fall victim to various voting irregularities [6, 9]. 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 [22, 25]. 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 [24]. As the computational social choice community continues to grow there is increasing attention on empirical results (see, e.g., [25]) and we hope to address this problem with our study.

A fundamental problem in research into properties of voting rules is the lack of large data sets to run empirical experiments [20, 24]. There have been studies of several distinct datasets but these are limited in both number of elections analyzed [6] and size of individual elections within the datasets analyzed [9, 24]. 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* [10] (a cyclical majority ordering) or *preference domain restrictions* such as *single peakedness* [4] (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 [6, 9, 20].

We begin in Section 2 with a survey of the datasets that are commonly used in the literature. We then detail in Section 3 our new dataset, including summary statistics and a basic overview of the data. We then move into Section 4 which is broken into multiple subsections where we attempt to answer many questions about voting. Section 4.1 details an analysis that attempts to answer the questions "How often does Concert's Paradox occur?", "How often does any voting cycle occur?", and a look at the prevalence of single peaked preferences and other domain restricted election profiles [4, 23]. Section 4.2 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 4.3 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. Section 5 details an experimentwe preform to investigate, empirically, the relationship between necessary coalition size and vote deficit for manipulations of the Borda rule. This paper reports on an expanded analysis in terms of number of tests and amount of data used from the previously published work by Mattei [13, 14].

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 [9, 24]. These problems can be attributed to the lack of large amounts of data from real elections [20]. Chamberlin et al. [6] 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 [21] and observed the same phenomena: a high degree of stability between elections rules. Felsenthal et al. [9] analyzed a dataset of 36 unique voting instances from unions and other professional organizations in Europe. Recently, data from a series of elections in Ireland have been studied in a variety of contexts in social choice [12]. 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 [6, 9, 16, 20, 21, 24] came to a similar conclusion: there is scant evidence for occurrences of Condorcet's Paradox [17]. 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 [16]) find almost no occurrences of either single-peaked preferences [4] or the more general value-restricted preferences [23].

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., [20, 24]). Gehrlein [10] 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 [10]

Tideman and Plassmann have undertaken the task of verifying the statistical cultures used to generate pseudo-election data [24]. 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.

3 The New Data

We have mined strict preference orders from the Netflix Prize Dataset [2]. The Netflix dataset offers a vast amount of preference data; compiled and publicly released by Netflix for its Netflix Prize [2]. 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 [24] 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 [24]. 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 ($\binom{17770}{3} \approx 1 \times 10^{12}$) so we have subdivided it. 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 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 1 and 2.

	Set 1	Set 2	Set 3	Set 4	Set 5
Median	610.0	592.0	597.0	583.0	581.0
Mean	964.8	880.6	893.3	843.3	829.9
Max.	18,270.0	19,480.0	19,040.0	17,930.0	12,630.0
Elements	1,453,012.0	1,640,584.0	1,737,858.0	1,495,316.0	1,388,892.0
	Set 6	Set 7	Set 8	Set 9	Set 10
Median	584.0	585.0	580.0	600.0	573.0
Mean	853.2	868.4	841.3	862.7	779.2
Max.	20,250.0	24,670.0	21,260.0	17,750.0	13,230.0
Elements	1,344,775.0	931,403	1,251,478	1,500,040	1,260,164

Table 1: Summary statistics for 3 candidate elections.

Using the notion of item-item extension [11] 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 2. 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.

	Set 1	Set 2	Set 3	Set 4	Set 5
Median	471.0	450.0	458.0	446.0	440.0
Mean	555.6	512.2	532.7	508.0	490.2
Max.	3,519.0	2,965.0	4,032.0	2,975.0	2,192.0
Elements	1,881,695.0	1,489,814.0	1,753,990	1,122,227.0	1,032,874
	Set 6	Set 7	Set 8	Set 9	Set 10
Median	449.0	454.0	447.0	432.0	424.0
Mean	512.2	521.3	513.0	475.8	468.2
Max.	3,400.0	3,511.0	3,874.0	2,574.0	2,143.0
Elements	1,082,377.0	642,537	811,130	1,117,798	427,916

Table 2: Summary statistics for 4 candidate elections.

