Received: 00 Month 0000 | Revised: 00 Month 0000 | Accepted: 00 Month 0000

DOI: xxx/xxxx

ARTICLE TYPE

Group Recommendation with Noisy Subjective Preferences

Amirali Salehi-Abari*[1] | Kate Larson[2]

[1]Faculty of Business and IT, Ontario Tech University, Ontario, Canada
[2]Cheriton School of Computer Science, University of Waterloo, Ontario, Canada

Correspondence
*Amirali Salehi-Abari. Email: abari@uoit.ca

Social choice theory provides a principled framework for the aggregation of individuals' preferences in support of group decision-making and recommendation. Much of this work, however, either assumes that individuals' subjective preferences (and thus, their votes) are correctly specified by the individuals themselves, or alternatively that the votes of individuals are noisy estimates of some underlying ground truth over rankings of alternatives. We argue that neither model appropriately addresses some of the issues which arise in the context of group-recommendation domains where individuals have subjective preferences but for some reason (e.g., the high cognitive burden, concerns about privacy, etc.) may instead vote using a noisy estimate of their subjective preference rankings. In this paper, we propose a general probabilistic framework for modeling noisy subjective preferences, and explore the accuracy and reliability of four well-studied voting rules under various noise models. Our results demonstrate that there is no single reliable method amongst the examined methods. Specifically, we observe the change in noise distribution can flip one method from being the most reliable to the least.

KEYWORDS:
Group Recommendation, Noisy Subjective Preferences, Computational Social Choice.

1 | INTRODUCTION

Group decision problems involve making decisions for a group of individuals who have their own personal and possibly conflicting preferences. Group decision problems are prevalent (e.g., decisions for social groups, business organizations, public policies, etc.). To assist group decision making, group recommender systems are developed in various domains such as tourism[1], crowdfunding[2], music[3], news/web pages[4], TV programs[5], and movies[6]. These group recommender system require understanding user preferences (through preference learning or elicitation) in order to effectively recommend an option to a group of users.

Individual subjective preferences (in the form of ratings, rankings over alternatives, etc.) are constantly elicited or learnt whether in the context of high-stakes political elections, when making dinner decisions amongst friends, or when being asked to rate/rank movies in order to improve suggestions made by recommender systems. Though individual preferences can be represented in various forms[7], preferences are usually given by rankings in much of social choice theory which provides principled frameworks for aggregating individuals' preferences in support of group-decision making and recommendations[8,9]. This is because preference rankings help circumvent, to some extent, the problem of interpersonal comparisons of utilities[10].

In this context, we initiate a study of uncertainty with respect to subjective preference rankings, with a particular focus on how noise in revealed *subjective preferences* influences different voting rules (or aggregation rules) from the social choice literature. Noisy reporting of subjective preferences has long been documented in recommender-system research[11,12,13]. Individuals may reveal preference rankings which differ from their true underlying preferences for numerous reasons, and thus revealed

preferences may sometimes be viewed as noisy samples of true underlying preferences. For example, *high cognitive cost* or simple inexperience across all alternatives, may result in an individual being incapable of accurately specifying a total ranking over alternatives when queried. Individuals may also simply make *mistakes* when reporting their preference rankings (e.g. erroneously changing the order of some option on forms or user interfaces), or individuals may purposefully misreport preference rankings, not with intent of manipulating the system, but for privacy reasons.[1] For example, an individual may be reluctant to reveal their passion for country music in certain social settings, or may prefer to not reveal certain preferences so as to avoid sharing too much information with advertisers.

We propose a general probabilistic framework for modeling *noisy subjective preferences*. We also present an empirical methodology for studying noises in subjective preferences and their impact on group decision making and aggregation tasks. Under this framework, we analyze four well-known voting rules (Borda, Plurality, Kemeny, and Copeland) on various noise models, preference data, group sizes, and preference distributions. Our findings demonstrate the degree by which different voting rules are tolerant to noise, and illustrate that each rule is highly sensitive to the underlying noise model. That is, it is not possible to rank the voting rules in terms of general robustness and reliability against noise. Our models and empirical findings raise a number of interesting discussion points and future research directions in both theory and practice.

2 | RELATED WORK

We review the related work on group recommendation methods, noisy preferences and robustness of voting rules, noisy objective preferences, preference elicitation and group recommendation with incomplete preferences, and preference ranking learning.

