Collecting, Classifying, Analyzing, and Using Real-World Ranking Data

Niclas Boehmer
Technische Universität Berlin
Berlin, Germany
niclas.boehmer@tu-berlin.de

Nathan Schaar
Technische Universität Berlin
Berlin, Germany
n.schaar@campus.tu-berlin.de

ABSTRACT

We present a collection of 7582 real-world elections divided into 25 datasets from various sources ranging from sports competitions over music charts to survey- and indicator-based rankings. We provide evidence that the collected elections complement other publicly available data from the PrefLib database [47]. Using the map of elections framework [66], we divide the datasets into three categories and conduct an analysis of the nature of our elections. To evaluate the practical applicability of previous theoretical research on (parameterized) algorithms and to gain further insights into the collected elections, we analyze different structural properties of our elections including the level of agreement between voters and election's distances from restricted domains such as single-peakedness. Lastly, we use our diverse set of collected elections to shed some further light on several traditional questions from social choice, for instance, on the number of occurrences of the Condorcet paradox and on the consensus among different voting rules.

KEYWORDS

elections; data generation; empirical analysis; map of elections; Kemeny score; restricted domains; consensus among voting rules

ACM Reference Format:
Niclas Boehmer and Nathan Schaar. 2023. Collecting, Classifying, Analyzing, and Using Real-World Ranking Data. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 10 pages.

1 INTRODUCTION

The area of computational social choice is concerned with the algorithmic and axiomatic analysis of collective decision-making problems, where given a set of agents with preferences over some alternatives the task is to select a "compromise" alternative [13]. One important part of computational social choice is the study of algorithmic aspects of election-related problems such as the computation and manipulation of voting rules [17, 19, 33, 43, 73]. While in the early years of the field the main focus lay on the study of the theoretical worst-case computational complexity of these problems, in recent years the focus has at least partially shifted towards the practical applicability of theoretical research (see e.g., [6, 34, 37, 41, 66, 68]). Two classical social choice questions which have been studied from an empirical perspective are the number of occurrences of voting paradoxes [14, 18, 35, 57] and the consensus among voting rules [18, 21, 35, 46, 58, 60, 61]. Nevertheless, there are still many subareas that lack empirical research. For instance, there are numerous theoretical papers designing parameterized algorithms for elections that are close to being single-peaked[1] (see e.g. [20, 30, 50, 64, 70–72]) with only Sui et al. [65] measuring the distance of real-world elections from being single-peaked and detecting that most elections are far away. Thus, the practical applicability of the developed algorithms is largely unclear.

One reason for the general rarity of experimental works in voting validating the applicability of theoretical research might be the lack of data. To tackle this issue, in 2013, Mattei and Walsh [47, 48] started the very useful PrefLib platform, a database for real-world election data. Many community members have contributed to this popular platform over the past years and before adding our data to it, PrefLib contained 701 real-world elections dived into 36 datasets (see Boehmer et al. [7, Table 5] for a recent overview of the datasets). Many elections from PrefLib are based on humans expressing opinions over alternatives, e.g., over candidates in an election, over movies, or types of sushi. However, due to this nature of these elections, most of them either have few candidates or voters express only partial preferences which can include many ties. In fact, as observed by Boehmer et al. [7, Table 5], there are only 165 elections from 8 sources on PrefLib with 10 or more candidates where votes include not too many ties. The goal of this paper is to contribute to the rise of experimental works in computational social choice by executing the following four steps:

Step 1: Collecting Data. In Section 3, we present our collection of 7582 real-world elections divided into 25 datasets. We preprocess the data by deleting candidates and voters until each voter ranks all candidates. Subsequently, to be able to better compare the properties of our elections, for each dataset we create 500 elections containing 30 voters over 15 candidates. Our real-world elections differ from most of the already publicly available ones in three aspects: First, they contain virtually no ties and are of various sizes (the average number of candidates varies from around 20 to above 800, while the average number of voters ranges from around 12 to over 1400). Moreover, even after deleting voters and candidates until all voters rank all candidates, most elections are still of at least medium size. As a majority of algorithms are designed for such so-called *complete* elections, this is a very important step to ensure the usefulness of our data for experimental works. In the past, elections have been often completed by appending missing candidates in random order or based on the preference of other voters [7, 22]. Our approach offers the clear advantage that preferences in the final election are not distorted in any way: Each pairwise ordering

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

[1] An election is single-peaked if there exists a societal order of the candidates and each voter ranks candidates that are closer to its top-choice according to the societal order above those which are further away.

of a voter represents its true opinion. Second, unlike a majority of elections on PrefLib, our datasets are not based on humans explicitly expressing preferences over alternatives. Traditionally, this might be considered as a drawback of our data as political elections are still often thought of as the prime application of social choice theory. However, we want to remark that this is no longer true, as voting is also relevant and already used in many other contexts, e.g., in multi-agent and recommender systems (as witnessed by social choice being an AAMAS area), or in sports, when aggregating the results of multiple competitions into a final ranking. Third, around half of our datasets arise from time-based preferences, i.e., capture in one form or another the changing preferences of agents over time. Time-based elections might not directly match ones intuition for an election; however, preferences obtained at different points in time are also frequently collected in an election (for instance, when deciding on the overall winner of multiple competitions). Notably, while there are already some theoretical works dealing with such time-evolving preferences [10, 16, 42, 56], as pointed out by Mattei and Walsh [48], there are only very few such elections currently publicly available.

Step 2: Classifying Data. In Section 4, we apply the map of elections framework of Szufa et al. [66] and Boehmer et al. [8] to visualize the collected elections as points on a map. Using this, we detect that one of our datasets seems to fall into a so-far vacant part of the "space of elections". Moreover, based on their positions on the map, we propose a classification of our datasets into three categories and observe in the subsequent experiments that datasets from one category typically have similar properties. This suggests that if one wants to run experiments on our data, it should be sufficient to use few datasets from each of the three categories.

Step 3: Analyzing Data. In Sections 5 and 6, we analyze various structural properties of the collected elections. This analysis serves three purposes: First, we aim for a better understanding of the collected elections. Second, we want to gain some insights into the relationship between the different properties. Third, we try to contribute to putting the research on parameterized algorithms for voting-related problems on an empirical basis by measuring already used parameters. Unfortunately, we find that most of them are typically quite large and thus that most algorithms developed for these parameters are probably not really practically usable on our data. Briefly put, in Section 5 we analyze the degree of similarity between voters in an election, while in Section 6 we check which of our elections are (close to) a restricted domain.

Step 4: Using Data. In Section 7, we use our collected elections to address some classical and already empirically researched questions from social choice, such as the frequency of Condorcet winners and the consensus among voting rules. While we partly confirm previous findings, for instance, that most elections have a Condorcet winner and that voting rules often return the same winner, we find contradicting evidence for others and also identify some datasets showing a distinct behavior. This indicates that our datasets are quite different from each other with some of them showing rarely observable and non-standard behavior, making them collectively well-suited for experimental research.

The full version of this paper containing additional discussions and experiments is available at arxiv.org/pdf/2204.03589.pdf [11]. The collected datsets are available at github.com/n-boehmer and preflib.org/BoSc22. We also collected some further datasets which we do not include in our analysis for the sake of conciseness.

