Empirical Evaluation of Voting Rules with Strictly Ordered Preference Data

Nicholas Mattei

University of Kentucky
Department of Computer Science
Lexington, KY 40506, USA
nick.mattei@uky.edu

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). 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.

1 Introduction

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 [17]. With this design shift comes the need to aggregate these (possibly disjoint) observations and preferences into a total, group 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) [12]. Questions about voting and preference aggregation have circulated in the math and social choice communities for centuries [1, 8, 18].

Many scholars wish to empirically study how often and under what conditions individual voting rules fall victim to various voting irregularities [7, 12]. 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 [21,24]. 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 [23].

A fundamental problem in research into properties of voting rules is the lack of large data sets to run empirical experiments [19,23]. There have been studies of some datasets but these are limited in both number of elections analyzed [7] and size of individual elections within the datasets analyzed [12, 23]. While there is little agreement about the frequency that voting paradoxes occur or the consensus between voting methods, all the studies so far have found little evidence of *Condorcet's Voting Paradox* [13] (a cyclical majority ordering) or *preference domain restrictions* such as *single peakedness* [5] (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 [7,12,19].

As the computational social choice community continues to grow there is increasing attention on empirical results (see, e.g., [24]). The empirical data will support and justify the theoretical concerns [10, 11]. Walsh explicitly called for the establishment of a repository of voting data in his COMSOC 2010 talk [25]. We begin to respond to this call through the identification, analysis, and posting of a new repository of voting data.

We evaluate a large number of distinct 3 and 4 candidate elections derived from a novel data set, under the voting rules: Plurality, Copeland, Borda, Repeated Alternative Vote, and k-Approval. Our research question is manifold: Do different voting rules often produce the same winner? How often does Condorcet's Voting Paradox occur? Do basic statistical models of voting accurately describe our domain? Do any of the votes we analyze show single-peaked preferences [5] or other domain restrictions [22]?

2 Related Work

The literature on the empirical analysis of large voting datasets is somewhat sparse and many studies use the same datasets [12, 23]. These problems can be attributed to the lack of large amounts of data from real elections [19]. Chamberlin et al. [7] provide 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. analyse APA data from later years [20] and observe the same phenomena: a high degree of stability between elections rules. Felsenthal et al. [12] analyze 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 find a high degree of consensus between voting rules (with the notable exception of Plurality).

All of the empirical studies surveyed [7, 12, 16, 19, 20, 23] come to a similar conclusion: that there is scant evidence for occurrences of Condorcet's Paradox [18]. Many of

these studies find no occurrence of majority cycles (and those that find cycles find them in rates of 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 homogenous population [16]) find almost no occurrences of either single-peaked preferences [5] or the more general value restricted preferences [22].

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., [19, 23]). Gehrlein [13] 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 [13]

Tideman and Plassmann have undertaken the task of verifying the statistical cultures used to generate pseudo-election data [23]. Using one of the largest datasets available Tideman and Plassmann find little evidence supporting the models currently in use to generate election data. Regenwetter et al. undertake a similar exercise and also find small support for the existing models of election generation [19]. The studies by both Regenwetter et al. and Tideman and Plassmann propose new statistical models with which to generate election pseudo-data that are better fits for their respective datasets.

3 The Data

We have mined strict preference orders from the Netflix Prize Dataset [3]. The Netflix dataset offers a vast amount of preference data; compiled and publically released by Netflix for its Netflix Prize [3]. 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 2004) 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; this is a random, sequential number 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 [23] 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 [23] and we use it to generate comparable results.

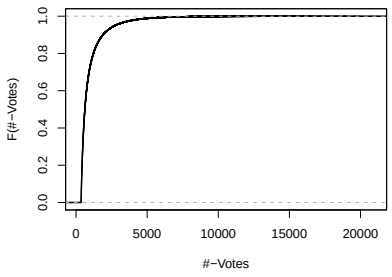

Fig. 1. Empirical CDF of Set 3A.