2.1 | Group Recommendation

Group recommendation methods can be widely classified as follows: (i) *Artificial/Virtual profile methods*[14], where joint artificial user profiles for each group of users are created to keep track of their joint revealed/elicited preferences; (ii) *Profile-merging methods*[5,15], which merge group members' individual profiles to form a group profile, based on which recommendations are made; (iii) *Recommendation/scoring aggregation* methods[16,17,18,19,20,21,22,23], which aggregate the recommendations (or inferred preferences) for each group member into single group recommendation list (or recommended option). This aggregation is usually conducted by a *group consensus function* (or social choice function). See Felfernig *et al.*[9] for a detailed overview of group recommender systems.

Our focus in this paper is on the third category. While assuming the individual preferences are inferred or elicited, we narrow our focus on the robustness of social choice functions (or group consensus functions) against noise in inferred/elicited subjective preferences.

2.2 | Noisy Preferences and Robustness

Procaccia *et al.*[24] studied worst-case robustness of voting rules for noisy preferences where their noise model chose k pairs of adjacent candidates (in a worst case preference profile) uniformly at random and swapped them. A similar noise model was also used by Shiryaev *et al.*[25]. Our work differs from this literature in several ways. Instead, we are interested in studying various forms of aggregation rules and methods for dealing with noisy subjective preferences in the average case (i.e., probabilistic framework), rather than in the worst case. To this end, we propose a general probabilistic generative model for noisy preferences, explaining how noisy preferences arise from true subjective preferences.

2.3 | Preference Elicitation and Group Recommendation with Incomplete Preferences

There has been recent interest in elicitation and aggregation of uncertain or incomplete preferences. Examples include elicitation of voter's preference distribution over rankings[26], aggregation of incomplete subjective preferences[27,28], and aggregation of incomplete preferences over social networks for group recommendation[21]. Our work differs in several ways. We do not expect voters to be aware of their uncertainty of their own preferences, nor do we assume they can explicitly report it. Rather than

[1]There is a large body of research on manipulation of voting rules[8]. In this paper, we do not address strategic issues and make the explicit assumption that the misreporting of preferences is non-manipulative in nature.

assuming that the preferences are noise-free, we assume that full-rankings are observed (or inferred through preference-learning methods) but are noisy.

2.4 | Noisy Objective Preference

An alternative interpretation of voting is to view votes as *noisy realizations* of some objective ground truth ranking, with voting rules interpreted as maximum likelihood estimators of the correct outcomes[29,30,31,32,33,34,35]. In this setting, there exists a "correct" ranking (i.e., ground truth ranking) and each voter's ranking corresponds to a noisy realization of this correct ranking. Thus, if a noise model is given, one can compute the maximum likelihood estimate of the correct outcome[29,36]. Conitzer and Sandholm[30] studied common voting rules to determine for which there exists a noise model such that the rule can be interpreted as maximum likelihood estimate. The maximum likelihood approach is extended to partial orders[31], and is studied in multi-issue domains[37], and for selecting a subset of alternatives[38]. Assuming that there is a ground truth ranking, Caragiannis *et al.*[32] studied the number of votes that a voting rule needs to reconstruct the true ranking. Also, voting rules under adversarial noise model are studied[39]. Our work is distinguished from this literature by lack of the assumption of the common correct ranking for all voters; i.e., individual rankings are *subjective* in our model rather than objective.

2.5 | Preference Ranking Learning

There is a rich literature on learning preference models[40,41,42,43]. Gormley and Murphy[40] develop learning algorithms for mixtures of both Placket-Luce models and Benter models[44]. A spatial model combined with Placket-Luce model is deployed for exploring voting data[45]. Murphy and Martin[46] employ a mixture of distance-based ranking models to describe individual preferences (in the form of full rankings) from a heterogeneous population. Similarly, Busse *et al.*[41] learn a mixture of ranking models for partial preferences of the top-t type (i.e., individuals have ranked their t favourites out of m items). Lu and Boutilier[42] relax the restriction on t-type partial rankings by representing partial rankings as pairwise comparisons. More recently, the learning of the mixtures of distance-based ranking models with the generalized weighted distance metric has been studied[47]. Azari *et al.*[43] studied conditions on exponential families of random utility models (e.g., Thurstone and its variants) under which fast inference within a Bayesian framework is possible. Salehi-Abari and Boutilier developed inference methods[21] and probabilistic models[48] of preference rankings correlated over social networks. Our focus is a little different than this literature: while we use some probabilistic machinery and frameworks from this field, our main focus is on understanding how robust preference aggregation rules are with respect to noise.