2 PRELIMINARIES

For a set S and an integer $k \in \mathbb{N}$, we denote as $\binom{S}{k}$ the set of all k-element subsets of S. For a set C of candidates, let $\mathcal{L}(C)$ denote the set of all total orders over C. We refer to the elements of $\mathcal{L}(C)$ as preference orders, votes, or voters. An election E is defined by a set of $C = \{c_1, \dots, c_m\}$ of m candidates and a collection $V = (v_1, \dots, v_n)$ of n voters with $v_i \in \mathcal{L}(C)$ for each $i \in [n]$. For a voter $v \in V$ and two candidates $a, b \in C$, we write $a \succ_v b$ to denote that v prefers a to b. We say that voter $v \in V$ ranks candidate $c \in C$ in position $i \in [m]$ if v prefers exactly $i-1$ candidates from $C \setminus \{c\}$ to c. We refer to the candidate which a voter ranks in the first position as its top-choice. For two votes $v, v' \in \mathcal{L}(C)$, their Kendall tau distance $\mathrm{KT}(v, v')$ is defined as the number of candidate pairs on which orderings v and v' disagree: $|\{c, c' \in \binom{C}{2} \mid (c \succ_v c' \wedge c' \succ_{v'} c) \vee (c \succ_{v'} c' \wedge c' \succ_v c)\}|$. Alternatively, $\mathrm{KT}(v, v')$ can be interpreted as the minimum number of swaps of adjacent candidates that need to be performed to transform v into v'.

Next, we define three different restricted domains. In single-peaked elections, there is an order of the candidates and each voter prefers candidates that are closer to its top-choice with respect to the order to those further away: Formally, an election $E = (C, V)$ is *single-peaked* [3] if there is a linear order $\rhd$ over C, sometimes called the societal order, such that for each three candidates $a, b, c \in C$ with $a \rhd b \rhd c$, for each $v \in V$, if $a \succ_v b$ then $b \succ_v c$. In single-crossing elections, there is an order of voters such that going through the voters according to the order, the ordering of each pair of candidates changes at most once: Formally, an election $E = (C, V)$ is *single-crossing* [51, 62] if there is a linear order $\rhd$ over V such that for each two candidates $c, c' \in C$, there do not exist three votes $v, v', v'' \in V$ with $v \rhd v' \rhd v''$ such that $c \succ_v c'$, $c' \succ_{v'} c$, and $c \succ_{v''} c'$. Lastly, an election $E = (C, V)$ is *group-separable* [38, 39] if each subset $A \subseteq C$ of candidates with $|A| \geq 2$ can be partitioned into two sets A' and A'' such that each voter $v \in V$ prefers either all candidates from A' to all candidates from A'' or the other way around.

3 COLLECTING REAL-WORLD ELECTIONS

In the following, we list the different data sources that we used to create our elections, ranging from results of sports competitions over music charts and expert assessments to survey- or indicator-based rankings. For each data source, we describe how we created elections from the data (if there happens to be a tie, we break it arbitrarily); for some sources, we created two types of elections.

From a methodological perspective, our elections are of one of two types: We say that an election is *time-based* if each vote corresponds to an evaluation of the candidates at different points in time. In contrast to this, we call an election *criterion-based* if each vote corresponds to some, in principle, independent criterion judging the candidates at the same point in time. In Table 1, we indicate for each dataset the type, the number of contained elections, and their average size before and after the preprocessing.

Boxing/Tennis (World) Rankings. The boxing data (collected by Jürisoo [40]) contains the Ultimate Fighting Championship rankings of the top 16 fighters in twelve different weight classes in different weeks between February 2013 and August 2021. The tennis data (collected by Wang [69]) contains weekly rankings of the top 100 male tennis players published by the ATP between January 1990 and September 2019. For each year (and weight class), we created a *tennis top 100 (boxing top 16)* election where each player (fighter) is a candidate and each vote corresponds to the ranking of the players (fighters) in one week.

American Football. The American football data (collected by Massey [45]) contains weekly power rankings of college football teams from different media outlets for each season between 1997 and 2021. We created two different types of elections with teams as candidates: First, for each season and each media outlet, we created a *football season* election where each vote corresponds to the power ranking of the teams in one week according to the media outlet. Second, for each week in one of the seasons, we created a *football week* election where each vote corresponds to the power ranking of the teams in this week according to one of the media outlets.

Formula 1. The Formula 1 data (collected by Rao [59]) contains the finishing times of each driver in each lap of a race between 1950 and 2020. From this we created two types of elections with drivers as candidates: First, for each year, we created a *Formula 1 season* election where each vote corresponds to a race in this year and ranks the drivers by their finishing time in this race. Second, for each race, we created a *Formula 1 race* election where each vote corresponds to a lap in the race and ranks the drivers by the time they spend in this lap.

Spotify. For each day between the 1st of January 2017 and 9th January 2018, the Spotify data (collected by Oliveira [53]) contains a daily ranking of the 200 most listened songs in one of 53 countries. We created two types of elections with songs as candidates: First, for each month and each country, we created a *Spotify month* election where each vote corresponds to the ranking of the songs on one day of the month in the country. Second, for each day, we created a *Spotify day* election where each vote corresponds to the ranking of the songs on this day in one of the 53 countries.

Tour de France. For each edition of the Tour de France between 1903 and 2021, the data contains the completion times of all riders for each stage. The dataset was crawled by us from the website procyclingstats.com. For each edition, we created one *Tour de France* election in which the riders are the candidates and each vote corresponds to a stage and ranks the riders by their completion time.

City Rankings. The city data (collected by Blitzer [4]) contains twelve quantitative indicators for the life quality in 216 different cities determined by movehub.com. We created a single *city ranking* election where each city is a candidate and each vote corresponds to the ranking of the cities with respect to one of the indicators.

Country Rankings. For each year between 2005 and 2016, the country ranking data (based on the popular world happiness report and collected by Oxa [55]) contains different quantitative indicators for the happiness of citizens from over 100 countries. For each year, we created a *country ranking* election where the countries are the candidates and each vote ranks them according to one indicator.

name	type	raw			relevant complete		
		#Elec.	Avg. #Voters	Avg. #Cand.	#Elec.	Avg. #Voters	Avg. #Cand.
boxing top 16	time	99	31.9	19.76	31	17.45	15.32
football season	time	2746	12.28	152.36	2422	12.6	156.71
Formula 1 race	time	454	61.3	20.46	396	47.2	17.93
Formula 1 season	time	71	14.58	43.97	42	13.38	21.57
Spotify month	time	645	29.78	306.64	632	29.91	109.28
tennis top 100	time	29	50.48	140	29	49.9	62.31
Tour de France	time	97	21.14	175.69	95	19.7	82.64
city ranking	crit.	1	12	216	1	12	216
country ranking	crit.	12	17.25	119.17	12	14.25	95.58
football week	crit.	415	83.28	219.67	415	77.35	98.45
Spotify day	crit.	362	53.06	247.74	375	49.06	20.73
university ranking	crit.	4	18.5	832.5	4	18.5	123.25

Table 1: Information about our election datasets.

University Rankings. For each year between 2012 and 2015, the university ranking data (collected by O'Neill [54]) contains rankings of universities according to different criteria provided by three systems. For each year, we created a *university ranking* election where the universities are the candidates and each vote ranks them according to one criterion used by one of the three systems.

