Guide to Numerical Experiments on Elections in Computational Social Choice

Niclas Boehmer, Piotr Faliszewski, Łukasz Janeczko, Andrzej Kaczmarczyk, Grzegorz Lisowski, Grzegorz Pierczyński, Simon Rey, Dariusz Stolicki, Stanisław Szufa, and Tomasz Wąs

Abstract

We analyze how numerical experiments regarding elections were conducted within the computational social choice literature (focusing on papers published in the IJCAI, AAAI, and AAMAS conferences). We analyze the sizes of the studied elections and the methods used for generating preference data, thereby making previously hidden standards and practices explicit. In particular, we survey a number of statistical cultures for generating elections and their commonly used parameters.

1 Introduction

Computational social choice is an interdisciplinary area that draws on artificial intelligence, computer science theory, economics, operations research, logic, social sciences, and many other fields [13]. Its main goal is algorithmic analysis of collective decision making processes, but over time noncomputational approaches, such as the axiomatic method or game-theoretic considerations, have also become popular and are pursued equally vigorously. Up to a few years ago, results in computational social choice were largely theoretical and only recently numerical experiments—not to mention actual empirical studies—have received more prominent attention. In this survey, our goal is to encourage further experimental studies on elections and voting, a prominent subarea of computational social choice, by presenting a *Guide*. Our Guide has two main components:

1. On the one hand, the Guide surveys how experiments were performed so far, what election sizes were considered, how data was obtained, and what parameters were considered. Such information is helpful when planning one's own experiments, e.g., to stay in sync with the literature. In this sense, the paper is akin to a *tourist guide*, which shows the richness of the landscape that one would see, e.g., upon visiting a city.
2. On the other hand, we want to point out good practices and make recommendations as to how experiments should be run. While each experiment is different and requires specific considerations, there are also general rules of thumb that one might want to follow (such as using at least several data sources, which in the past has often been neglected). In this sense, our guide takes a role of a *"how to"* document, giving advice.

To achieve these goals, initially we have gone over all papers published in the AAAI, IJCAI, and AAMAS conference series between 2010 and 2023 and collected those that discuss elections and voting (or some very similar structures; see Section 2 for details on the collection process). As we continue working on this process, the current version also includes papers published in these conferencs in 2024. We intend to update the survey annually.

For each of the collected papers, we have analyzed how the authors obtained preference data for their experiments, which statistical cultures (i.e., models of generating synthetic data) they used and with which parameters, and what election sizes they considered. A large part of the survey is discussing the

conclusions from this analysis. This includes providing general statistics (such as the number of papers that include experiments in various years, or the number of data sources used by the papers) and an overview of popular statistical cultures. We contrast these observations with the *map of elections*, as introduced by Szufa et al. [44], which shows relations between various statistical cultures and real-life data sets, as well as with the *microscope* of Faliszewski et al. [22], which visualizes specific elections (and, effectively, specific synthetic models). We use these tools to give some advice as to which statistical models are possibly more appealing than others.

We complement our work by providing a Python package with implementations of the most popular models for sampling approval and ordinal elections https://github.com/COMSOC-Community/prefsampling and a website with access to our database of papers https://guide.cbip.matinf.uj.edu.pl/. Due to limited space, we mostly focus on ordinal elections.

2 Collecting Data

We have collected all papers that were published in the AAAI, IJCAI, and AAMAS conference series between 2010 and 2023 (in case of IJCAI we have also collected the papers from 2009). For the Guide, we selected papers that contained numerical experiments on elections (or very related structures).

By an *election*, we mean a pair $E = (C, V)$, where $C = \{c_1, \ldots, c_m\}$ is a set of candidates and $V = (v_1, \ldots, v_n)$ is a sequence of voters that express preferences over these candidates. In an *ordinal election* each voter v_i has a preference order, i.e., a strict ranking $\succ_{v_i}$ of the candidates, from the one that v_i likes most to the one that he or she likes least. In an *approval election*, each voter v_i has a set $A(v_i) \subseteq C$ of candidates that he or she approves. Occasionally, authors consider variants of elections where, for example, the preference orders are either weak or partial, or are expressed over some combinatorial domain (e.g., see the literature on CP-nets [33]). We include papers that study such elections as well.

We restrict our attention to papers that include elections with at least three candidates. Indeed, two-candidate elections are very different from those with at least three.[1] As a consequence, we do not include numerous papers that study, e.g., a setting where two parties compete (as, e.g., the work of Borodin et al. [12]) or which are motivated by presidential elections with two candidates (as, e.g., the paper of Wilder and Vorobeychik [46]), or which focus on liquid democracy and voting over two options (as examples, see the works of Colley et al. [14] and Bloembergen et al. [4]).

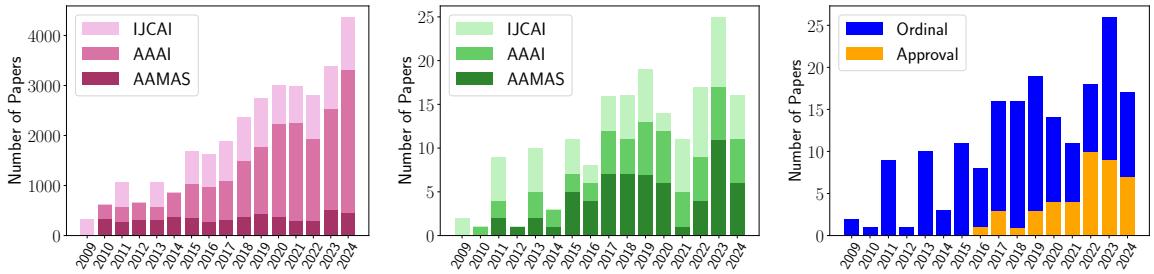
Occasionally we ran into gray areas and bent (or not) our rules on an individual basis.[2] We believe that most readers would agree with most of our choices. We list and cite all the 163 papers that we included in the Guide, together with meta-data about their experiments, in the full version of the paper.

Collecting Papers. We have downloaded the papers from the respective conferences in September 2023, using the links from the DBLP website.[3] This way we included all tracks of the conferences, including, e.g., demo or doctoral consortium papers, etc. We skipped 34 papers, whose links were missing or were corrupted and which could not be downloaded manually from any official source. Then, we performed an automated screening to select a long list of papers that might contain experimental studies of elections. Specifically, for each paper we checked whether it included keywords related to elections and experiments (the keywords were `election`, `vote`, and `ballot` for elections, and `experiment`, `simulation`, and `empirical` for the experiments; to pass the screening, a paper had to include words from both groups, on at least two pages). We looked at each paper that passed the keyword-based screening and checked if it indeed considered elections and included experiments. While our sets of keywords were

[1] Naturally, we include papers that consider two candidates as a special case, in addition to larger candidate sets.

[2] For example, we did not include the work of Peters et al. [39] in the Guide as in the conference versions the authors mention conclusions from experiments, but do not describe their details.

[3] Source: https://dblp.org/xml/release/dblp-2023-09-01.xml.gz



(a) No. of papers from the considered conference series downloaded for the Guide.

(b) No. of papers in the Guide from AAAI, IJCAI, and AAMAS conference series.

(c) No. of papers in the Guide that consider either ordinal or approval elections.