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.

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 [9]. 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 [17]. 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 [15]. In Section 4.2 we see that the voting rules exhibit a high degree of Condorcet Efficiency in our dataset. The results in Section 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 4.3 indicate that several statistical models currently in use for testing new voting rules [22] do not reflect the reality of our dataset. All of these results are in keeping with the analysis of other, distinct, datasets [6, 9, 16, 20, 21, 24] and provide support for their conclusions.

4.1 Preference Cycles and Domain Restrictions

Condorcet's Paradox of Voting is the observation that rational group preferences can be aggregated, through a voting rule, into an irrational total preference [17]. 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 [9].

Table 3 summarize the rates of occurrence of the different types of voting cycles found in 4 candidate set (3 candidate table is omitted for space). 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. [9]: 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.

	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%)

Table 3: Number of elections demonstrating various types of voting cycles for 4 candidate elections.

Black first introduced the notion of single-peaked preferences [4], 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 typical 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 preferred that volume is.

This is expressed in an election as the scenario when some candidate, in a three candidate election, is never ranked last. The notion of restricted preference profiles was extended by Sen [23] 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 [1]. Preference restrictions have important theoretical applications and are widely studied in the area of election manipulation. Many election rules become easy to affect through bribery or manipulation when electorates preferences are single-peaked [5].

Table 4 summarizes our results for the analysis of different restricted preference profiles when $m = 3$. 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.

4.2 Voting Rules

We analyze our dataset under the voting rules Plurality, Borda, 2-Approval, and Repeated Alternative Vote (RAV). We assume the reader is familiar with the normal voting rules discussed here. We

	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 4: Number of 3 candidate elections demonstrating preference profile restrictions.

note that RAV is an extension of the alternative vote (AV) where the process is repeated (removing the winning candidate at each step) to generate a total order over all the candidates. A more complete treatment of voting rules and their properties can be found in Nurmi [17] or Arrow, Sen, and Suzumura [1].

We follow the analysis outlined by Felsenthal et al. [9]. 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 [9, 17]. 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 [9].

We have omitted the tables of our results for space considerations, see Mattei [13, 14] for additional details and results. For the elections with $m = 3$ and $m = 4$ we have Borda and RAV agreeing with Copeland $\approx 98\%$ of the time, on average. For Plurality, when $m = 3$ we have $\approx 92\%$ agreement with Copeland. This correlation drops to $\approx 87\%$ when we move to $m = 4$. Plurality performs the worst as compared to Copeland across all the datasets. 2-Approval does fairly poorly when $m = 3$ ($\approx 90\%$) but does surprisingly well ($\approx 96\%$) 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. All sets had a median value of 1.0 and small standard error 0.2 for plurality and much less for all rules. Our analysis supports other empirical studies in the field that find a high consensus between the various voting rules [6, 9, 21].

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 [15]. 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. For plurality, Borda, and RAV we have a Condorcet Efficient of $\approx 95\%$, on average. The worst case is 2-Approval, when $m = 3$, as it results in the lowest Condorcet Efficiency in our dataset ($\approx 88\%$). 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.

Overall, we find a consensus between the various voting rules in our tests. This supports the findings of other empirical studies in the field [6, 9, 21]. Merrill finds much lower rates for Condorcet Efficiency than we do in our study [15]. 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 [10].

4.3 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 [24]. 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 [24]).