From Raw to Normalized Elections

In our experiments, we do not use the raw elections created as described above but instead apply some postprocessing. As a first step, by deleting voters and candidates, we converted each created election into a *complete* election, i.e., an election where every voter ranks all candidates (see our full version for more details [11]). As in our experiments we are interested in elections with at least 15 candidates[2], we call each election with 15 or more candidates (and an arbitrary number of voters) *relevant.* We display information about the number and size of the relevant complete elections from each dataset in Table 1.

Next, similarly as done by Boehmer et al. [8], to be able to meaningfully compare the results of our experiments within datasets and between datasets, we created *normalized* elections. For each dataset, we created 500 elections with 15 candidates and 30 voters as follows. To create an election $E = (C, V)$, we uniformly at random selected one relevant complete election $F = (D, W)$ from the respective dataset. Subsequently, we sampled a subset of 15 candidates C uniformly at random from D. After that, to create V, we sampled 30 times a vote uniformly at random from W with replacement. This means that a vote from W can occur potentially multiple times in V and that different normalized elections might be based on F. In all our experiments presented in the following sections we only use normalized elections and will no longer explicitly specify this. We refer to the dataset containing all normalized elections from all datasets as the *aggregated* dataset.

[2] We chose this number to be as large as possible while still being able to include most of our elections.

4 DRAWING A MAP OF OUR ELECTIONS

To get a feeling for the type of our elections and to be able to better relate the datasets to each other, we apply the "map of elections" framework. In this framework, which has been developed by Szufa et al. [66] and Boehmer et al. [8] (see also [5, 9]), we take a set of elections and compute for each pair their so-called "positionwise" distance.[3] Afterward, using the embedding algorithm from Fruchterman and Reingold [36], we draw a map of our elections where each election is represented by a dot with the Euclidean distance between two dots being as similar as possible to the distance between the respective two elections. Note that the position of an election on the map thus naturally depends on the set of depicted elections.

To give a meaning to the absolute position of an election on the map, Boehmer et al. [8] introduced what they call a compass consisting of four types of "extreme" elections capturing different kinds of (dis)agreement between voters and their convex combinations:

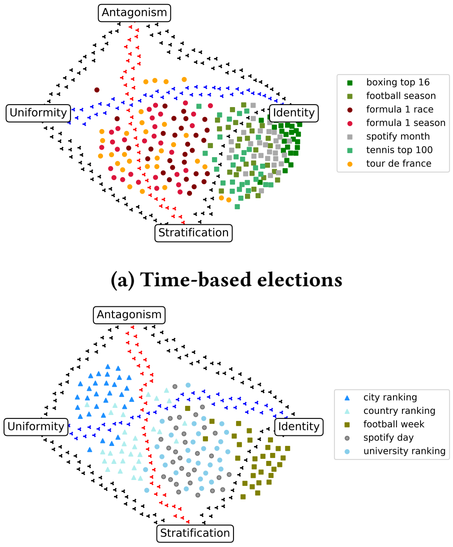
Identity All voters have the same preference order.

Uniformity Each possible preference order appears exactly once.

Antagonism Half of the voters rank the candidates in the same order, while the other half ranks them in the opposite order.

Stratification There is a partitioning of the candidates into two sets A and B of equal size and all possible preference orders where all candidates from A are ranked before those from B appear once.

Setup. We created two maps of elections (Figure 1) where each election is represented by a point whose shape and color indicate the dataset to which it belongs. To make the created maps not too crowded, we created a separate map for time-based (Figure 1a) and criterion-based (Figure 1b) elections. For each map, we included 30 elections sampled uniformly at random from each normalized dataset and the compass elections introduced by Boehmer et al. [8] together with their convex combinations appearing as "paths".



(a) Time-based elections

(b) Criterion-based elections

Figure 1: Visualization of our elections as map of elections.

Classifying Datasets. Examining Figure 1, it is possible to divide our datasets into three groups: The first group of datasets (boxing top 16, football season, Spotify month, tennis top 100, and football week) drawn as squares all contain elections somewhat close to identity. Notably, except for football week[4], these are all time-based datasets. For all of them except Spotify month, the ranking at a certain point in time partly depends on information on candidates that also already influenced previous votes. As a result, in some sense, votes are "by design" not independent here.[5] In contrast to this, in time-based elections from the other datasets (Formula 1 race, Formula 1 season, and Tour de France), which do not belong to this group and are further away from identity, one vote only depends on the performance of a candidate at some point in time (and not on previous performances).

The second group of datasets (Formula 1 race, Formula 1 season, Tour de France, Spotify day, and university rankings) drawn as circles constitute the "middle" part of our maps: This is also reflected in them being roughly at the same distance from identity and uniformity (while all are clearly closer to stratification than to antagonism). What is particularly striking here is that despite the fact that these elections are seemingly not all simply close to a canonical extreme election like identity, there are surprising similarities between the datasets: In particular, university, Formula 1 race, and Spotify day elections all fall in exactly the same area of the space of elections (the average distance of two elections from one of these datasets is very close to the average distance of two elections picked from two different of these datasets). The same also holds for Tour de France and Formula 1 season elections. Remarkably, Tour de France and Formula 1 season elections are also by design of a very similar nature in the sense that in both datasets players compete in a similar task on different days. The similarity of these datasets indicates that whether players drive in cars or ride bicycles seems to be not so crucial for the resulting election (similar observations apply to boxing top 16 and tennis top 100, and city rankings and country rankings).

The third group of datasets consists of city and country rankings and is drawn as triangles. Both are clearly different from the rest as they are significantly closer to uniformity than identity. Remarkably, the city ranking dataset is the only one of our datasets and the first known dataset which is significantly closer to antagonism (distance 29) than stratification (distance 43). Considering the underlying data which provides ratings of cities according to different indicators, the "closeness" to antagonism is quite plausible, as some of the studied indicators seem to capture in some sense contradicting objectives, e.g., big cities where inhabitants typically have access

[3] The positionwise distance is based on the notion of frequency matrices. In the frequency matrix of an election, each column corresponds to a candidate and each row to a position and an entry captures the fraction of voters ranking the respective candidate in the respective position. The distance between two elections then corresponds to the summed earth mover's distance between the columns of their frequency matrices with columns being rearranged to minimize this distance (see [8, 66] for details).

[4] Recalling that in football week elections the strength of college football teams at one point are judged by different systems (votes), it is also quite intuitive that these elections are close to identity, as one could argue that there exists a "ground truth".

[5] For Spotify month this is not really the case "by design". However, also here similar effects are present. E.g. users often listen to playlists that only change slowly over time, implying that what users listened to on one day in some sense "predicts" what they will listen to on the next day.

to a variety of healthcare facilities (being one of the indicators) are typically also quite polluted (being another indicator).

Captured Part of the Space of Elections. It seems that our datasets contain elections of a different nature than those available on PrefLib: Boehmer et al. [8, Figure 2b] drew a map of elections including representatives of all PrefLib datasets with at least 10 candidates, 10 votes, and not too many ties. They observed that most elections are closer to uniformity than identity and closer to stratification than antagonism, thereby ending up in the bottom left quadrant of the map. In contrast to this, our elections are mostly located in the bottom right quadrant. Nevertheless, we can confirm the observation of Boehmer et al. [8] that real-world elections typically end up closer to stratification than antagonism (we also do not provide any elections that are in the top right quadrant).