Figure 1: Statistics regarding the numbers of papers in the Guide.

selected to limit the number of papers that we had to analyze manually, they were also meant to not be very restrictive. For example, IJCAI-2023 included 846 papers of which 41 passed the initial screening, but only 7 passed manual checking and made it to the Guide.

Recording Experiments. Finally, we have analyzed the experiments that the collected papers included. For each experiment, we recorded the type of elections used (ordinal or approval), how the votes were obtained (e.g., if they were generated from some statistical culture or were based on real-life data), the sizes of the considered elections (expressed as numbers of candidates and voters), and the number of samples used to obtain each "data point" (the notion of a data point is paper-specific; in most cases it meant the number of elections generated for each datapoint on some plot). For each of these parameters we recorded additional notes, if we felt that some further comments would be helpful.

Remark 2.1. *Authors often consider elections where some parameter—such as the number of voters—changes with a particular step (e.g., from* 20 *to* 100 *voters, with a step of* 5*). In such cases, we recorded the range of election sizes considered, but omitted the step parameter. Indeed, we felt that availability of such data would not affect our analysis too strongly, but would hinder data collection.*

We stress that our notion of what counts as *one* experiment is quite distinctive. For example, if some hypothetical paper described two "experiments," where in the former it considered the running time of some algorithm and in the latter it analyzed whether some property is satisfied, but it used the same (or, identically generated) data for both, then we would have recorded this as a single experiment. Similarly, if a paper included a single "experiment," such as, e.g., testing manipulability of some voting rule, but within this "experiment" it first focused on one statistical culture and a range of election sizes, and then it moved to a different culture and a different range of sizes, then we would record this as two experiments.

3 Bird's Eye View of The Guide

At the time of writing this version of this survey, the Guide included 179 papers. In Figure 1a we plot the number of papers that we downloaded for each of the considered conferences, and in Figure 1b we show how many papers in each of the conferences included numerical experiments on elections. Generally, the trend is that the number of experimental works was increasing in the period between 2010 and 2016, but now has largely stabilized, albeit with a notable drop in 2020 and 2021 and a peak in 2023. Indeed, it tempting to say that the additional papers published in 2023 were the ones "missing" in 2020 and 2021. One might also speculate that the decrease in 2021 was due to the COVID-19 pandemics but, as Figure 1a shows, the overall number of papers in the conferences has not decreased as dramatically).

In Figure 1c we plot the number of papers in the Guide that consider experiments on either ordinal or

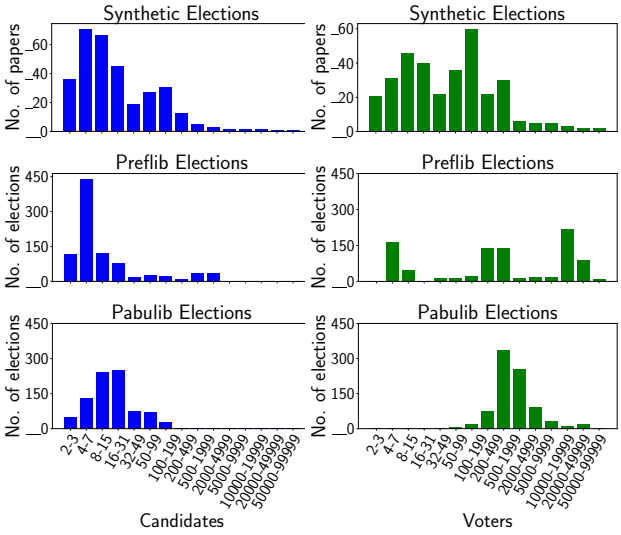


Figure 2: Histograms of the numbers of candidates and voters of synthetic elections used in the papers from the Guide (top), and in Preflib (middle) and Pabulib (bottom).

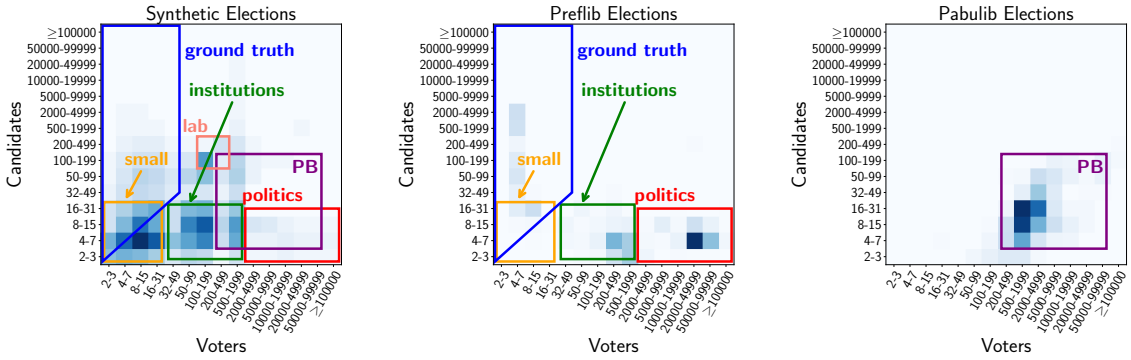


Figure 3: Heatmaps of the sizes of synthetic elections used in the papers from the Guide (left), real-life elections from Preflib (middle), and real-life elections from Pabulib (right). Preflib plot omits the elections provided by Boehmer and Schaar [6] (including them would create an overwhelming spike in the area for 8-31 voters and 100-499 candidates). Darker cells mean more papers with elections of a given size.

approval elections. While, so far, ordinal elections have received far greater attention (altogether 140 papers consider them, whereas only 42 papers include experiments on approval ones; with some papers including both types of elections), it is evident that in recent years approval elections have become popular. One of the reasons for this partial shift of interest is that approval elections are very natural in the context of multiwinner elections [19, 32] and in participatory budgeting [41], two topics that received a lot of attention in recent years.

3.1 Sizes of Elections in Experiment

Next, we analyze the sizes of elections studied in the papers from the Guide. In Figure 2 we plot histograms showing how many papers consider particular numbers of candidates and voters, and in Figure 3 we show heatmaps illustrating the popularity of different combinations of these parameters. We also include analogous data for elections from the Preflib [35] and Pabulib [21] databases of real-life elections (the former mostly contains ordinal elections, whereas the latter mostly includes approval ones, only regarding participatory budgeting; Pabulib plots omit "Artificial Mechanical Turk" datasets).

regime	candidates (m)	voters (n)
small elections	$2 - 30$	$2 - 30$
political elections	$2 - 20$	≥ 2000
voting in institutions	$2 - 30$	$30 - 2000$
participatory budgeting	$4 - 200$	$200 - 100000$
ground truth	$m \geq n$	≤ 50
multiwinner lab	$100 - 500$	$100 - 500$

Table 1: Rough classification of the ranges of numbers of candidates and voters in various types of elections in the papers from the Guide.