3 | A MODEL FOR NOISY SUBJECTIVE PREFERENCES

We here present our stochastic noise process in which we differentiate true subjective preferences from revealed noisy subjective preferences.[2]

3.1 | Ordinal Preferences

We consider a set of m alternatives $\mathcal{A} = \{a_1, \dots, a_m\}$ and a set of n individuals (or agents) $\mathcal{N} = \{1, \dots, n\}$. A *strict preference* relation $\succ_i$ for agent $i \in \mathcal{N}$ over $\mathcal{A}$ is a binary, *transitive*, *asymmetric*, and *total (or complete)* relation, where the notation $x \succ_i y$ is interpreted as alternative x is preferred to alternative y by agent i. Given a strict preference ordering, $\succ_i$, this uniquely defines a *ranking*, r_i, over $\mathcal{A}$, where for any $x, y \in \mathcal{A}$, if $x \succ_i y$ then x is ranked above y in r_i. Similarly, given a ranking $r_i \in \Omega(\mathcal{A})$ where $\Omega(\mathcal{A})$ is the set of permutations over $\mathcal{A}$, there is a corresponding unique strict preference ordering. Thus, in the rest of this paper we work in the space of rankings over $\mathcal{A}$, and use the term ranking, preference ranking, and preference ordering interchangeably.

3.2 | Noisy Subjective Preferences

Given individual $i \in \mathcal{N}$, we distinguish between its *true preference ranking* $r_i \in \Omega(\mathcal{A})$ and its *observed/revealed preference ranking* $\tilde{r}_i \in \Omega(\mathcal{A})$. We assume that for each $i \in \mathcal{N}$, its true preference ranking r_i is independent and identically

[2] Through this paper, a true subjective preference refers to a noise-free subjective preference.

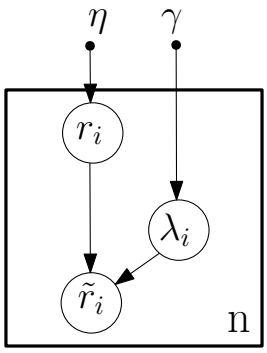


FIGURE 1 Probabilistic model: true preferences $\{r_i\}$, revealed preferences $\{\tilde{r}_i\}$, and noise parameters $\{\lambda_i\}$.

distributed by a parameterized ranking distribution $p(r|\boldsymbol{\eta})$ where $\boldsymbol{\eta}$ is the vector of parameters: $r_i \sim p(r|\boldsymbol{\eta})$. Many possible options exist for this ranking distribution such as distance-based models (*e.g.*, ϕ-Mallows), multistage models (*e.g.*, Plackett-Luce), paired comparison models, mixture models, and spatial models. See Marden[49] for an excellent overview on models of ranking distributions.

We assume that individual i's *revealed preference ranking*, $\tilde{r}_i$, is a noisy observation of its true preference r_i. These noisy observations might arise due to various reasons; for example, individuals might make mistakes when reporting their preference rankings or may purposefully misreport their preference for privacy reasons. In particular, we assume that $\tilde{r}_i$ is drawn from a conditional distribution $g(r|r_i, \lambda_i)$:

$$\tilde{r}_i \sim g(r|r_i, \lambda_i),$$

where λ_i is the parameter controlling the extent of noise for individual i. Note that λ_i can be different for each individual, thus allowing us to model a population with heterogeneous levels of noise in observed preferences. We consider a Bayesian framework by assuming λ_i are drawn independently and identically from a prior distribution $h(x|\boldsymbol{\gamma})$ with the vector of parameters $\boldsymbol{\gamma}$. Figure 1 depicts the graphical representation of our probabilistic model.

One plausible class of ranking distributions well-suited for characterizing this conditional distribution is *distance-based ranking distributions*[50,49]. These distributions have ranking probabilities that decrease with increasing distance from a "modal" or "reference" ranking $\sigma \in \Omega(\mathcal{A})$:

$$\mathbb{P}(r|\sigma, \omega) = \frac{1}{\psi(\omega)} \exp(-\omega d(r, \sigma)), \tag{1}$$

where $\omega \in [0, \infty)$ is a concentration or dispersion parameter, $\psi(\omega)$ is a normalizing constant, and $d(r, \sigma)$ is a distance between r and σ. As $\omega \to \infty$, P becomes concentrated at the reference ranking σ, whereas for $\omega = 0$, P is the uniform distribution over $\Omega(\mathcal{A})$. Distance-based ranking models differ in the choice of distance metrics.