5 SIMILARITY MEASURES AND THEIR CORRELATION

In addition to our analysis from the previous section based on the map of elections, in this section, we focus on one structural property of our elections, i.e., the similarity of different votes in one election, and compare four measures capturing different facets of similarity. Notably, similarity measures are a potentially attractive parameter to develop parameterized algorithms because they can be understood as a "distance from triviality" parameterization, as most computational problems are easy if all votes are the same. We evaluate the practicability of parameterized algorithms from Betzler et al. [1] on our data.

Setup. We consider four measures for each election (C, V):

Maximum KT-distance The maximum KT-distance among all pairs of votes: $\max_{v,v' \in \binom{V}{2}} \mathrm{KT}(v, v')$.

Average KT-distance The average KT-distance among all pairs of votes: $\sum_{v,v' \in \binom{V}{2}} \mathrm{KT}(v,v')/|\binom{V}{2}|$.

Disagreeing pairs The number of candidate pairs for which not all votes agree on their ordering: $|\{\{c, c'\} \in \binom{C}{2} \mid \exists v, v' \in V : c \succ_v c' \land c' \succ_{v'} c\}|$.

Kemeny score The minimum summed KT-distance of a central order to all votes: $\min_{v^* \in \mathcal{L}(C)} \sum_{v \in V} \mathrm{KT}(v, v^*)$.

Note that the number of disagreeing pairs is always at least as large as the maximum KT-distance, which in turn is at least as large as the average KT-distance (all three values range from 0 to $|\binom{C}{2}|$ so from 0 to 105 in our case).

Values of Similarity Measures. In Figure 2, for all four measures, we depict for each dataset the value of the similarity measure averaged over all 500 elections from the dataset. Concerning the results on the aggregated dataset, what stands out is that the maximum KT-distance and the number of disagreeing pairs is quite high and in particular much higher than the average KT-distance (and comparing normalized values also than the Kemeny score). However, this is also quite intuitive in the sense that both the maximum distance and the number of disagreeing pairs might in the end only depend on two voters and are thus very sensitive to "outliers" (as soon as there are two voters with reversed preferences orders in an election, both values are at the maximum). Considering the results on the different datasets, especially the average number of disagreeing pairs clearly divides them (in line with our groups proposed in Section 4): Unsurprisingly, the datasets close to identity have a "low" average number of disagreeing pairs (always below 50). The number is the lowest for boxing top 16 and Spotify month with 11.43 and 27.72, respectively. This is quite remarkable as it means that *all* voters agree on the ordering of 89.1% and 73.6% of all candidate pairs, respectively. For the "middle" datasets, the average number of disagreeing pairs is much higher and lies between 89.8 and 99.45 (this means that the voters only agree on the ordering of between 5.4% and 14.4% of all candidate pairs). For the two "outliers", city and country ranking, the average number of disagreeing pairs is very close to the maximum possible value of 105 with 104.93, respectively, 104.73. As already discussed in Section 4 one reason for this might be that in the two "outlier" datasets votes correspond to sometimes contradicting and opposing indicators, which can lead to two close-to-reversed votes.

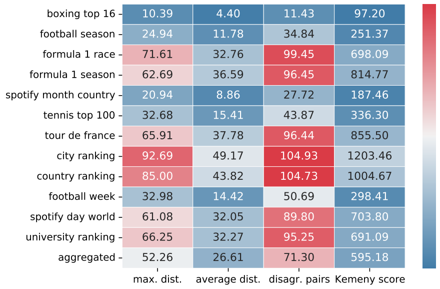


Figure 2: Average values for four different similarity measures. Colors encode the values normalized by the theoretically possible maximum value.

Similarity Measures for Parameterized Algorithms. Betzler et al. [1] developed different parameterized algorithms for computing the central order minimizing the Kemeny score: One algorithm running in $O^*(2^m)$, where m is the number of candidates. Another algorithm running in $O^*(1.53^k)$ where k is the Kemeny score, and an algorithm running in $O^*(16^d)$ where d is the average KT-distance (they also considered the maximum KT-distance between two votes as a parameter for a related problem). Considering the average values on the aggregated dataset, the exponential part of the running time of these algorithms evaluate as follows. 2^m is 32768, 1.53^k is 7.79×10^{109}, and 16^d is 1.1×10^{32}. Even on boxing top 16, where votes are most similar to each other, the number of candidates still leads to the best results (2^m is 32768, 1.53^k is 8.2×10^{17}, 16^d is 198668), partly questioning the practical usefulness of the algorithms for the two similarity parameterizations. Overall, it seems that the number of candidates is nearly always the best of the considered parameters to use. Considering the different similarity measures, the average KT-distance is clearly the smallest, which is also theoretically guaranteed; however, the gap to the other parameters might be seen as unexpectedly large.

Further Considerations. In our full version [11], we discuss that on the aggregated dataset all pairs of similarity measures are strongly correlated. In particular, on each dataset, the NP-hard to compute Kemeny score is very strongly correlated with the average KT-distance. Further, we analyze whether the top part (positions 1 to 8), middle part (positions 5 to 12) or bottom part (positions 8 to 15) of different votes are more similar to each other. Our two main observations are: First, voters typically agree more on which candidates should be considered as high-quality or low-quality candidates than who should be considered as medium-quality candidates. Second, voters tend to rank candidates at the top more consistently in the same ordering than candidates at the bottom. We also take a closer look at time-based elections and analyze the similarity of successive votes in those elections.

6 RESTRICTED DOMAINS

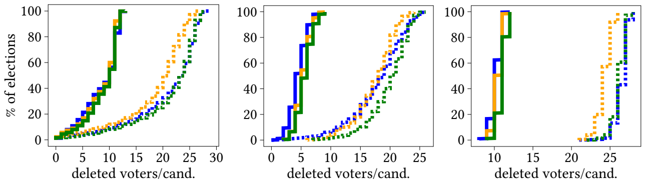
In this section, we analyze which of our elections are part of a restricted domain. There are numerous papers analyzing the computational complexity of various problems on elections from different types of restricted domains (see e.g., [2, 12, 28, 31, 32, 44, 64, 67] and Elkind et al. [25, 26, 27] for surveys). Possible motivations for these works are typically that restricted domains allow for nice combinatorial algorithms and the belief that they capture (close-to) realistic situations. We focus on the three arguably most popular restricted domains of single-peaked [3], single-crossing [51, 62], and group-separable elections [38, 39].

We check here which of our elections fall into one of these domains and afterwards consider the candidate deletion and voter deletion distance of all elections from them.

Members in Restricted Domains. Overall, only very few of our elections fall into a restricted domain. That is, for the 500 boxing top 16 elections, where votes are very similar to each other, the number of single-peaked/singe-crossing/group-separable elections is 77/138/101. Moreover, we have one single-peaked election in the football season dataset and one in the Spotify month dataset. So overall, only 1.3%, 2.3%, 1.6% of our elections are single-peaked, single-crossing, and group-separable, respectively. Some other works have also analyzed the occurrences of elections from restricted domains and found even less evidence: Regenwetter et al. [61] analyzed five-candidate American Psychological Association (APA) presidential elections and found no evidence of restricted domains. Mattei [46] considered three- and four-candidate elections based on a Netflix price competition and found that 0.03% of elections are single-peaked.