Remark 3.1. *In Figures 2 and 3, for each paper we record each election size that occurs in its experiments only once, even if it appears in several experiments (if we recorded each election size once per experiment, the figures would not change much). Further, if an experiment considers elections of different sizes (for example, analyzing how its result changes as we vary the numbers of candidates or voters), then we record an election with a given size for each bucket in the histogram/heatmap to which it fits.*

We identify six main regimes in which many of the papers operate, listed in Table 1. The classification is due to us, but it is inspired by what we have seen in the papers, and it takes into account the data from Preflib and Pabulib. Hence, the boundaries of the regimes are somewhat arbitrary and fluid, and papers sometimes mention other motivations for the election sizes they consider (or often omit such motivation altogether). Further, the classification is naturally not perfectly accurate and rather focuses on capturing general trends and pragmatics. For example, it is possible that there is some (fairly atypical) real-life political election with 30 candidates and 500 voters, even though we classify such elections as having between 2 and 20 candidates, and at least 2000 voters. As many papers that consider elections from a given regime do not mention this explicitly as their motivation or goal, it is reassuring that, nonetheless, the community focused on elections that match natural, realistic settings (with the possible exception of the *multiwinner lab* one, which is not particularly realistic, but has other redeeming features). Below we discuss the regimes in detail.

Small Elections. This regime includes the smallest elections and captures, e.g., groups of friends voting on where to have lunch or small committees within companies, e.g., deciding who to hire (given a shortlist). However, generally, papers using this type of data do not explicitly state their motivation. Experiments over small elections are sometimes conducted to provide illustrations for theoretical results, rather than to get new insights. Notably, small elections are often chosen due to technical challenges, for instance when the studied problems are computationally difficult. They also often arise in studies done on human subjects.

Politics. Our next type of elections regards various forms of *political elections*, which contain a limited number of candidates ($m \leq 20$) and a comparably high number of voters ($n \geq 2000$). Papers that use elections of these sizes and point to specific motivations indeed typically mention some form of political elections, such as parliamentary, city board, referendum, or presidential (nominee) ones. Accordingly, political elections from Preflib (such as the Irish dataset) are particularly popular in such papers. The only other application scenario that is occasionally mentioned is crowdsourcing, e.g., in the form of large-scale surveys (such as the Sushi survey on Preflib) or peer grading.

Voting in Institutions. Our next regime involves fairly small groups of up to 30 candidates and slightly larger numbers of voters (up to 2000), which can be seen as the sizes of a typical election in an institution

such as, e.g., a professional association.[4] However, papers using these election sizes often do not focus on particular applications and simply find this setting appealing. Indeed, elections from this regime are sometimes used due to the hardness of computational problems studied, as they often allow for sufficiently realistic, but manageable experiments. Papers using such elections focused on a wide range of topics, involving matching, party elections, iterative voting, or randomized voting rules. It is also worth mentioning that many papers in this category included other (smaller or larger) election sizes.

PB Elections. Instances in this group are mostly real-life participatory budgeting elections from Pabulib. They typically contain hundreds (up to 220) of candidates and more than 200, but up to tens of thousands, of voters. There is no canonical way of using the resources from Pabulib. Authors usually consider either (i) all elections that are available at the time they access Pabulib; (ii) elections that satisfy certain size criteria (e.g., have at least 10 candidates); or (iii) elections that are of high enough quality (i.e., large-sized elections with a high average number of approvals per voter), such as PB elections from Warsaw from the years 2020–2023. As Pabulib is constantly growing, it is important to mention in such papers either the time of downloading the data or some basic statistics of the used data. Recently, Faliszewski et al. [23] have analyzed Pabulib data in detail and found that the properties of elections from different cities often differ quite significantly. Consequently, one might want to use data from at least several cities in experiments.

Multiwinner Lab. This type of election contains mid-sized instances that are characteristic of experimental analysis of *multiwinner* voting rules (with very few exceptions). Papers, many of which are written by some of the coauthors of this survey, often argue that the considered numbers of candidates and voters, both between 100 and 500, balance the trade-off between running times of algorithms and the structural complexity of the preferences. Briefly put, these elections are big enough to be interesting in the context of studied properties, but small enough to be analyzed by the respective computational techniques. Elections with equal numbers of voters and candidates, specifically $m = n = 100$ and $m = n = 200$, are particularly prevalent. Sometimes, the number m of candidates is determined by the desired committee size k with the goal to obtain a certain (e.g., integral) value of m/k. Naturally, these specific elections are typically generated using synthetic models.

Search for Ground Truth. This class of elections is slightly more vague. It contains elections where there are different "credible" sources of information ($n \leq 50$) ranking a variety of candidates ($m > n$) and typically the goal is to aggregate these sources to recover an objective quality ranking of the candidates. These elections appear in many papers with a range of mentioned application scenarios including aggregating the opinions of experts (e.g., judges or funding panel members), aggregating rankings of items according to different criteria (e.g., price, outward appearance,...), aggregating rankings of athletes in different types of competitions (e.g., Olympic climbing), aggregating the outputs of different computer systems (e.g., machine translation systems or search engines), or deciding which items to select for a small group. Elections of these sizes are typically generated from the impartial culture model (even more frequently than in the other regimes), whereas the Mallows model, which would be a natural choice for such scenarios, and real-world data are rarely used (see Section 4 for a discussion of statistical cultures). Real-world datasets from Preflib that fall into this category include different sports competitions (such as Formula 1 and speed skating), criteria-based rankings (e.g., of cities, countries and universities), and rankings output by different search engines according to the same query.

[4] Elections to the IFAAMAS Board of Trustees, with over 300 eligible voters, are a possible real-life example, and ERS data from Preflib is another. On the other hand, presidential elections of the American Psychological Association (APA) that are available on Preflib have around 5 candidates and 17,000 voters and are thus perhaps closer to the political setting.

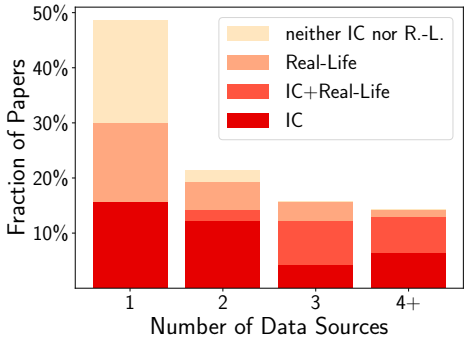
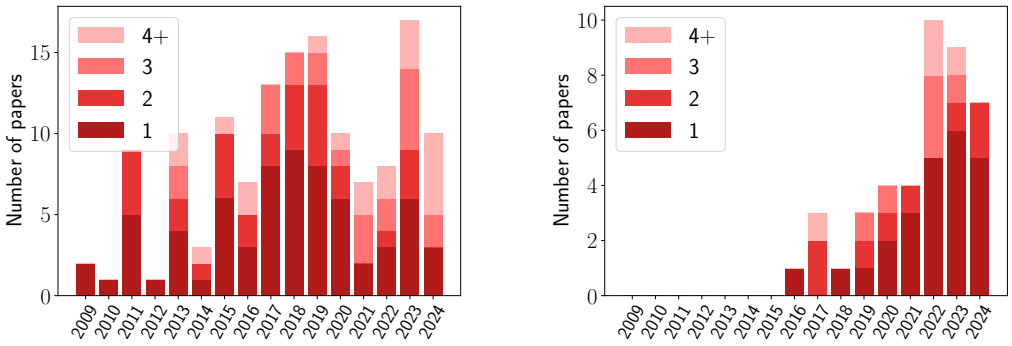


Figure 4: Numbers of data sources used in the papers that consider ordinal elections. "Neither IC nor R.-L." means papers that used neither impartial culture (IC) nor real-life data, "Real-Life" means using real-life data but not IC, "IC + Real-Life" means using both IC and real-life data, and "IC" means using IC but not real-life data.