The widely-used Mallows ϕ-model is an example of distance-based ranking distribution with d being Kendall's τ distance d_τ, measuring the minimum number of pairwise swaps required to transform one ranking to another one. One can write the probability of ranking r under Mallows ϕ-model by

$$\mathbb{P}(r|\sigma, \phi) = \frac{1}{Z(\phi)} \phi^{d_\tau(r,\sigma)}, \tag{2}$$

where $\phi = exp(-\omega)$ is dispersion parameter for controlling the extent of noise, and $Z(\phi)$ is a normalizing constant. The ϕ-mallows model has been widely studied for modeling noise in objective preferences[29,30,31,37,32].

Example: ϕ-Mallows noise model. We can model an individual i's noisy revealed preference $\tilde{r}_i$ of his/her true preference ranking r_i using Mallows' ϕ-model. To do so, we assume that the modal ranking σ is i's true ranking, r_i, and thus the probability of having some observed ranking $\tilde{r}_i$ is given by the conditional distribution

$$g(\tilde{r}_i|r_i, \phi_i) = \frac{1}{Z(\phi_i)} \phi_i^{d_\tau(\tilde{r}_i, r_i)}, \tag{3}$$

where ϕ_i is dispersion parameter for controlling the extent of noise in i's revealed preference.

4 | EXPERIMENTAL METHODOLOGY

Our proposed stochastic process can be deployed for studying the robustness of preference aggregation methods or voting rules against noisy subjective preference rankings. We describe a general experimental methodology, which can be used for such studies. The general idea of our methodology is that for each aggregation method, we measure its robustness by comparing the group preference aggregated from noisy preferences with the group preference aggregated from true preferences. The larger this difference is, the less robust the aggregation method is. In Section 5, we apply this methodology and study four well-known voting rules.

Our methodology starts with generating true preferences for a group of n agents. We assume that individual true preferences over $\mathcal{A}$ are drawn independently from a ranking distribution (e.g., Mallow's model or Plackett-Luce model) or a real-world preference dataset (e.g., Sushi dataset). We let R denote the set of true preference ranking of those n agents. We then generate a *noisy* ranking for each individual in the group using an instance of noise model class (e.g., distance-based ranking models discussed above). We let $\tilde{R}$ represent the set of observed noisy preference ranking for agents in the group.

To examine the effect of noise for aggregation method F, we distinguish between the aggregated ranking under true preferences $r_{Agg} = F(R)$ and aggregated ranking under observed preferences $\tilde{r}_{Agg} = F(\tilde{R})$. We compute the extent to which these two aggregated rankings differ from each other using two different metrics.

The *Scaled Kendall-Tau Distance (SKTD)*,

$$SKTD(r_{Agg}, \tilde{r}_{Agg}) = \frac{2}{m(m-1)} d_\tau \left(r_{Agg}, \tilde{r}_{Agg}\right),$$

measures how close the aggregated ranking given true preferences r_{Agg} is to the aggregated ranking given observed noisy preferences $\tilde{r}_{Agg}$. Here, d_τ is Kendall's τ distance, measuring the minimum number of pairwise swaps required to transform one ranking to another one. We note that $SKTD(r_{Agg}, \tilde{r}_{Agg}) \in [0, 1]$ where 0 is the case where $r_{Agg} = \tilde{r}_{Agg}$, and 1 represents the maximum possible difference. Thus, the lower the $SKTD$ is for a particular voting rule or aggregation method, the more robust that rule is to noise.

Our SKTD metric is of special importance, when aggregation method is used in practice for outputting the group preference ranking over all alternatives. However, sometimes, aggregation method only intend to find the top ranked item for the group. So, we introduce another comparison metrics for comparing the top ranked item in the aggregated rankings to suit better top-ranked recommendation. The *Disaggreement Distance (DD)* is defined as

$$DD(r_{Agg}, \tilde{r}_{Agg}) = \mathbb{1}[r^{-1}_{Agg}(1) = \tilde{r}^{-1}_{Agg}(1)],$$

where, $\mathbb{1}[.]$ is the indicator function and $r^{-1}(1)$ represents the item ranked first in the ranking r. Note that $DD(r_{Agg}, \tilde{r}_{Agg}) \in \{0, 1\}$ where 0 represents the case where the top ranked item is the same in both r_{Agg} and $\tilde{r}_{Agg}$. We note that DD is only sensitive to the agreement of top-ranked items and is not impacted with the order of other items in the aggregated rankings.