We follow the general outline in Tideman and Plassmann to guide us in this study [24]. 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. For UUP we need to calibrate the parameters over the entire dataset. We follow the method described in Tideman and Plassmann [24]: first re-label each empirical election in the dataset such that the order with the most votes becomes the labeling for all the other votes. This requires reshuffling the vector so that the most likely vote is always $A > B > C$. Then, over all the reordered vectors, we maximize the log-likelihood of

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

where $N_1,\ldots,N_6$ is the number of votes received by a vote vector and $p_1,\ldots,p_6$ are the probabilities of observing a particular order over all votes (we expand this equation to 24 vectors for the $m = 4$ case). To compute the error between the culture's distribution and the empirical observations, we re-label the culture distribution so that 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 [3]. We use a setup as described by Walsh [25]; 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 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.

	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: 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	0.2394 (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 6: Mean Euclidean distance between the empirical data set and different statistical cultures (standard error in parentheses) for elections with 4 candidates.

Table 5 and Table 6 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 [24]

5 Manipulation of Borda Elections

In this section, we present empirical results for experiments involving algorithms given by Zuckerman et al. to manipulate elections under the Borda voting rule [27]. Much of the analysis of manipulation and algorithms for manipulation takes place in the theoretical domain, including looking at the frequency of manipulation relative to the total election size for scoring rules given by Xia and Conitzer [26]. Additionally, Pritchard et al. have looked at the asymptotic and average set sizes necessary to manipulate elections under a variety of rules [18, 19]. Unfortunately, Pritchard's analysis is under the Impartial Culture assumption, which is an election distribution that we have seen does not match our data.

Our experiment takes ballot data for an election under the Borda rule, and a non-winning candidate, and adds manipulators one by one until the distinguished candidate wins. The question we ask is, how many manipulators are needed? The algorithm greedily calculates the ballot for each manipulator, given all of the unmanipulated ballots and the ballots of the previous manipulators. The next manipulator's ballot has the distinguished candidate first, and then lists the rest of the candidates in reverse order of their total points so far [27]. This algorithm by Zuckerman et. al has been proven to either find the optimal coalitional manipulation, or over-guess by one voter [27]. In a furthur empirical study Davies et al. compared two additional algorithms for finding Borda manipulations to Zuckerman et al.'s [8]. Davies et al. found that, while all three algorithms found the optimal manipulation over 75% of the time, Davies et al.'s AVERAGE FIT algorithm found the optimal manipulation over 99% of the time.

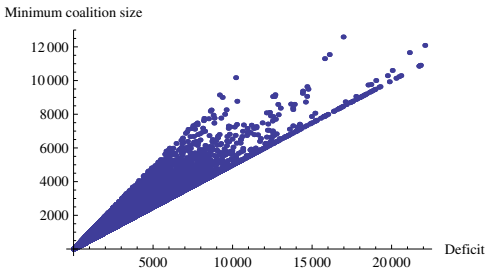

Figure 1: Deficit vs. minimum coalition size for Zuckerman's algorithm

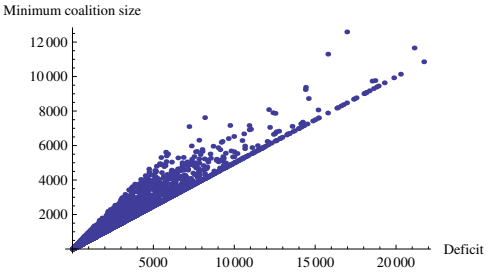

Figure 2: Deficit vs. minimum coalition for promoting third-place candidates

The size of the coalition is determined both by the distribution of votes and by the *deficit* of the distinguished candidate, namely, the difference between the number of points assigned to the current winner and the number of points assigned to the distinguished candidate. We ask a fundamentally different question than the earlier experiments on Borda manipulation algorithms. At minimum, a Borda manipulation requires a coalition size linear in the deficit size, d [8]. We want to know how

often, and under what conditions, do we have a linear coalition requirement versus when we require a super-linear coalition.