(a) Ordinal Elections

(b) Approval Elections

Figure 5: Numbers of data sources used in the papers from the Guide that consider either ordinal (left) or approval (right) elections in particular years.

3.2 Statistics of Data Sources

Overall, in 140 papers we identified 228 experiments that were using ordinal elections. Most of them (59.3%) used only synthetic data. It is a bit worrisome that 15.7% of the papers relied solely on the highly unrealistic impartial culture model (where we choose votes uniformly at random). About 14.3% of the papers used only real-life elections (mostly from Preflib), with the Sushi dataset being the most popular. We include aggregated statistics about the number of data sources for ordinal elections in Figure 4, and in Figure 5 (left) we show the numbers of papers that use a given number of data sources depending on a year.[5] We see that in the last few years more and more papers use more than just a single source of data, which certainly is a positive trend.

Regarding approval elections, in 42 papers we recorded 58 experiments. In Figure 5 (right) we see how many papers use a given number of data sources. As opposed to the ordinal case, we see that majority of papers use only a single source of data. However, in this case it is not as worrisome as most typically this means using real-life data from Pabulib (or real-life data from Preflib, adapted to the approval setting). Altogether, 61.9% of the papers that study the approval setting used real-life data (see Figure 8).

[5]We treat different statistical cultures as different data sources, but we view "real-life data" as a single one, irrespective of whether a paper is using just a single dataset from Preflib (such as the Sushi one) or multiple different ones.

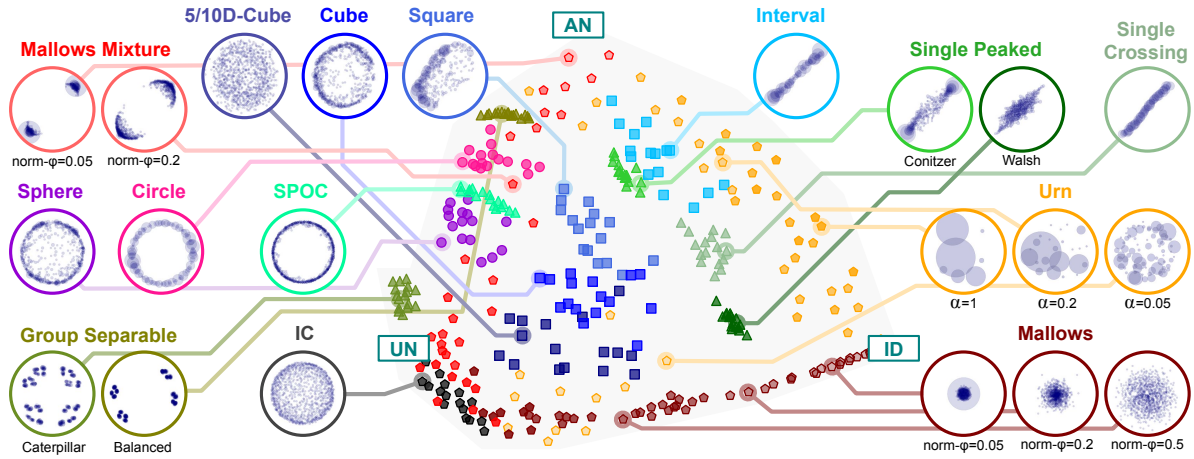


Figure 6: Map of elections and the microscope. Elections in the map have 8 candidates and 96 voters (for computational reasons) and the ones in the microscope have 10 candidates and 1000 voters (for better visualization). Hence, the connections between the elections on the map and their microscopes are meant to show a general behavior, not the exact compositions of the given election.

4 Statistical Cultures for Ordinal Elections

In this section we take a closer look at the most popular statistical cultures, i.e., models of generating synthetic preference data, for ordinal elections (75% of the papers that study ordinal elections use at least one of the cultures that we describe, and this fraction grows to over 92.1% if we include real-life data). Below we provide their definitions and discuss their use in the papers from the Guide, including common parameter settings. Further, in Figure 6 we illustrate elections that these models generate, as well as the relations between the models, on a map of elections. The swap distance between two preference orders u and v, denoted $\kappa(u, v)$, is the number of pairs of candidates a and b such that u and v disagree on their ranking (i.e., one of them ranks a above b, and the other ranks b above a).

Maps of elections are a way to visualize an election dataset and have been introduced by Szufa et al. [44]. Specifically, for each two elections in the dataset we measure their similarity (using the isomorphic swap distance [20]) and visualize them as points on a plane, so that the Euclidean distances between the points resemble these similarities (we use the MDS embedding [30]). Crucially, the maps use distances that are invariant to renaming the candidates and voters and, hence, illustrate structural similarities between the elections. Further, our maps include three special elections as reference points: Identity (ID), where all votes are the same, antagonism (AN), which has two equal-sized groups of voters with opposite preference orders, and an approximation of a uniformity (UN) election, where each possible vote appears once. Maps help in designing diverse datasets as one may choose elections that are far from each other and, hence, likely have different features. A map of Preflib was shown by Faliszewski et al. [24].

Following Faliszewski et al. [22], we also include "microscope" maps of specific types of elections. To form such a microscope, we take a single election, measure the swap distance between each pair of its votes, and then draw a picture where each disc represents a single vote (with the radii of the discs representing the number of identical votes) and the Euclidean distances between the discs resemble the swap distances between the votes. Microscope maps show internal structure of the considered elections and we recommend that one looks at them whenever generating data from a new source.

In Figure 7 we give statistics as to how popular are particular statistical cultures for ordinal elections (note that the exact fractions given in this plot can be slightly different from the ones given in the following paragraph headings; for example, in Figure 7 we distinguish between 1D-Euclidean models and Euclidean models for higher dimensions, but we do not do so in the text below).

Impartial Culture (Used in 55% of the Papers). Under the impartial culture (IC) model we generate votes one-by-one, choosing each preference order uniformly at random. Consequently, there is no apparent structure among the votes, as seen in Figure 6. While by now the model is part of the folklore, its first use dates back to the work of Guilbaud [26], who studied the probability of the Condorcet paradox. It is commonly agreed that impartial culture does not generate realistic elections but, nonetheless, it is used in 55% of the papers. Indeed, the model is extremely simple and does not require setting any parameters. This means that every experiment that uses IC, uses the very same distribution. Consequently, it has become the baseline that many researchers evaluate their results against.[6] We largely agree with this use of IC as a common yardstick, but we very strongly encourage the use of further models in experiments, to get a broader view of the studied phenomena.

Impartial anonymous culture (IAC), introduced by Kuga and Nagatani [31] and Fishburn and Gehrlein [25], is a variant of IC where each voting situation is equiprobable (a voting situation associates each vote with the number of voters that cast it). Impartial anonymous and neutral culture (IANC) further abstracts away from candidate names [18]. Unless there are very few candidates or the number of voters is huge, IAC and IANC generate elections that are very similar to IC.