For each setting (e.g., fixed noise model, true preference model, group size n, etc.), one can generate large number of instances and report the average of SKTD and DD over those instances for the aggregation method under investigation.

5 | EMPIRICAL EXPERIMENTS

We report on a series of experiments where we measure the noise tolerance of several well-studied voting rules. Our general goal is to analyse their robustness under various settings including true preference distributions, noise models, group sizes, the number of alternatives. We intend to understand which voting rule is the most robust rule for each setting and if any of our examined voting rules is dominant under all settings (i.e., it is always more robust than others).

5.1 | Experimental Setup

Before presenting our experiments and their results, we discuss our experimental setup including examined true preference distributions, noise models, voting rules, and group sizes.

True Preferences. We assume that individual true preferences over $\mathcal{A}$ are drawn independently from a ranking distribution or a real-world preference data set.

We consider two variants of ranking distributions: *unimodal* and *bimodal*. For unimodal, we consider a ϕ-Mallows model parameterized by dispersion ϕ_D and reference ranking σ_D (see Eq. 3). Our unimodal distribution captures the scenarios in which there is one most popular ranking (i.e., reference ranking), and the popularity of other rankings decreases with their distances to the reference ranking. We also consider bimodal distributions. We specifically focus on a two-component mixture of ϕ-Mallows model specified by probability distribution

$$\mathbb{P}(r|\boldsymbol{\pi}, \boldsymbol{\sigma}, \boldsymbol{\phi}) = \sum_{i=1}^{2} \frac{\pi_i}{Z(\phi_i, m)} \phi_i^{d_\tau(r,\sigma_i)},$$

where π_i, ϕ_i, σ_i are the mixing proportion, dispersion parameter and reference ranking for component i, respectively. To reduce the number of parameters in our experiments, we let $\pi_1 = \pi_2$, $\sigma_2 = \overleftarrow{\sigma_1}$, and $\phi_1 = \phi_2$, where $\overleftarrow{\sigma_1}$ is the reverse ranking of reference ranking σ_1 (i.e., the ranking with maximum distance to σ_1). Thus, we can re-parameterize our bimodal distribution by dispersion ϕ_D and reference ranking σ_D; so

$$\mathbb{P}(r|\phi_D, \sigma_D) = \frac{0.5}{Z(\phi_D, m)} (\phi_D^{d_\tau(r,\sigma_D)} + \phi_D^{d_\tau(r,\overleftarrow{\sigma_D})}).$$

Our bimodal distribution captures the scenarios when there are two popular maximally distinct rankings, and the popularity of other rankings decrease with the distance from these two modals.

For unimodal and bimodal distributions, we fix $\sigma_D = (1, \cdots, m)$ but vary ϕ_D over $\{0.5, 0.75, 0.9, 1.0\}$. For $\phi_D = 1.0$, the ranking distribution is *impartial culture*, in which all rankings are equally likely. In contrast, with $\phi_D = 0.5$, rankings will be distributed very close to reference ranking(s). We note that our choice of σ_D is arbitrary, and the results are the same for any other σ_D due to the symmetric nature of ranking space.

We also run extensive experiments on true preferences drawn from real-world preferences from the 2002 Irish Election[51] and Sushi data sets[52]. The Sushi and Irish datasets were obtained from the websites of the original owner of the datasets. The Irish Election data consists of two data sets: Dublin West and Dublin North. Dublin West (resp. Dublin North) consists 29,989 (resp. 43,942) ballots of the top-t form, of which 3800 (resp. 3662) are complete rankings. The Sushi dataset consists of 5000 complete preference rankings over varieties of sushi. For all three data sets, we created preference data sets with various values m from their complete preferences, by choosing m candidates and limiting each individual's preferences to these m options. While it is true that we have no way of checking whether the revealed preferences in these data sets correspond exactly with people's true preferences, we argue that testing our techniques on these observed preferences is still enlightening.