Distance to a Restricted Domain. Given that only a few of our elections fall into a restricted domain, our goal now is to check whether more are at least close to one. In particular, we consider the voter deletion and candidate deletion distance, i.e., the minimum number of voters/candidates that need to be deleted such that the resulting election falls into the restricted domain. Notably, there are also many more distance measures (see, e.g., [20, 23, 29]). Motivated by the many polynomial-time results on restricted domains, there are several papers developing parameterized algorithms for election-related problems for different distance measures to restricted domains (see [30, 50, 52] for algorithms parameterized by the voter and candidate deletion distance and [20, 64, 70–72] for examples for other distance measures).

For each of our elections, we computed the voter and candidate deletion distance from single-peakedness, single-crossingness, and group-separability.[6] In Figure 3a, we show the results on the aggregated dataset as a cumulative distribution function. For the candidate deletion distance, the picture is very similar for all three restricted domains: There are around 15% of elections within distance 4, around 28% within distance 6, around 56% within distance 10, and around 99% within distance 12. Considering that we have seen in the previous part that there are considerably more single-crossing elections than single-peaked or group-separable elections, the similarity between the domains here is partly unexpected.



(a) aggregated **(b) Spotify month** **(c) Spotify day**

Figure 3: For different datasets, fraction of elections within a given candidate deletion (solid) or voter deletion distance (dashed) from single-peakedness (blue), single-crossingness (orange), and group-separability (green).

For voter deletion, there is some difference between the restricted domains: For all three restricted domains, 15% of elections are within distance 14 and are more or less uniformly distributed within this distance. For single-peakedness and group-separability, 25% of all elections are within a distance of 18, 50% within a distance of 23, and 99% within a distance of 27. For single-crossingness, distances are typically one smaller, as 25% of all elections fall within distance 17 and 50% within distance 20. This slight difference might be because in contrast to the other two domains, for single-crossingness an ordering of the voters is needed which might be easier to construct if we can choose which voters to delete (however, for single-peakedness the same is true for candidate deletion, yet no such effect is visible). Comparing the normalized voter deletion distance to the normalized candidate deletion distance it seems that the latter is typically slightly smaller. Nevertheless, there is a strong linear correlation between the candidate deletion and voter deletion distance of an election.

Examining the results on the dataset level, there are significant differences: The general trend here is that the higher the average Kemeny score of a dataset is the further is the dataset on average from a restricted domain. One dataset from our close to identity group which contains many elections with a low Kemeny score are Spotify month election, and in Figure 3b we depict the cumulative distribution for this dataset. Notably, for all three restricted domains, 50% of the Spotify month elections have a candidate deletion distance of 5 and smaller. In contrast to this, in Figure 3c we show the plot for Spotify day elections which belong to the middle datasets and have higher Kemeny scores. Here for all elections, at

[6] For single-peaked candidate deletion we used the polynomial-time algorithm from Erdélyi et al. [29] and for single-crossing voter deletion the polynomial-time algorithm from Bredereck et al. [15]. For single-peaked voter deletion and single-crossing candidate deletion distance and for voter and candidate deletion distance to group separability, we used the fixed-parameter tractable algorithms based on conversions to hitting set by Elkind and Lackner [24].

least 9 candidates or at least 21 voters need to be deleted to make it fall into one of our three restricted domains, indicating that this dataset is far away from a restricted domain. Given that one can see the whole Spotify data as one huge election, the opposite behavior of Spotify day and Spotify month elections highlights the natural fact that depending on which votes from a large election are taken into account very different elections arise. To sum up, we have found only little evidence of elections from restricted domains and also only a few elections at a small distance (recall that 5 candidates are not really a small number here, as this corresponds to 33% of candidates). Thus, both the voter and candidate deletion distance are probably too large on many real-world elections for the usage of parameterized algorithms.

Further Considerations. In our full version [11], we check the overlap between elections from different restricted domains. We find that the different restricted domains and their closer environment heavily overlap and that it is, for instance, possible to apply algorithms for (close to) single-peaked or single-crossing elections to an overwhelming majority of (close to) group-separable elections. Further, we analyze the properties of elections that are (close to) being single-peaked or single-crossing and observe that they are typically quite degenerate, meaning that they have a low Kemeny score and that they fall into a small part of the space of all elections from the respective restricted domain. Moreover, we find that value-restricted elections [63] occur quite frequently and that in the characterization of single-peaked, single-crossing, and group-separable elections via forbidden configurations one of the two configurations is redundant on our data.

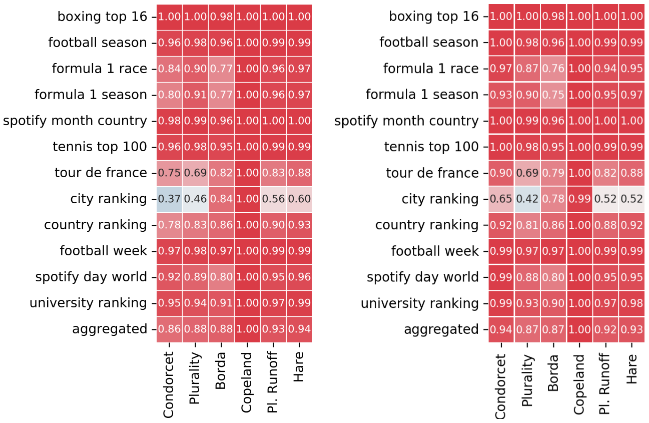
7 CASE STUDY: HOW DIFFERENT ARE DIFFERENT VOTING RULES?

We now use our datasets to shed some further light on traditional questions from social choice. While there is already quite some empirical research on the considered questions, nearly all of these works considered elections with 3 to 5 candidates coming from a single data source. Thus, our rich data allows us to take a broader look.

One popular question arises around the notion of a Condorcet winner. A candidate c is a strong (weak) Condorcet winner if for each other candidate d more than (at least) half of the voters prefer c to d. Previous research has found that strong Condorcet winners nearly always exist, i.e., the so-called Condorcet paradox occurs rarely, and that the strong Condorcet efficiency, i.e., how often the strong Condorcet winner is selected as a winner, of all rules is high [18, 21, 46, 57, 58]. We investigate these issues in Section 7.1.

In Section 7.2, we analyze the level of agreement between different voting rules. While from a theoretical and axiomatic perspective, voting rules significantly differ from each other, various authors provided evidence that most of them are very similar in practice [18, 21, 35, 46, 49, 58, 60, 61].

Overall, while parts of our results in this section are in line with previous studies, we also find evidence that suggests that the established consensus in the literature according to which in practice all voting rules are more or less the same should be relativized, as it seems to only apply if we have elections with a Condorcet winner and/or the number of voters divided by the number of candidates is large. Notably, all our rules may return multiple tied winners.



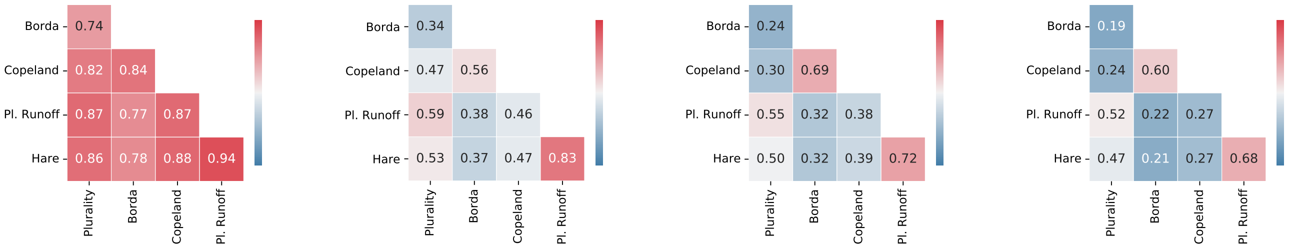
(a) Strong Condorcet winner **(b) Weak Condorcet winner**

Figure 4: In the first column, fraction of elections admitting a strong/weak Condorcet winner. In the other columns, strong/weak Condorcet efficiency of different voting rules.