Mallows Model (Used in 30% of the Papers). Using the Mallows model [34] is the second most popular way to generate synthetic elections in the Guide. This is quite positive as recent work indicated that it provides a good coverage of the space of real-life elections [5, 7]. In Figure 6, Mallows elections form a line between ID and UN. The basic idea is that there is an underlying "ground truth" ordering v^* of the candidates and that the probability of sampling a vote from the model decreases with the vote's distance from v^*. The expected distance can be controlled by a dispersion parameter $\phi \in [0, 1]$. Formally, the probability of sampling a vote v is proportional to $\phi^{\kappa(v,v^*)}$. (Occasionally authors express the probability of sampling a vote v as proportional to $e^{-\phi \cdot \kappa(v,v^*)}$, as done, e.g., in the work of Doucette and Cohen [16]. This is correct, but yields a different range of ϕ values.)

Authors often consider multiple values of the dispersion parameter at equal distances from each other (e.g., $\phi \in \{0.1, 0.2, \ldots\}$), but single values (e.g., $\phi = 0.8$ or $\phi = 0.5$) appear as well. Generally, there is a trend toward using larger values. Another strategy is not to consider specific, fixed values and, instead, generate elections by first sampling a value of the dispersion parameter uniformly from some pre-specified range and then drawing votes from the Mallows model with the drawn dispersion parameter (see e.g., the works of Bachrach et al. [1], Boehmer et al. [8], Faliszewski et al. [22]). This procedure creates a diverse dataset without the need for separate evaluations. Mixtures of Mallows models combining multiple models with different central orders and dispersion parameters with some weight function on top have also been used, but less frequently (an example of such a mixture, with the voters equally split between two Mallows models with equal noise and opposite central orders, is visible in Figure 6).

Recently, Szufa et al. [44] and Boehmer et al. [9, 5] argued that there are certain issues when using the Mallows model. In particular, they showed that equally-spaced values of the dispersion parameter do not provide a uniform coverage of the space between ID and UN elections: For larger numbers of candidates, parameter values below, say, 0.8 will result in elections where votes are fairly similar to each other (this, indeed, justifies the use of high ϕ values in previous works). Moreover, they argued that fixing a dispersion parameter and changing the number of candidates fundamentally changes the nature of the sampled elections, thus rendering results for different numbers of candidates incomparable. They provided a new parameter, $\text{norm-}\phi$, that ensures that uniformly-selected parameter values provide uniform coverage of the space between ID and UN (indeed, to generate Mallows elections for Figure 6, we were choosing $\text{norm-}\phi \in [0, 1]$ uniformly at random): Given a value of $\text{norm-}\phi \in [0, 1]$, one computes classic ϕ so that the expected swap distance between the central vote and one generated using the Mallows model is $\text{norm-}\phi \cdot 1/4 \cdot m(m-1)$ (where m is the number of candidates). We point to their paper(s) for further explanations, intuitions, and ways of computing ϕ given $\text{norm-}\phi$.

[6] This view is spelled out, e.g., by Reijngoud and Endriss [40].

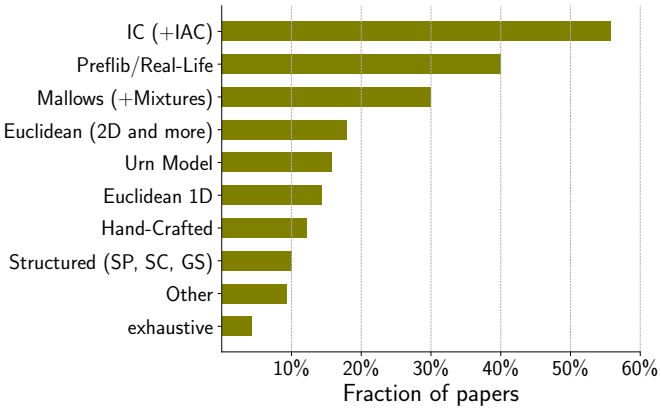


Figure 7: Fractions of papers that use given data sources for ordinal elections. "Hand-Crafted" refers to models designed specifically for a given paper. "Exhaustive" means generating all elections of a given size.

Pólya-Eggenberger Urn Model (Used in 15.7% of the Papers). The Pólya-Eggenberger urn model [17, 2] uses a nonnegative parameter of contagion $\alpha \in \mathbb{R}$, which corresponds to the level of correlation between the votes. Votes are generated iteratively as follows: We imagine an urn which initially contains one copy of each possible order; to generate a vote, we draw one from the urn, include its copy in the election, and return it to the urn, together with $\alpha \cdot m!$ copies, where m is the number of candidates.[7] For $\alpha = 0$ we get IC, and for $\alpha = 1/m!$ we get IAC [18].

Among the considered papers, 22 conducted experiments on the urn model. Typical values of α were $10/m!$, 0.05, 0.1, 0.2, 0.5, and 1. In a few papers, particularly regarding the map of elections, α was derived from the Gamma distribution with shape parameter $k = 0.8$ and scale parameter $\theta = 1$ (and this is how we generated the urn elections for Figure 6).

Euclidean Elections (Used in 23.6% of the Papers). Under a Euclidean model, we assume that the candidates and voters are represented as points in some d-dimensional Euclidean space. Typically, these points are sampled uniformly at random from a d-dimensional cube (usually $[0, 1]^d$, for $d = 1$ this is the Interval model, for $d = 2$ the Square model, and for $d = 3$ the Cube model). Occasionally other distributions are considered (such as various forms of Gaussian distributions and uniform distribution over a d-dimensional sphere; for $d = 2$ this is the Circle model and for $d = 3$ the Sphere model). Each voter's ranking is constructed so that he or she ranks candidates whose points are closer to his or hers higher than those whose points are further away.

Among the considered papers, 33 conducted experiments on Euclidean preferences. The most popular choice was the 2D setting (25 papers), followed by the 1D one (20 papers). Some papers additionally investigated higher dimensions, reaching up to the 20D model (e.g., Boehmer et al. [8] and conference papers that lead to the work of Szufa et al. [44]).

Single-Peaked Elections (Used in 10% of the Papers). Single-peakedness is one of the most prominent structured domains. An election is single-peaked [3] if there is an ordering of the candidates—the societal axis—such that for each voter, sweeping through the axis from left to right, the position of the corresponding candidates in the voter's ranking first increases and then decreases. Single-peaked elections are usually motivated by the fact that they cover applications in which there is an objective order of candidates; a typical example being the political left-to-right spectrum.

In practice, authors use two main methods to generate such elections. Both of them first select an axis

[7] This normalized variant is due to McCabe-Dansted and Slinko [36]; in the unnormalized variant the parameter gives the absolute number of the additional copies put back into the urn.

uniformly at random. The model proposed by Walsh [45] uses a uniform distribution over the votes that are single-peaked for the selected axis. In the model proposed by Conitzer [15], to generate a vote we first pick uniformly at random its top choice. Then, to fill the next position in the ranking, we flip a symmetric coin and either select the first unused candidate to the right or to the left of the top-choice one. We repeat the procedure until all positions are filled (or the remaining positions are uniquely determined).

While the Walsh approach seems more appealing as a single-peaked variant of impartial culture, the Conitzer approach is interesting because it gives elections very similar to the 1D-Euclidean ones (where the candidate and voter points are sampled uniformly at random from an interval). Consequently, multiple papers with experiments on both Walsh and Conitzer models show that they tend to give qualitatively different elections. Thus, when studying single-peaked elections, we recommend using both approaches.