Figure 1 shows the relationship of the initial deficit to the coalition size. For our experiment we used 296,553 elections, ranging in size from 350 to 18,269 voters, from Set 1 (detailed in Section 3). The average number of voters per election in this size is 991.68, and the median is 621. Each point in the graph in Figure 1 represents the a coalition size for an election with that deficit, regardless of which candidate was promoted. For 99% of the elections we tested, it took $\lfloor \frac{d}{2} \rfloor + 1$ coalition members. Figure 2 shows coalition sizes as a function of deficit for promoting the third-place candidate to a winner.

For those elections where promoting the 3rd-place candidate took a coalition of more than $\lfloor \frac{d}{2} \rfloor + 1$, the average deficit for promoting the *second-place* candidate is 306, and the average corresponding coalition size is 154 ($= \lfloor \frac{d}{2} \rfloor + 1$). For those elections, the average deficit for promoting the *third-place* candidate is 873, and the average corresponding coalition size is 572.

6 Conclusion

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; a lack of support for existing statistical models which are used to generate election pseudo-data; and some interesting differences between the sizes of coalitions needed to promote a 2nd-place candidate and a 3rd-place candidate, using Zuckerman's algorithm for Borda. Our study is significant as it adds more results to the current discussion of what is an election and how often do voting irregularities occur? Voting is a common method by which agents make decisions both in computers and as a society. Understanding the unique 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 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.

The collection and public dissemination of the datasets is a central point our work. We plan to establish a repository of election data so that theoretical researchers can validate with empirical data. We plan to identify several other free, public datasets that can be viewed as "real world" voting data. The results reported in our study imply that our data is reusable as real world voting data. Therefore, it seems that the Netflix dataset, and its $> 10^{12}$ possible elections, can be used as a source of election data for future empirical validation of theoretical voting studies. We would like to, instead of comparing how voting rules correspond to one another, evaluate their power as maximum likelihood estimators [7]. Additionally, we would like to expand our evaluation of statistical models to include several new models proposed by Tideman and Plassmann, and others [24]. We will continue to analyze manipulation algorithms from the literature on elections from this data set.

Acknowledgements

Thanks to Dr. Florenz Plassmann for his helpful discussions on this paper and guidance on calibrating statistical models and to Elizabeth Mattei and Tom Dodson for their helpful discussion and comments on preliminary drafts of this paper. We thank the reviewers who have made this a better paper through their careful reviews. This work is supported by the National Science Foundation,

under EAGER grant CCF-1049360. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] K. Arrow, A. Sen, and K. Suzumura, editors. *Handbook of Social Choice and Welfare*, volume 1. North-Holland, 2002.

[2] J. Bennett and S. Lanning. The Netflix Prize. In *Proceedings of KDD Cup and Workshop*, 2007. `www.netflixprize.com`.

[3] S. Berg. Paradox of voting under an urn model: The effect of homogeneity. *Public Choice*, 47(2):377–387, 1985.

[4] D. Black. On the rationale of group decision-making. *The Journal of Political Economy*, 56(1), 1948.

[5] F. Brandt, M. Brill, E. Hemaspaandra, and L. A. Hemaspaandra. Bypassing combinatorial protections: Polynomial-time algorithms for single-peaked electorates. In *Proc. of the 24th AAAI Conf. on Artificial Intelligence (AAAI 2010)*, pages 715 – 722, 2010.

[6] J. R. Chamberlin, J. L. Cohen, and C. H. Coombs. Social choice observed: Five presidential elections of the American Psychological Association. *The Journal of Politics*, 46(2):479 – 502, 1984.

[7] V. Conitzer and T. Sandholm. Common voting rules as maximum likelihood estimators. In *Proc. of the 21st Annual Conf. on Uncertainty in AI (UAI)*, pages 145–152, 2005.

[8] J. Davies, G. Katsirelos, N. Narodytska, and T. Walsh. Complexity of and algorithms for borda manipulation. In *Proc. of the 25th AAAI Conf. on Artificial Intelligence (AAAI 2011)*, pages 657–662, 2011.