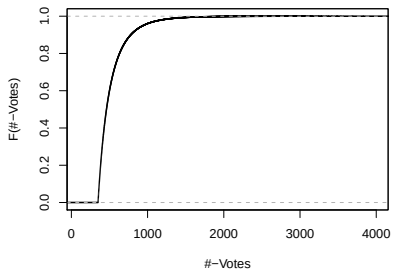

Fig. 2. Empirical CDF of Set 4A.

The dataset is too large to use completely ($\binom{17770}{3} \approx 1 \times 10^{12}$). Therefore, we have drawn 3 independent (non-overlapping with respect to movies) samples of 2000 movies randomly from the set of all movies. We then, for each sample, search all the $\binom{2000}{3} \approx 1.33 \times 10^9$ possible elections for those with more than 350 votes. This search generated 1,553,611, 1,331,549, and 2,049,732 distinct movie triples within each of the respective samples. 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 22,079 votes; metrics of central tendency are presented in Table 1. Figures 1 and 2 show the empirical cumulative distribution functions (ECFD) for Set3A and 4A respectively. All of the datasets show similar ECDF's to those pictured.

Using the notion of item-item extension [14] 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

	3 Candidate Sets			4 Candidate Sets		
	Set 3A	Set 3B	Set 3C	Set4A	Set 4B	Set 4C
Min.	350.0	350.0	350.0	350.0	350.0	350.0
1st Qu.	444.0	433.0	435.0	394.0	393.0	384.0
Median	617.0	579.0	581.0	461.0	461.0	438.0
Mean	963.8	881.8	813.4	530.9	530.5	494.6
3rd Qu.	1,041.0	931.0	901.0	588.0	591.0	539.0
Max.	22,079.0	18,041.0	20,678.0	3830.0	3396.0	3639.0
Elements	1,553,611.0	1,331,549.0	2,049,732.0	2,721,235.0	1,222,009.0	1,243,749.0

Table 1. Summary Statistics for the election data.

which was a valid voting instance. For each set we only searched for extensions within the same draw of 2000 movies, making sure to remove any duplicate 4-item extensions. The results of this search are also summarized in Table 1. We found no 5-item extensions with more than 350 votes in the >30 billion possible extensions. Our constructed dataset 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 distinct election.

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 24 hours (parallelized over the two cores) for each of the three independently drawn sets. The four movie extension searches took approximately 168 hours per dataset while the five movie extensions took about 240 hours per dataset. 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 15 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.

4 Analysis and Discussion

We have found a large correlation between each of the 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 [12]. 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 [18]. 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]. 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. In Section 4.2 we see that

the voting rules exhibit a high degree of Condorcet Efficiency in our dataset. Finally, the experiments in Section 4.3 indicate that several statistical models currently in use for testing new voting rules [21] do not reflect the reality of our dataset. All of these results are in keeping with the analysis of other, distinct, datasets [7, 12, 16, 19, 20, 23] and provide support for their conclusions.

4.1 Domain Restrictions and 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 [18]. 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 [12].

		Partial Cycle	Partial Top	Total
$m = 3$	Set 3A	635 (0.041%)	635 (0.041%)	635 (0.041%)
	Set 3B	591 (0.044%)	591 (0.044%)	591 (0.044%)
	Set 3C	1,143 (0.056%)	1,143 (0.056%)	1,143 (0.056%)
$m = 4$	Set 4A	3,837 (0.141%)	2,882 (0.106%)	731 (0.027%)
	Set 4B	1,864 (0.153%)	1,393 (0.114%)	462 (0.035%)
	Set 4C	3,233 (0.258%)	2,367 (0.189%)	573 (0.046%)

Table 2. Number of elections demonstrating various types of voting cycles.

Table 2 is a summary of the rates of occurrence of the different types of voting cycles found in our data set. 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 ($< 0.06\%$ 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 $< 0.2\%$ and therefore seem of little practical importance in our dataset as well. Our results for cycles that do not include the winner mirror those of Felsenthal et al. [12]: 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.

Black first introduced the notion of single-peaked preferences [5]; 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, it is the idea that some candidate, in a three candidate election, is never ranked last. The notion of restricted preference profiles was extended by Sen [22]

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 trivially easy to manipulate when electorates preferences are single-peaked [6].