Single-peakedness on a circle (SPOC) is a variant of single-peakedness where the axis is cyclic [38]. Sampling SPOC elections using the Conitzer's approach leads to a uniform distribution of such votes.

Single-Crossing Elections (Used in 5.7% of the Papers). An election is single-crossing if we can order either all the votes in a way that for every pair of candidates all the voters either prefer one of them to the other, or the relative preference between them changes exactly once when going from the first to the last vote in the ordering [37, 42]. It is unknown how to sample such votes uniformly at random in polynomial time (and, indeed, doing so might be challenging). Szufa et al. [44] give a sampling heuristic which seems reasonable, but makes no guarantees about its distribution (we use it in Figure 6).

Group-Separable Elections (Used in 3.6% of the Papers). A group-separable election [27, 28] can be characterized by a rooted, ordered tree whose leaves are candidates (Inada's definition was different, we follow an approach of Karpov [29]). Then, each vote in such an election must be obtainable by, first, reversing the order of children of arbitrary internal nodes of the tree (possibly none), and then reading the candidates from leaves from left to right. In the considered experiments, only group-separable elections with balanced or caterpillar trees were considered and the votes were drawn uniformly at random. Such elections do not resemble real-life data, but are different from elections given by any other culture (which is visible by their distinct position in the map), thus they can capture unusual phenomena, which might be hard to spot otherwise.

Which Models to Use? There is no clear answer as to which statistical cultures are the *best* in some objective sense. However, there are three natural approaches to choosing which models to use in a paper: First, one might want to cover as much of the space of elections as possible (this might mean including elections from structured domains, in addition to more common models). Second, one might know the nature of the real-life data that appears in a given phenomenon and might want to choose model(s) that generate similar elections. Finally, one might want to stick to realistic data, but without focusing on its specific type. In this case, results on the map of elections [44, 7, 22] suggest choosing cultures that land in a triangle between ID, UN, and Euclidean elections (for dimension 2 or higher). This might mean, e.g., using the Mallows model, urn models with fairly low contagion parameters, and Euclidean models (such as, e.g., the 5D-Cube).

5 Approval Elections

For an analysis of approval elections, we point to the full version of the paper [11]. Briefly put, we observed that real-life data is used much more often than in the ordinal case, i.e., in over 61.9% of the papers. Regarding synthetic elections, variants of Euclidean and IC models are clearly dominant. We suggest using at least one of them, for comparison. Other models received notably less attention, even

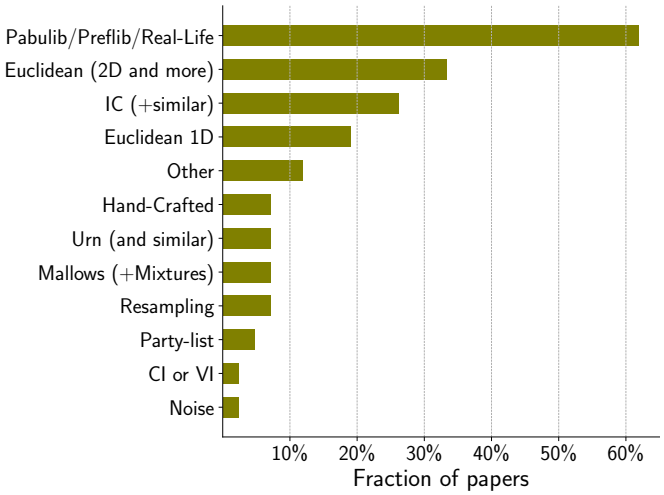


Figure 8: Fractions of papers that use given data sources for approval elections. "Hand-Crafted" refers to models specifically designed for a given paper. Mallows and urn models refer to generating ordinal elections according to these models and considering some top candidates from each vote as the approved ones.

though the resampling model of Szufa et al. [43] and the independent approval model of Faliszewski et al. [23] are quite natural (but also introduced so recently that they might have not caught on yet).

In Figure 8 we illustrate fractions of papers that use particular statistical cultures (or real-life data).

Which Models to Use? For approval elections, real-life data is the most common data source, appearing in a majority of papers with approval elections from the Guide. Pabulib [21] provides a rich collection of real-life elections from participatory budgeting exercises. Thus, especially for works on participatory budgeting, Pabulib is the most attractive and relevant data source. However, if one's work is not tailored to participatory budgeting, it might be a good idea to also consider other data sources, such as the real-world data from Boehmer et al. [10] or suitable elections from Preflib. For synthetic data, one may also consult the maps of Szufa et al. [43] to check for which parameter choices Euclidean models as well as the resampling model generate elections that are, in some sense, similar to those from Pabulib. Indeed, we feel that the resampling model could be quite an appealing model of generating synthetic approval elections, but so far there is little evidence to back this view (as the model was only introduced recently). It would also be interesting to analyze how elections generated according to various statistical cultures compare to approval elections from scenarios other than participatory budgeting.

6 Conclusions

Looking back, we see that impartial culture and real-life data are popular both in the ordinal and approval settings. While the ordinal world uses real-life data less frequently and fairly often considers structured domains, in the approval world the situation is the opposite. We hope that our analysis will help researchers to see current trends and approaches, and will allow them to design more conclusive experiments. For the ordinal setting, we suggest the use of real-life data, Euclidean models (especially with higher dimensions), normalized Mallows model, and urn elections (with small contagion parameter). Impartial culture is a yardstick to measure against previous papers, and structured domains can give otherwise difficult-to-spot insights. For approval elections, Pabulib is a natural and appealing source of real-life data (for participatory budgeting). As far as synthetic data goes, Euclidean models and the resampling models (and, possibly, their mixtures) seem appealing.

Acknowledgments This work was funded in part by the French government under the management of Agence Nationale de la Recherche as part of the France 2030 program, reference ANR-23-IACL-0008. T. Wąs was partially supported by EPSRC under grant EP/X038548/. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 101002854). The research presented in this paper is supported in part from the funds assigned by Polish Ministry of Science and Technology to AGH University.

References

[1] Y. Bachrach, O. Lev, Y. Lewenberg, and Y. Zick. Misrepresentation in district voting. In *Proceedings of IJCAI-2016*, pages 81–87, 2016.

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

[3] D. Black. *The Theory of Committees and Elections*. Cambridge University Press, 1958.

[4] D. Bloembergen, D. Grossi, and M. Lackner. On rational delegations in liquid democracy. In *Proceedings of AAAI-2019*, pages 1796–1803, 2019.

[5] N. Boehmer. *Application-oriented collective decision making: Experimental toolbox and dynamic environments*. PhD thesis, Technical University of Berlin, Germany, 2023. URL https://nbn-resolving.org/urn:nbn:de:101:1-2023112900583621809963.

[6] N. Boehmer and N. Schaar. Collecting, classifying, analyzing, and using real-world ranking data. In *Proceedings of AAMAS-2023*, pages 1706–1715, 2023.

[7] N. Boehmer, R. Bredereck, E. Elkind, P. Faliszewski, and S. Szufa. Expected frequency matrices of elections: Computation, geometry, and preference learning. In *Proceedings of NeurIPS-2022*, 2022.