[9] D. S. Felsenthal, Z. Maoz, and Rapoport A. An empirical evaluation of six voting procedures: Do they really make any difference? *British Journal of Political Science*, 23:1 – 27, 1993.

[10] W. V. Gehrlein. Condorcet's paradox and the likelihood of its occurance: Different perspectives on balanced preferences. *Theory and Decisions*, 52(2):171 – 199, 2002.

[11] J. Han and M. Kamber, editors. *Data Mining*. Morgan Kaufmann, 2006.

[12] T. Lu and C. Boutilier. Robust approximation and incremental elicitation in voting protocols. In *Proc. of the 22nd Intl. Joing Conf. on Artificial Intelligence (IJCAI 2011)*, 2011.

[13] N. Mattei. Empirical evaluation of voting rules with strictly ordered preference data. In *Proceedings of the 2nd International Conference on Algorithmic Decision Theory (ADT 2011)*, 2011.

[14] N. Mattei. *Decision Making Under Uncertainty: Theoretical and Empirical Results on Social Choice, Manipulation, and Bribery*. PhD thesis, University of Kentucky, 2012.

[15] S. Merrill, III. A comparison of efficiency of multicandidate electoral systems. *American Journal of Politial Science*, 28(1):23 – 48, 1984.

[16] R. G. Niemi. The occurrence of the paradox of voting in university elections. *Public Choice*, 8(1):91–100, 1970.

[17] H. Nurmi. Voting procedures: A summary analysis. *British Journal of Political Science*, 13:181 – 208, 1983.

[18] G. Pritchard and A. Slinko. On the average minimum size of a manipulating coalition. *Social Choice and Welfare*, 27(2):263–277, 2006.

[19] G. Pritchard and M.C. Wilson. Asymptotics of the minimum manipulating coalition size for positional voting rules under impartial culture behaviour. *Mathematical Social Sciences*, 58(1):35–57, 2009.

[20] M. Regenwetter, B. Grogman, A. A. J. Marley, and I. M. Testlin. *Behavioral Social Choice: Probabilistic Models, Statistical Inference, and Applications*. Cambridge Univ. Press, 2006.

[21] M. Regenwetter, A. Kim, A. Kantor, and M. R. Ho. The unexpected empirical consensus among consensus methods. *Psychological Science*, 18(7):629 – 635, 2007.

[22] R. L. Rivest and E. Shen. An optimal single-winner preferential voting system based on game theory. In V. Conitzer and J. Rothe, editors, *Proc. of the 3rd Intl. Workshop on Computational Social Choice (COMSOC 2010)*, pages 399 – 410, 2010.

[23] A. K. Sen. A possibility theorem on majority decisions. *Econometrica*, 34(2):491–499, 1966.

[24] N. Tideman and F. Plassmann. Modeling the outcomes of vote-casting in actual elections. In D.S. Felsenthal and M. Machover, editors, *Electoral Systems: Paradoxes, Assumptions, and Procedures*. Springer, 2012.

[25] T. Walsh. An empirical study of the manipulability of single transferable voting. In *Proc. of the 19th European Conf. on AI (ECAI 2010)*, pages 257–262. IOS Press, 2010.

[26] L. Xia and V. Conitzer. Generalized scoring rules and the frequency of coalitional manipulability. In *Proceedings of the 9th ACM conference on Electronic commerce*, pages 109–118. ACM, 2008.

[27] M. Zuckerman, A. D. Procaccia, and J. S. Rosenschein. Algorithms for the coalitional manipulation problem. *Artificial Intelligence*, 173(2):392 – 412, 2009.

Nicholas Mattei, James Forshee, and Judy Goldsmith
Department of Computer Science
University of Kentucky
Lexington, KY 40506, USA
Email: nick.mattei@uky.edu,james.forshee@uky.edu,goldsmit@cs.uky.edu