		Single-Peak	Single-Mid	Single-Bottom
$m = 3$	Set 3A	342 (0.022%)	0 (0.0%)	198 (0.013%)
	Set 3B	227 (0.017%)	0 (0.0%)	232 (0.017%)
	Set 3C	93 (0.005%)	0 (0.0%)	100 (0.005%)
$m = 4$	Set 4A	1 (0.022%)	0 (0.000%)	1 (0.013%)
	Set 4B	0 (0.000%)	0 (0.000%)	0 (0.000%)
	Set 4C	0 (0.000%)	0 (0.000%)	0 (0.000s%)

Table 3. Number of elections demonstrating various value restricted preferences.

Table 3 summarizes our results for the analysis of different restricted preference profiles. There is (nearly) a complete lack of preference profile restrictions when $m = 4$ and near lack ($< 0.03\%$) 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.[1]

4.2 Voting Rules

The variety of voting rules and election models that have been implemented or "improved" over time is astounding. For a comprehensive history and survey of voting rules see Nurmi [18]. 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 [18] and in Arrow, Sen, and Suzumura [1].

Plurality: Plurality is the most widely used voting rule [18] (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

[1] Set 3B contains the movies *Star Wars: Return of the Jedi* and *The Shawshank Redemption*. Both are widely considered to be "good" movies; all but 15 of the 227 elections exhibiting single-peaked preferences share one of these two movies.

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 [18] the first place candidate receives $c-1$ points, with each candidate later in the ranking receiving 1 less point 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 [12]. 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 is repeated up to $c-1$ times [12]. 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. [12]. 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 [12, 18]. 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 [12]. This procedure has the disadvantage of demonstrating if voting rules fail to correspond closely to the results from Copeland. Another method, not used in this paper, would be to consider each of the voting rules as a maximum likelihood estimator of some "ground truth." We leave this track for future work [9].

Table 4 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 [7, 12, 20]. 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. Unfortunately, our data is not split into enough independent samples to accurately perform any statistical hypothesis testing. Computing a paired t-test with all $> 10^6$ elections within a sample set would provide trivially significant results due to the extremely large sample size.

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]. Table 5 shows the proportion of Condorcet Winners selected

		Plurality	2-Approval	Borda	RAV
	Mean	0.9300	0.9149	0.9787	0.9985
Set 3A	SD	0.1999	0.2150	0.1029	0.0336
	Mean	0.9324	0.9215	0.9802	0.9985
Set 3B	SD	0.1924	0.2061	0.0995	0.0341
	Mean	0.9238	0.9177	0.9791	0.9980
Set 3C	SD	0.208	0.2130	0.1024	0.0394
	Mean	0.9053	0.9578	0.9787	0.9978
Set 4A	SD	0.1691	0.0956	0.0673	0.0273
	Mean	0.9033	0.9581	0.9798	0.9980
Set 4B	SD	0.1627	0.0935	0.0651	0.0263
	Mean	0.8708	0.9516	0.9767	0.9956
Set 4C	SD	0.2060	0.1029	0.0706	0.0404

Table 4. Voting results (Spearman's ρ) for Sets A,B, and C.

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. 2-Approval, when $m = 3$, results in the lowest rate of Condorcet Winner selection in our dataset.

		Condorcet Winners	Plurality	2-Approval	Borda	RAV
	Set 3A	1,548,553	0.9665	0.8714	0.9768	0.9977
$m = 3$	Set 3B	1,326,902	0.9705	0.8842	0.9801	0.9980
	Set 3C	2,041,756	0.9643	0.8814	0.9795	0.9971
	Set 4A	2,701,464	0.9591	0.9213	0.9630	0.9966
$m = 4$	Set 4B	1,212,370	0.9626	0.9290	0.9693	0.9971
	Set 4C	1,241,762	0.9550	0.9253	0.9674	0.9940

Table 5. Condorcet Efficiency of the various voting rules.

Overall, we find a consensus between the various voting rules in our tests. This supports the findings of other empirical studies in the field [7,12,20]. Merrill finds much different 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 [13].