7.1 Condorcet Paradox and Condorcet Efficiency

In line with the literature, we first focus on strong Condorcet winners. In Figure 4a, in the first column, we depict for each of our datasets the fraction of elections admitting a strong Condorcet winner. While for all datasets from our first group of close to identity datasets around 96% of elections admit a strong Condorcet winner, for the other datasets this fraction is (considerably) below 100%. The most extreme case are city ranking elections where only 37% of the elections admit a strong Condorcet winner. Moreover, overall “only” 86% of all our elections admit a strong Condorcet winner. This is in contrast to previous works. For instance, Popov et al. [58] reported that in one of their studied datasets 93.3% of elections admit a strong Condorcet winner, while for all others this value is above 99.7%.

Concerning the strong Condorcet efficiency of the different voting rules, results again significantly depend on the considered dataset. For close to identity datasets all voting rules have a very high Condorcet efficiency of 0.95 and above (note that Copeland’s voting rule is guaranteed to select a strong Condorcet winner if one exists). Mattei [46] and Popov et al. [58] also reported a Condorcet efficiency of 0.95 and above for different rules. However, on our other datasets, the Condorcet efficiency can be much lower: For Plurality, Plurality with Runoff, and Hare, their Condorcet efficiency is the lowest on the city ranking dataset with 0.46, 0.59, and 0.6, respectively. For Borda, the minimum Condorcet efficiency is 0.77 on Formula 1 race and Formula 1 season elections. Interestingly, the other voting rules achieve a much higher efficiency on these two sets. Considering the results on the aggregated dataset, Hare and Plurality with Runoff have the highest Condorcet efficiency with 0.94 and 0.93 respectively, while Plurality and Borda both have a Condorcet efficiency of 0.88. Given that Borda takes much more information into account than Plurality, it is slightly unexpected that both perform so similarly here.



(a) All elections

(b) All non-city ranking elections without a strong Condorcet winner

(c) All city ranking elections

(d) All city ranking elections without a strong Condorcet winner

Figure 5: For pairs of voting rules, fraction of elections where rules return same winner after lexicographic tie-breaking.

In Figure 4b, we depict the same statistics for the notion of weak Condorcet winners: A substantial fraction of our elections, i.e., 8% of all elections, 15% of Tour de France, and 28% of city ranking elections, admit a weak but no strong Condorcet winner. This is quite remarkable given that the distinction between a weak and a strong Condorcet winner is about tie-breaking. The Condorcet efficiency of our rules slightly decreases when moving from strong to weak Condorcet winners. This is something to be expected because weak Condorcet winners, which are now also taken into account, have in general a slightly weaker general standing in the election than strong Condorcet winners.

7.2 Consensus among Voting Rules

We analyze the consensus among winners returned by different voting rules. In Figure 5a, we depict the average lexicographic agreement of each pair of rules. The average lexicographic agreement of some pair of rules is the fraction of all elections where the winner returned by the two rules is the same if we apply lexicographic tie-breaking during the execution of both rules. In general, the consensus among the different voting rules is quite high, ranging from 0.96 for the only two iterative rules, Hare and Plurality with runoff, to 0.74 for Borda and Plurality. However, the reason for this generally high agreement between voting rules might be connected to our observation from Section 7.1 that most of our elections have a strong Condorcet winner and that in case a strong Condorcet winner exists, most of the time rules return it as a winner. To verify this, in Figure 5b, we depict the average lexicographic agreement of pairs of voting rules on all elections without a strong Condorcet winner (we also excluded elections from the city ranking dataset, since, as argued later, they consistue outliers). Indeed, the consensus among voting rules is significantly lower in this case: For all pairs of rules except for Hare and Plurality with runoff, whose average lexicographic agreement is still 0.83, the average lexicographic agreement drops by between 0.33 and 0.41 when moving from the full election dataset to elections without strong Condorcet winner. Figure 5b further suggests that there exist two groups of voting rules: Plurality, Plurality with Runoff, and Hare on the one hand, and Borda and Copeland on the other hand. This partition is also quite intuitive, as all rules from the first group use Plurality scores in some way or the other, while Copeland and Borda in some sense always take into account the full election. Overall, our results indicate that a main reason why voting rules seem to typically exhibit a high consensus on real-world elections is because they all favor strong Condorcet winners which often exist. This could also explain why previous research [18, 21, 35, 58, 60, 61] has found a higher consensus among rules than what we have observed: On their data strong Condorcet winners exist more often than on ours.

On the dataset level, results are again very different and correlate with our grouping: On the one hand, on close to identity datasets the consensus of voting rules is very high, while, on the other hand, on city rankings it is lowest. In Figures 5c and 5d, we display the average lexicographic agreement on all city ranking elections and on all city ranking elections without strong Condorcet winners. Both Figures 5c and 5d look quite similar (as many city ranking elections do not admit a strong Condorcet winner). Again, we can find the already observed partitioning of the rules into groups. Here, both the consensus between Plurality with Runoff and Hare and the consensus between Borda and Copeland is particularly high.

8 CONCLUSION

We have collected, classified, analyzed, and used a diverse collection of real-world elections and provided various evidence hinting at their usefulness for experimental research. To the best of our knowledge, this is the first work that systematically compares elections from numerous different sources.

For future work, it would be interesting to analyze the relationship of the collected elections to elections drawn from various statistical cultures. Moreover, also performing our experiments on such synthetic elections could be useful to get a better understanding of their properties. In addition, examining the collected elections (even) more carefully would be of great use: While we have been able to provide intuitive explanations for some phenomena we observed, the reasons for others remain unclear. Furthermore, as we have found only little evidence to support the large-scale practical applicability of already developed parameterized algorithms, identifying new properties that are shared by many elections and that allow for the development of tractable algorithms would be extremely valuable. Finally, the main purpose of this project is to provide a helpful source of real-world election datasets. In light of our empirical and structural results, we recommend to use city ranking, football week, Spotify day, and Tour de France elections as a smaller, yet still diverse dataset for testing.

Acknowledgments

NB was supported by the DFG project MaMu (NI 369/19) and by the DFG project ComSoc-MPMS (NI 369/22). NS was supported by the DFG project MaMu (NI 369/19). The authors would like to thank Piotr Faliszewski for constructive criticism of the manuscript

REFERENCES

[1] Nadja Betzler, Michael R. Fellows, Jiong Guo, Rolf Niedermeier, and Frances A. Rosamond. 2009. Fixed-parameter algorithms for Kemeny rankings. *Theor. Comput. Sci.* 410, 45 (2009), 4554–4570.

[2] Nadja Betzler, Arkadii Slinko, and Johannes Uhlmann. 2013. On the Computation of Fully Proportional Representation. *J. Artif. Intell. Res.* 47 (2013), 475–519.