[8] N. Boehmer, J.-Y. Cai, P. Faliszewski, A. Z. Fan, Ł. Janeczko, A. Kaczmarczyk, and T. Wąs. Properties of position matrices and their elections. In *Proceedings of AAAI-2023*, pages 5507–5514, 2023.

[9] N. Boehmer, P. Faliszewski, and S. Kraiczy. Properties of the Mallows model depending on the number of alternatives: A warning for an experimentalist. In *Proceedings of ICML-2023*, pages 2689–2711. PMLR, 2023.

[10] N. Boehmer, M. Brill, A. Cevallos, J. Gehrlein, L. Sánchez Fernández, and U. Schmidt-Kraepelin. Approval-based committee voting in practice: A case study of (over-)representation in the Polkadot blockchain. In *Proceedings of AAAI-2024*, 2024. Accepted for publication.

[11] N. Boehmer, P. Faliszewski, Ł. Janeczko, A. Kaczmarczyk, G. Lisowski, G. Pierczyński, S. Rey, D. Stolicki, S. Szufa, and T. Wąs. Guide to numerical experiments on elections in computational social choice. Technical Report arXiv:2402.11765, arXiv.org, 2024.

[12] A. Borodin, O. Lev, N. Shah, and T. Strangway. Big city vs. the great outdoors: Voter distribution and how it affects gerrymandering. In *Proceedings of IJCAI-2018*, pages 98–104, 2018.

[13] F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. Procaccia, editors. *Handbook of Computational Social Choice*. Cambridge University Press, 2016.

[14] R. Colley, T. Delemazure, and H. Gilbert. Measuring a priori voting power in liquid democracy. In *Proceedings of IJCAI-2023*, pages 2607–2615, 2023.

[15] V. Conitzer. Eliciting single-peaked preferences using comparison queries. *Journal of Artificial Intelligence Research*, 35:161–191, 2009.

[16] J. Doucette and R. Cohen. A restricted markov tree model for inference and generation in social choice with incomplete preferences. In *Proceedings of AAMAS-2017*, pages 893–901, 2017. URL https://dl.acm.org/doi/10.5555/3091125.3091251.

[17] F. Eggenberger and G. Pólya. Über die statistik verketteter vorgänge. *ZAMM-Journal of Applied*

Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik, 3(4):279–289, 1923.

[18] Ö. Eğecioğlu and A. Giritligil. The impartial, anonymous, and neutral culture model: A probability model for sampling public preference structures. *Journal of Mathematical Sociology*, 37(4):203–222, 2013.

[19] P. Faliszewski, P. Skowron, A. Slinko, and N. Talmon. Multiwinner voting: A new challenge for social choice theory. In U. Endriss, editor, *Trends in Computational Social Choice*. AI Access Foundation, 2017.

[20] P. Faliszewski, P. Skowron, A. Slinko, S. Szufa, and N. Talmon. How similar are two elections? In *Proceedings of AAAI-2019*, pages 1909–1916, 2019. doi: 10.1609/AAAI.V33I01.33011909.

[21] P. Faliszewski, J. Flis, D. Peters, G. Pierczynski, P. Skowron, D. Stolicki, S. Szufa, and N. Talmon. Participatory budgeting: Data, tools and analysis. In *Proceedings of IJCAI-2023*, pages 2667–2674, 2023. doi: 10.24963/IJCAI.2023/297.

[22] P. Faliszewski, A. Kaczmarczyk, K. Sornat, S. Szufa, and T. Wąs. Diversity, agreement, and polarization in elections. In *Proceedings of IJCAI-2023*, pages 2684–2692, 2023.

[23] P. Faliszewski, Ł. Janeczko, A. Kaczmarczyk, M. Kurdziel, G. Pierczyński, and S. Szufa. Learning real-life approval elections. In *Proceedings of AAMAS-25*, 2025.

[24] P. Faliszewski, J. Mertlova, P. Nunn, S. Szufa, and T. Wąs. Distances between top-truncated elections of different sizes. In *Proceedings of AAMAS-25*, 2025.

[25] P. Fishburn and W. Gehrlein. Condorcet paradox and anonymous preference profiles. *Public Choice*, 26:1–18, 1978.

[26] G. Guilbaud. Les théories de l'intéret genéral et le problémelogique de l'agrégation. *Economie Appliquée*, 5:501–584, 1952.

[27] K. Inada. A note on the simple majority decision rule. *Econometrica*, 32(32):525–531, 1964.

[28] K. Inada. The simple majority decision rule. *Econometrica*, 37(3):490–506, 1969.

[29] A.5 Karpov. On the number of group-separable preference profiles. *Group Decision and Negotiation*, 28(3):501–517, 2019.

[30] J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[31] K. Kuga and H. Nagatani. Voter antagonism and the paradox of voting. *Econometrica*, 42(6): 1045–1067, 1974.

[32] M. Lackner and P. Skowron. *Multi-Winner Voting with Approval Preferences*. Springer, 2023.

[33] J. Lang and L. Xia. Voting in combinatorial domains. In F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia, editors, *Handbook of Computational Social Choice*, chapter 9, pages 197–222. Cambridge University Press, 2016.

[34] C. Mallows. Non-null ranking models. *Biometrica*, 44:114–130, 1957.

[35] N. Mattei and T. Walsh. Preflib: A library for preferences. In *Proceedings of ADT-2013*, pages 259–270, 2013.

[36] J. McCabe-Dansted and A. Slinko. Exploratory analysis of similarities between social choice rules. *Group Decision and Negotiation*, 15:77–107, 2006.

[37] J. Mirrlees. An exploration in the theory of optimal income taxation. *Review of Economic Studies*, 38:175–208, 1971.

[38] D. Peters and M. Lackner. Preferences single-peaked on a circle. *Journal of Artificial Intelligence Research*, 68:463–502, 2020.

[39] D. Peters, G. Pierczynski, N. Shah, and P. Skowron. Market-based explanations of collective decisions. In *Proceedings of AAAI-2021*, pages 5656–5663. AAAI Press, 2021.

[40] A. Reijngoud and U. Endriss. Voter response to iterated poll information. In *Proceedings of AAMAS-2012*, pages 635–644, 2012. URL http://dl.acm.org/citation.cfm?id=2343787.

[41] S. Rey and J. Maly. The (computational) social choice take on indivisible participatory budgeting. Technical Report arXiv.2303.00621 [cs.GT], arXiv.org, 2023.

[42] K. Roberts. Voting over income tax schedules. *Journal of Public Economics*, 8(3):329–340, 1977.

[43] S. Szufa, P. Faliszewski, Ł. Janeczko, M. Lackner, A. Slinko, K. Sornat, and N. Talmon. How to sample approval elections? In *Proceedings of IJCAI-2022*, pages 496–502, 2022.

[44] S. Szufa, N. Boehmer, R. Bredereck, P. Faliszewski, R. Niedermeier, P. Skowron, A. Slinko, and N. Talmon. Drawing a map of elections. *Artificial Intelligence*, 343:Article 104332, 2025.

[45] T. Walsh. Generating single peaked votes. Technical Report arXiv:1503.02766 [cs.GT], arXiv.org, March 2015.

[46] B. Wilder and Y. Vorobeychik. Defending elections against malicious spread of misinformation. In *Proceedings of AAAI-2019*, pages 2213–2220, 2019.