Noise Models for Preferences. Given the true ranking, r_i, of individual i we generate a *noisy* ranking for the individual using three different classes of noise models. The classes of noise models differ based on how they distribute the noise. The *entire* model draws noisy rankings for individual i from a ϕ-Mallows distribution with dispersion parameter ϕ_N and reference ranking r_i. The *top* noise model fixes the bottom $\frac{m}{2}$ items in the ranking r_i and then applies a ϕ-Mallows model with dispersion ϕ_N on the top $\frac{m}{2}$ items, whereas the *bottom* noise model does the reverse. In both these models noise is isolated to only part of the ranking. We believe each of these models can be a valid model and their validity is context dependent.[3]

We vary ϕ_N over $\{0.25, 0.5, 0.75, 0.8, 0.9\}$ in our experiments. Unless noted otherwise, all experiments use the entire model.

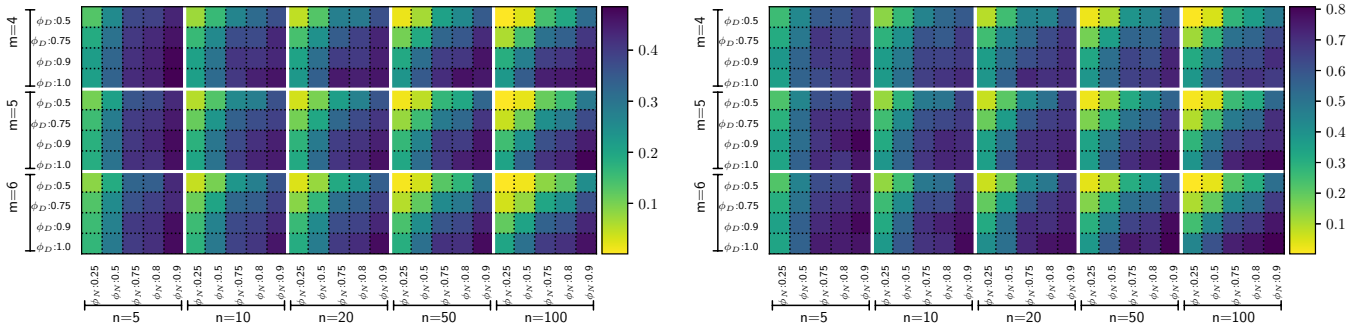
Group Size and Number of Alternatives. In addition to true preference distribution and noise models, group size and the number of alternatives might be impacting factor in determining robustness of voting rules. For all experiments, we vary the number of alternatives m over $\{4, 5, 6\}$ and the group size n over $\{5, 10, 20, 50, 100\}$.

Voting Rules. We consider four preference aggregation methods (or voting rules) in our experiments: Plurality, Borda, Copeland and Kemeny[8].

Plurality and Borda are examples of positional scoring rules where a positional scoring rule is defined by a scoring vector $\alpha = (\alpha_1, \ldots, \alpha_m)$ where $\alpha_1 \geq \ldots \geq \alpha_m$. For each voter $i \in \mathcal{N}$, an alternative a_j receives α_k points if it is ranked in the k^{th} position by i. These scoring vectors or "votes" are aggregated by summing across the scores provided by all voters for each alternative, with the final aggregated ranking corresponding to the alternatives ordered based on decreasing aggregated scores. Plurality corresponds to the scoring vector $(1, 0, \ldots, 0)$ while Borda corresponds to the scoring vector $(m-1, m-2, \ldots, 1, 0)$.

Copeland and Kemeny are examples of Condorcet consistent voting rules[8]. The Copeland rule orders alternatives based on the number of their pairwise victories, given a set of preference rankings (a tie is a half of a victory). The Kemeny rule returns

[3] One future direction is to validate these noise models in various contexts through a set of lab studies. We conjecture that noise in ranking boils down to the underlying utilities of paired items. If two items have utilities close to each other, the probability of a switch should be higher as the user is more indecisive. This can happen for items on the top or bottom. Development of such models is an interesting future direction.