4.3 Statistical Models of Elections

We evaluate our dataset to see how it matches up to different probabilistic 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 [23]. There are other election generating cultures that we do not analyze because we found no support for restricted preference profiles (either single-peaked or single-bottomed). These cultures, such as weighted Independent Anonymous Culture, generate preference profiles that are skewed towards single-peakedness or single-bottomness (a further discussion and additional election generating statistical models can be found in [23]). We follow the general outline in Tideman and Plassmann to guide us in this study. 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 $1/m!$

Dual Culture (DC): The dual culture assumes that the probability of opposite preference orders is equal. So, $pr(ABC) = pr(CAB)$, $pr(ACB) = pr(BCA)$ 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 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 [23]: 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 [4]. We use a setup as described by Walsh [24]; 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 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 is equally likely. 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 1,639,070 and 1,718,532 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
$m = 3$	Set 3A	0.3304 (0.0159)	0.2934 (0.0126)	0.1763 (0.0101)	0.3025 (0.0372)
	Set 3B	0.3192 (0.0153)	0.2853 (0.0121)	0.1685 (0.0095)	0.2959 (0.0355)
	Set 3C	0.3041 (0.0151)	0.2709 (0.0121)	0.1650 (0.0093)	0.2767 (0.0295)
$m = 3$	Urn	0.6226 (0.0249)	0.4744 (0.0225)	0.4743 (0.0225)	0.4909 (0.1054)
	IAC	0.2265 (0.0056)	0.1690 (0.0056)	0.1689 (0.0056)	0.2146 (0.0063)
	IAC-Fit	0.0372 (0.0002)	0.0291 (0.0002)	0.0260 (0.0002)	0.0356 (0.0002)
$m = 4$	Set 4A	0.2815 (0.0070)	0.2282 (0.0042)	0.1141 (0.0034)	0.3048 (0.0189)
	Set 4B	0.2596 (0.0068)	0.2120 (0.0041)	0.1011 (0.0026)	0.2820 (0.0164)
	Set 4C	0.2683 (0.0080)	0.2149 (0.0049)	0.1068 (0.0034)	0.2811 (0.0166)
$m = 4$	Urn	0.6597 (0.0201)	0.4743 (0.0126)	0.4743 (0.0126)	0.6560 (0.1020)
	IAC	0.1257 (0.0003)	0.0899 (0.0003)	0.0899 (0.0003)	0.1273 (0.0004)
	IAC-Fit	0.0528 (0.0001)	0.0415 (0.0001)	0.3176 (0.0001)	0.0521 (0.0001)

Table 6. Mean Euclidean distance between the empirical data set and different statistical cultures (standard error in parentheses).

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. UC has the lowest error in predicting the distributions found in our empirical data. 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 IC. We, like Tideman and Plassmann, find little support for the static cultures' ability to model real data [23]

5 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; 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 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.

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. A clearing house for data was discussed at COMSOC 2010 by Toby Walsh and others in attendance [25]. 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.

There are many directions for future work that we would like to explore. We plan to evaluate how many of the elections in our data set are manipulable and evaluate the frequency of occurrence of easily manipulated elections. We would like to, instead of comparing how voting rules correspond to one another, evaluate their power as maximum likelihood estimators [9]. Additionally, we would like to expand our evaluation of statistical models to include several new models proposed by Tideman and Plassmann, and others [23].

Acknowledgements Thanks to Dr. Florenz Plassmann for his helpful discussions on this paper and guidance on calibrating statistical models. Also thanks to Dr. Judy Goldsmith and Elizabeth Mattei for their helpful discussion and comments on preliminary drafts of this paper. We gratefully acknowledge the support of NSF EAGER grant CCF-1049360.