[3] Duncan Black. 1948. On the rationale of group decision-making. *J. Polit. Econ.* 56, 1 (1948), 23–34.

[4] Blitzer. 2017. Movehub City Rankings. kaggle.com/blitzr/movehub-city-rankings. Data obtained from movehub.com.

[5] Niclas Boehmer, Robert Bredereck, Edith Elkind, Piotr Faliszewski, and Stanisław Szufa. 2022. Expected Frequency Matrices of Elections: Computation, Geometry, and Preference Learning. In *Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems (NeurIPS '22)*. To appear.

[6] Niclas Boehmer, Robert Bredereck, Piotr Faliszewski, and Rolf Niedermeier. 2021. Winner Robustness via Swap- and Shift-Bribery: Parameterized Counting Complexity and Experiments. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI '21)*. ijcai.org, 52–58.

[7] Niclas Boehmer, Robert Bredereck, Piotr Faliszewski, Rolf Niedermeier, and Stanislaw Szufa. 2021. Putting a Compass on the Map of Elections. *CoRR* abs/2105.07815 (2021). arXiv:2105.07815 https://arxiv.org/abs/2105.07815

[8] Niclas Boehmer, Robert Bredereck, Piotr Faliszewski, Rolf Niedermeier, and Stanislaw Szufa. 2021. Putting a Compass on the Map of Elections. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI '21)*. ijcai.org, 59–65.

[9] Niclas Boehmer, Piotr Faliszewski, Rolf Niedermeier, Stanisław Szufa, and Tomasz Wąs. 2022. Understanding Distance Measures Among Elections. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI '22)*. ijcai.org, 102–108.

[10] Niclas Boehmer and Rolf Niedermeier. 2021. Broadening the Research Agenda for Computational Social Choice: Multiple Preference Profiles and Multiple Solutions. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '21)*. ACM, 1–5.

[11] Niclas Boehmer and Nathan Schaar. 2022. Collecting, Classifying, Analyzing, and Utilizing Real-World Elections. *CoRR* abs/2204.03589 (2022). arXiv:2204.03589 https://arxiv.org/abs/2204.03589

[12] Felix Brandt, Markus Brill, Edith Hemaspaandra, and Lane A. Hemaspaandra. 2015. Bypassing Combinatorial Protections: Polynomial-Time Algorithms for Single-Peaked Electorates. *J. Artif. Intell. Res.* 53 (2015), 439–496.

[13] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia (Eds.). 2016. *Handbook of Computational Social Choice.* Cambridge University Press.

[14] Felix Brandt, Christian Geist, and Martin Strobel. 2016. Analyzing the Practical Relevance of Voting Paradoxes via Ehrhart Theory, Computer Simulations, and Empirical Data. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS '16)*. ACM, 385–393.

[15] Robert Bredereck, Jiehua Chen, and Gerhard J. Woeginger. 2016. Are there any nicely structured preference profiles nearby? *Math. Soc. Sci.* 79 (2016), 61–73.

[16] Robert Bredereck, Till Fluschnik, and Andrzej Kaczmarczyk. 2020. Multistage Committee Election. *CoRR* abs/2005.02300 (2020). arXiv:2005.02300 https://arxiv.org/abs/2005.02300

[17] Ioannis Caragiannis, Edith Hemaspaandra, and Lane A. Hemaspaandra. 2016. Dodgson's Rule and Young's Rule. In *Handbook of Computational Social Choice*, Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia (Eds.). Cambridge University Press, 103–126.

[18] John R Chamberlin, Jerry L Cohen, and Clyde H Coombs. 1984. Social choice observed: Five presidential elections of the American Psychological Association. *J. Polit.* 46, 2 (1984), 479–502.

[19] Vincent Conitzer and Toby Walsh. 2016. Barriers to Manipulation in Voting. In *Handbook of Computational Social Choice*, Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia (Eds.). Cambridge University Press, 127–145.

[20] Denis Cornaz, Lucie Galand, and Olivier Spanjaard. 2013. Kemeny Elections with Bounded Single-Peaked or Single-Crossing Width. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI '13)*. IJCAI/AAAI, 76–82.

[21] Andreas Darmann, Julia Grundner, and Christian Klamler. 2019. Evaluative voting or classical voting rules: Does it make a difference? Empirical evidence for consensus among voting rules. *Eur. J. Polit. Econ.* 59 (2019), 345–353.

[22] John A. Doucette. 2014. Imputation, Social Choice, and Partial Preferences. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI '14)*. AAAI Press, 3069–3070.

[23] Edith Elkind, Piotr Faliszewski, and Arkadii M. Slinko. 2012. Clone structures in voters' preferences. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC '12)*. ACM, 496–513.

[24] Edith Elkind and Martin Lackner. 2014. On Detecting Nearly Structured Preference Profiles. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI '14)*. AAAI Press, 661–667.

[25] Edith Elkind, Martin Lackner, and Dominik Peters. 2016. Preference Restrictions in Computational Social Choice: Recent Progress. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI '16)*. IJCAI/AAAI Press, 4062–4065.

[26] Edith Elkind, Martin Lackner, and Dominik Peters. 2017. Structured Preferences. In *Trends in Computational Social Choice*, Ulle Endriss (Ed.). AI Access, Chapter 10, 187–208.

[27] Edith Elkind, Martin Lackner, and Dominik Peters. 2022. Preference Restrictions in Computational Social Choice: A Survey. *CoRR* abs/2205.09092 (2022). https://doi.org/10.48550/arXiv.2205.09092

[28] Edith Elkind, Evangelos Markakis, Svetlana Obraztsova, and Piotr Skowron. 2016. Complexity of Finding Equilibria of Plurality Voting Under Structured Preferences. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS '16)*. ACM, 394–401.

[29] Gábor Erdélyi, Martin Lackner, and Andreas Pfandler. 2017. Computational Aspects of Nearly Single-Peaked Electorates. *J. Artif. Intell. Res.* 58 (2017), 297–337.

[30] Piotr Faliszewski, Edith Hemaspaandra, and Lane A. Hemaspaandra. 2014. The complexity of manipulative attacks in nearly single-peaked electorates. *Artif. Intell.* 207 (2014), 69–99.

[31] Piotr Faliszewski, Edith Hemaspaandra, Lane A. Hemaspaandra, and Jörg Rothe. 2011. The shield that never was: Societies with single-peaked preferences are more open to manipulation and control. *Inf. Comput.* 209, 2 (2011), 89–107.

[32] Piotr Faliszewski, Alexander Karpov, and Svetlana Obraztsova. 2020. The Complexity of Election Problems with Group-Separable Preferences. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI '20)*. ijcai.org, 203–209.

[33] Piotr Faliszewski and Jörg Rothe. 2016. Control and Bribery in Voting. In *Handbook of Computational Social Choice*, Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia (Eds.). Cambridge University Press, 146–168.

[34] Piotr Faliszewski, Arkadii Slinko, Kolja Stahl, and Nimrod Talmon. 2018. Achieving fully proportional representation by clustering voters. *J. Heuristics* 24, 5 (2018), 725–756.

[35] Dan S Felsenthal, Zeev Maoz, and Amnon Rapoport. 1993. An empirical evaluation of six voting procedures: do they really make any difference? *Br. J. Polit. Sci.* 23, 1 (1993), 1–27.