Appendix

A Paper Screening Process

All the papers downloaded from our conferences had to pass initial screening. For a paper to pass this screening it had to include words related to both elections and experiments. For both categories it had to include some keyword on at least two pages. We have used the following keywords:[8]

> Elections: `electi`, `lection`, `vot`, `ballo`, `allot`.
>
> Experiments: `experime`, `periment`, `empiri`, `piric`, `pirical`, `simulatio`, `mulation`, `mulations`

However, if a matching word also contained as a substring one of the following forbidden words, then it was disregarded (to decrease the number of false positives):

> `accumul`, `balloon`, `vot20`, `vottir`, `preselection`, `flection`, `formulation`, `collection`, `selection`, `pivot`, `devot`, `prelection`, `allotment`, `allotted`.

The somewhat strange form of our keywords is due to the fact that extracting text from PDF files is sometimes inaccurate and words can be broken into parts at unexpected places. Thus we selected keywords that avoided many such problems.

A fairly large number of papers passed our basic screening criterion without actually considering numerical experiments on elections. This was intended: We wanted our filter to be fairly nonrestrictive, so that we would have as few *false negatives* as possible. Below we list typical reasons why many papers were *false positives*:

1. Authors make a passing remark to voting.
2. Word "election" is recognized as part of "selection." (Due to inaccurate text extraction, this happens even though we put "selection" on our list of forbidden words.)
3. Studying text data that includes political discussions.
4. An election-related word appears commonly in another subarea, such as, e.g., "VOTER" in some community detection papers, "voted-perceptron" in connection to learning, or the "VOT" dataset studied in some papers.
5. Using majority voting as a tool for classifiaction or to aggregate data.
6. A form of voting is used by the authors to gather some sort of data, or to aggregate data from a questionaire, but in a way that is not relevant to our work.

The above reasons typically apply to papers that clearly are out of scope for our work. Below we mention several reasons to not include computational social choice papers in the Guide:

1. Papers looking at two candidates only.

[8] Text recovered from PDF files is often faulty in the sense that two words may end up glued together, or may have some unexpected symbols added before or after. This is why we consider prefixes and suffixed and our keywords have somewhat specific form.

2. Papers that do not actually include experiments, but simply discuss their possibility and/or desirability.

3. Papers discussing issues that are close to voting, but nonetheless the model used is too far from the kind of elections that we consider (examples included aggregating graphs or dependence structures in multiissue domains).

B Analyzed Papers

In the full version of the paper [11] we list all the papers that made it to the Guide, together with some notes about the experiments that they include. For each paper we include:

1. The title, authors, year of publication, and the conference where it appeared, together with a reference to the bibliography.

2. A list of experiments that we recorded for it (each experiments starts either with letter "O" for ordinal and "A" for approval, followed by experiment number and colon).

3. For each experiment we list the statistical cultures/real-life data sources used, numbers of samples per data point, followed by the considered election sizes (see explanation below for the format used). For some experiments we include additional notes, e.g., related to the parameters used in various statistical cultures, or comments regarding the paper/experiment.

To record the sizes of the considered elections, we write `CxV`, where `C` is a string describing the number of candidates and `V` is the string describing the number of voters. Each such string can either be of the form `{a,b,c, ...}`, in which case it is simply an enumeration of possible values, or of the form `[a,b]`, in which case it represents an interval of values $\{a, a+1, \ldots, b\}$. For example, string `{5,10}x[20,100]` refers to a set of elections that either have 5 or 10 candidates and between 20 and 100 voters. Authors often consider elections where some parameter—such as the number of voters—changes with a particular step (e.g., one could consider between 20 and 100 voters, with a step of 5). We have decided to omit such details (on the one hand, this simplified the process of recording data and, on the other hand, we felt that availability of such data would not affect our analysis too strongly and interested readers would consult specific papers when needed).

Remark B.1. *In the table below we present the main contents of the Guide (i.e., our database). For all the papers we tried to find as much relevant information as we could, mostly relying on the paper itself (but occasionally we referred to the full version, if it were available). For some of the papers we recorded some details that we found interesting, but we did not follow any specific rule in this regard. Hence, for some papers we are (most likely) missing such comments. We stress that many comments/details for the papers are written in a very concise way. We expect to extend (some of) them as the Guide project progresses.*

Remark B.2. *For some papers we omit certain details, such as the number of samples per data points or election sizes. This happens, e.g., if such data is not relevant to a given paper (e.g., a paper evaluating some property of every real-life election from some set would have to list "one" as the number of samples, which would feel silly) or if it too difficult/cumbersome to obtain this data (e.g., recording precisely the sizes of elections from a number of considered real-life datasets).*

We occasionally write that some details are unclear in a given paper. This means that we tried to identify the respective bit of information and we failed. We will update the Guide as we learn such information (provided it is indeed included in the given paper and we missed it).

Remark B.3. *Whenever we could easily find a journal version of a paper, we included a reference to it. In some cases we knew of journal versions with different titles than their conference predecessors, but we generally did not seek them explicitly. For papers without journal versions, we attempted to locate their arXiv versions (but we could have failed whenever the authors changed the title).*

C Skipped Papers

Occasionally, links associated to the papers in the DBLP website were either missing or corrupted. It was often easy to download the papers manually after finding them (by title) on the official webpages of the respective venues. However, for 34 such troublesome papers we could not find trustworthy sources related to the corresponding proceedings publisher to download them from.

In the full version of the paper [11] we list these papers, including titles, authors, venues, tracks, and reasons for skipping the papers. The list contains:

- 14 papers from the Student Abstracts track of AAAI-2010;
- 15 papers from the Doctoral Consortium of AAAI-2011;
- 2 papers from the Special Track on AI and the Web of AAAI-2011;
- 2 papers from the Special Track on Computational Sustainability and AI of AAAI-2011;
- 1 paper from the Machine Learning Applications track of IJCAI-2019.

Out of the above papers, for 32 the DBLP webpage contained links to the Wayback Machine—a crawler that archives webpages—as the original links were expired. However, we were unable to access the respective PDF files from the Wayback links; instead, we only could read the abstracts. The remaining 2 papers had no links at all to the respective PDF files on the corresponding official proceedings webpage.

Niclas Boehmer
Hasso Plattner Institute
Potsdam, Germany
Email: niclas.boehmer@hpi.de

Piotr Faliszewski
AGH University of Krakow
Kraków, Poland
Email: faliszew@agh.edu.pl

Łukasz Janeczko
AGH University of Krakow
Kraków, Poland
Email: ljaneczk@agh.edu.pl

Andrzej Kaczmarczyk
University of Chicago
Chicago, USA
Email: akaczmarczyk@uchicago.edu

Grzegorz Lisowski
AGH University of Krakow
Kraków, Poland
Email: glisowski@agh.edu.pl

Grzegorz Pierczyński
University of Warsaw
Warsaw, Poland
Email: g.pierczynski@mimuw.edu.pl

Simon Rey
Currently unaffiliated

Dariusz Stolicki
Jagiellonian University
Kraków, Poland
Email: dariusz.stolicki@uj.edu.pl

Stanisław Szufa
Université Paris Dauphine-PSL
Paris, France
Email: s.szufa@gmail.com

Tomasz Wąs
University of Oxford
Oxford, UK
Email: tomasz.was@cs.ox.ac.uk