References

1. Arrow, K., Sen, A., Suzumura, K. (eds.): Handbook of Social Choice and Welfare, vol. 1. North-Holland (2002)

2. Arrow, K.: Social choice and individual values. Yale Univ Press (1963)
3. Bennett, J., Lanning, S.: The Netflix Prize. In: Proceedings of KDD Cup and Workshop (2007), www.netflixprize.com
4. Berg, S.: Paradox of voting under an urn model: The effect of homogeneity. Public Choice 47(2), 377–387 (1985)
5. Black, D.: On the rationale of group decision-making. The Journal of Political Economy 56(1) (1948)
6. Brandt, F., Brill, M., Hemaspaandra, E., Hemaspaandra, L.A.: Bypassing combinatorial protections: Polynomial-time algorithms for single-peaked electorates. In: Proc. of the 24th AAAI Conf. on Artificial Intelligence. pp. 715 – 722 (2010)
7. Chamberlin, J.R., Cohen, J.L., Coombs, C.H.: Social choice observed: Five presidential elections of the American Psychological Association. The Journal of Politics 46(2), 479 – 502 (1984)
8. Condorcet, M.: Essay sur l'application de l'analyse de la probabilit des decisions: Redues et pluralit des voix. Paris (1785)
9. Conitzer, V., Sandholm, T.: Common voting rules as maximum likelihood estimators. In: Proc. of the 21st Annual Conf. on Uncertainty in AI (UAI). pp. 145–152 (2005)
10. Conitzer, V., Sandholm, T., Lang, J.: When are elections with few candidates hard to manipulate? Journal of the ACM 54(3), 1 – 33 (June 2007)
11. Faliszewski, P., Hemaspaandra, E., Hemaspaandra, L.A., Rothe, J.: A richer understanding of the complexity of election systems. In: Ravi, S., Shukla, S. (eds.) Fundamental Problems in Computing: Essays in Honor of Professor Daniel J. Rosenkrantz, pp. 375 – 406. Springer (2009)
12. Felsenthal, D.S., Maoz, Z., A, R.: An empirical evaluation of six voting procedures: Do they really make any difference? British Journal of Political Science 23, 1 – 27 (1993)
13. Gehrlein, W.V.: Condorcet's paradox and the likelihood of its occurance: Different perspectives on balanced preferences. Theory and Decisions 52(2), 171 – 199 (2002)
14. Han, J., Kamber, M. (eds.): Data Mining. Morgan Kaufmann (2006)
15. Merrill, III, S.: A comparison of efficiency of multicandidate electoral systems. American Journal of Politial Science 28(1), 23 – 48 (1984)
16. Niemi, R.G.: The occurrence of the paradox of voting in university elections. Public Choice 8(1), 91–100 (1970)
17. Nisan, N., Roughgarden, T., Tardos, E., Vazirani, V. (eds.): Algorithmic Game Theory. Cambridge Univ. Press (2007)
18. Nurmi, H.: Voting procedures: A summary analysis. British Journal of Political Science 13, 181 – 208 (1983)
19. Regenwetter, M., Grogman, B., Marley, A.A.J., Testlin, I.M.: Behavioral Social Choice: Probabilistic Models, Statistical Inference, and Applications. Cambridge Univ. Press (2006)
20. Regenwetter, M., Kim, A., Kantor, A., Ho, M.R.: The unexpected empirical consensus among consensus methods. Psychological Science 18(7), 629 – 635 (2007)
21. Rivest, R.L., Shen, E.: An optimal single-winner preferential voting system based on game theory. In: Conitzer, V., Rothe, J. (eds.) Proc. of the 3rd Intl. Workshop on Computational Social Choice (COMSOC). pp. 399 – 410 (2010)
22. Sen, A.K.: A possibility theorem on majority decisions. Econometrica 34(2), 491–499 (1966)
23. Tideman, N., Plassmann, F.: Modeling the outcomes of vote-casting in actual elections. To appear in Springer publshed book, http://bingweb.binghamton.edu/~fplass/papers/Voting_Springer.pdf
24. Walsh, T.: An empirical study of the manipulability of single transferable voting. In: Proc. of the 19th European Conf. on AI (ECAI 2010). pp. 257–262. IOS Press (2010)
25. Walsh, T.: Where are the hard manipulation problems? In: Conitzer, V., Rothe, J. (eds.) Proc. of the 3rd Intl. Workshop on Computational Social Choice (COMSOC). pp. 9 – 11 (2010)