[36] T. Fruchterman and E. Reingold. 1991. Graph drawing by force-directed placement. *Software Pract. Exper.* 21, 11 (1991), 1129–1164.

[37] Judy Goldsmith, Jérôme Lang, Nicholas Mattei, and Patrice Perny. 2014. Voting with Rank Dependent Scoring Rules. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI '14)*. AAAI Press, 698–704.

[38] Ken-ichi Inada. 1964. A note on the simple majority decision rule. *Econometrica* (1964), 525–531.

[39] Ken-ichi Inada. 1969. The simple majority decision rule. *Econometrica* (1969), 490–506.

[40] Mart Jürisoo. 2021. UFC Rankings. kaggle.com/martj42/ufc-rankings. Data obtained from ufc.com.

[41] Orgad Keller, Avinatan Hassidim, and Noam Hazon. 2019. New Approximations for Coalitional Manipulation in Scoring Rules. *J. Artif. Intell. Res.* 64 (2019), 109–145.

[42] Martin Lackner. 2020. Perpetual Voting: Fairness in Long-Term Decision Making. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI '20)*. AAAI Press, 2103–2110.

[43] Jérôme Lang and Lirong Xia. 2016. Voting in Combinatorial Domains. In *Handbook of Computational Social Choice*, Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia (Eds.). Cambridge University Press, 197–222.

[44] Krzysztof Magiera and Piotr Faliszewski. 2017. How hard is control in single-crossing elections? *Auton. Agents Multi Agent Syst.* 31, 3 (2017), 606–627.

[45] Ken Massey. 2021. College Football/Basketball/Baseball Rankings. kaggle.com/masseyratings/rankings. Data obtained from masseyratings.com/.

[46] Nicholas Mattei. 2011. Empirical Evaluation of Voting Rules with Strictly Ordered Preference Data. In *Proceedings of the Second International Conference on Algorithmic Decision Theory (ADT '11)*. Springer, 165–177.

[47] Nicholas Mattei and Toby Walsh. 2013. PrefLib: A Library for Preferences http://www.preflib.org. In *Proceedings of the Third International Conference on Algorithmic Decision Theory (ADT '13)*. Springer, 259–270.

[48] Nicholas Mattei and Toby Walsh. 2017. A preflib.org Retrospective: Lessons Learned and New Directions. In *Trends in Computational Social Choice*, Ulle

Endriss (Ed.). AI Access, Chapter 15, 289–305.

[49] John C McCabe-Dansted and Arkadii Slinko. 2006. Exploratory analysis of similarities between social choice rules. *Group Decis. Negot.* 15, 1 (2006), 77–107.

[50] Vijay Menon and Kate Larson. 2016. Reinstating Combinatorial Protections for Manipulation and Bribery in Single-Peaked and Nearly Single-Peaked Electorates. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI '16)*. AAAI Press, 565–571.

[51] James A Mirrlees. 1971. An exploration in the theory of optimum income taxation. *Rev. Econ. Stud.* 38, 2 (1971), 175–208.

[52] Neeldhara Misra, Chinmay Sonar, and P. R. Vaidyanathan. 2017. On the Complexity of Chamberlin-Courant on Almost Structured Profiles. In *Proceedings of the 5th International Conference on Algorithmic Decision Theory (ADT '17)*. Springer, 124–138.

[53] Eduardo M. R. Oliveira. 2018. Spotify's Worldwide Daily Song Ranking. kaggle.com/edumucelli/spotifys-worldwide-daily-song-ranking.

[54] Myles O'Neill. 2019. World University Rankings. kaggle.com/mylesoneill/world-university-rankings.

[55] Alcides Oxa. 2019. World happiness report 2005 2018. kaggle.com/alcidesoxa/world-happiness-report-2005-2018.

[56] David C. Parkes and Ariel D. Procaccia. 2013. Dynamic Social Choice with Evolving Preferences. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI '13)*. AAAI Press.

[57] Florenz Plassmann and T. Nicolaus Tideman. 2014. How frequently do different voting rules encounter voting paradoxes in three-candidate elections? *Soc. Choice Welf.* 42, 1 (2014), 31–75.

[58] Sergey V Popov, Anna Popova, and Michel Regenwetter. 2014. Consensus in organizations: Hunting for the social choice conundrum in APA elections. *Decision* 1, 2 (2014), 123.

[59] Rohan Rao. 2021. Formula 1 World Championship (1950 - 2021). kaggle.com/rohanrao/formula-1-world-championship-1950-2020. Data obtained from ergast.com/mrd/.

[60] Michel Regenwetter, Bernard Grofman, Ilia Tsetlin, and Anthony AJ Marley. 2006. *Behavioral social choice: probabilistic models, statistical inference, and applications.* Cambridge University Press.

[61] Michel Regenwetter, Aeri Kim, Arthur Kantor, and Moon-Ho R Ho. 2007. The unexpected empirical consensus among consensus methods. *Psychol. Sci.* 18, 7 (2007), 629–635.

[62] Kevin WS Roberts. 1977. Voting over income tax schedules. *J. Polit. Econ.* 8, 3 (1977), 329–340.

[63] Amartya K Sen. 1966. A possibility theorem on majority decisions. *Econometrica* (1966), 491–499.

[64] Piotr Skowron, Lan Yu, Piotr Faliszewski, and Edith Elkind. 2015. The complexity of fully proportional representation for single-crossing electorates. *Theor. Comput. Sci.* 569 (2015), 43–57.

[65] Xin Sui, Alex Francois-Nienaber, and Craig Boutilier. 2013. Multi-Dimensional Single-Peaked Consistency and Its Approximations. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI '13)*. IJCAI/AAAI, 375–382.

[66] Stanislaw Szufa, Piotr Faliszewski, Piotr Skowron, Arkadii Slinko, and Nimrod Talmon. 2020. Drawing a Map of Elections in the Space of Statistical Cultures. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '20)*. IFAAMAS, 1341–1349.

[67] Toby Walsh. 2007. Uncertainty in Preference Elicitation and Aggregation. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence (AAAI '07)*. AAAI Press, 3–8.

[68] Jun Wang, Sujoy Sikdar, Tyler Shepherd, Zhibing Zhao, Chunheng Jiang, and Lirong Xia. 2019. Practical Algorithms for Multi-Stage Voting Rules with Parallel Universes Tiebreaking. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI '19)*. AAAI Press, 2189–2196.

[69] Mimi Wang. 2019. ATP Men Singles Tennis Rankings 1990 to 2019. kaggle.com/mimoopoo/atp-tennis-rankings-1990-to-2019. Data obtained from atptour.com.

[70] Yongjie Yang and Jiong Guo. 2014. Controlling elections with bounded single-peaked width. In *Proceedings of the 2014 International conference on Autonomous Agents and Multi-Agent Systems (AAMAS '14)*. IFAAMAS/ACM, 629–636.

[71] Yongjie Yang and Jiong Guo. 2015. How Hard is Control in Multi-Peaked Elections: A Parameterized Study. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '15)*. ACM, 1729–1730.

[72] Yongjie Yang and Jiong Guo. 2017. The control complexity of r-Approval: From the single-peaked case to the general case. *J. Comput. Syst. Sci.* 89 (2017), 432–449.

[73] William S. Zwicker. 2016. Introduction to the Theory of Voting. In *Handbook of Computational Social Choice*, Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia (Eds.). Cambridge University Press, 23–56.