(a) Avg. SKTD for Kemeny, unimodal distribution

(b) Avg. DD for Kemeny, unimodal distribution

FIGURE 2 Average SKTD (left) and DD (right) for Kemeny under our unimodal preference distribution, various m, n, ϕ_D, and ϕ_N. The x-axis changes over n and ϕ_N whereas the y-axis changes over m and ϕ_D. Best viewed in colour

the ranking that minimizes the sum of Kendall-Tau distances given a set of preference rankings $\{r_i | i = 1, \cdots, n\}$:

$$R_{Agg}^{Kemeny} = \arg\min_{r \in \Omega(\mathcal{A})} \sum_{i=1}^{n} d_\tau(r, r_i), \tag{4}$$

where $d_\tau(r, r_i)$ is Kendall's τ distance between ranking r and the individual i's preference ranking, and n is the group size.

Repetition and Statistical Analyses. For each setting (e.g., ϕ_D, ϕ_N, m, n), we generate 500 instances and report the average of SKTD and DD over those instances for all four voting rules. To mitigate the statistical noise of our experiments, we ran paired t-tests (at confidence level 0.05) for any pairwise comparison of voting rules, and all reported findings are significant with $p = 0.05$.

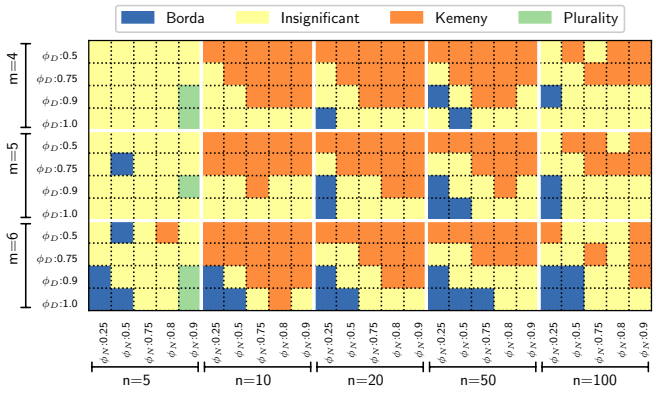
5.2 | Empirical Results

We discuss our results for analysing the robustness of our examined voting rules under various class of true preferences and noise models.

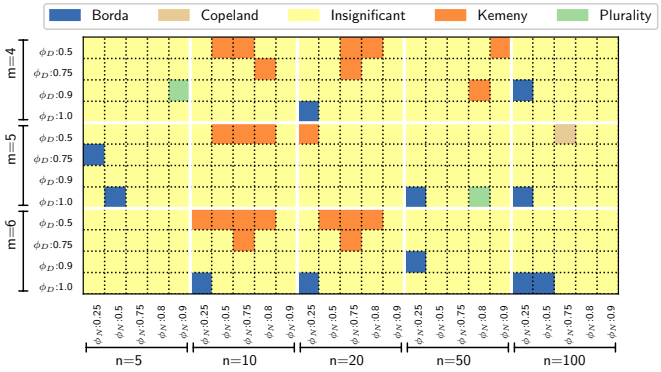
Unimodal Distributions for True Preferences. Figure 2 shows the average SKTD and DD (over 500 instances) for the Kemeny voting rule under the unimodal distribution, while varying m, n, ϕ_D, ϕ_N. The x-axis changes over n and ϕ_N whereas the y-axis changes over m and ϕ_D. We note, unsurprisingly, that average SKTD increases as the noise in the revealed preferences (i.e. controlled by ϕ_N) and preference diversity (i.e controlled by ϕ_D) increase. Also, the average SKTD decreases as either n or m increases. Our findings for Borda, Copeland and Plurality were qualitatively similar.

Figure 3 illustrates the best (minimum avg. error) and worst (maximum avg. error) voting rules of each configuration, under our unimodal distribution model, using both DD and SKTD as the judgment criteria. Figure 3(a) and Figure 3(b) show the best voting methods for each configuration under SKTD and DD respectively. We observe that, in both cases, Kemeny and Borda generally outperform other methods. For SKTD, Kemeny seems to be the best method when $n \geq 10$ and $\phi_D \leq 0.75$, or when the population size is medium to large and the preferences are similar. Borda outperformed other voting methods when $\phi_D \geq 0.9$, that is, in situations where individuals' preferences were not strongly correlated. We also observed that Plurality performed well if the population is very small, and both ϕ_D and ϕ_N are high. Using average DD as the criteria of interest, we note that for many cases it was impossible to find a clear best voting method. However, Kemeny and Borda were dominant in some settings, with Kemeny tending to do well on preference distributions that were fairly peaked (i.e. $\phi_D \leq 0.75$) and Borda doing well otherwise.

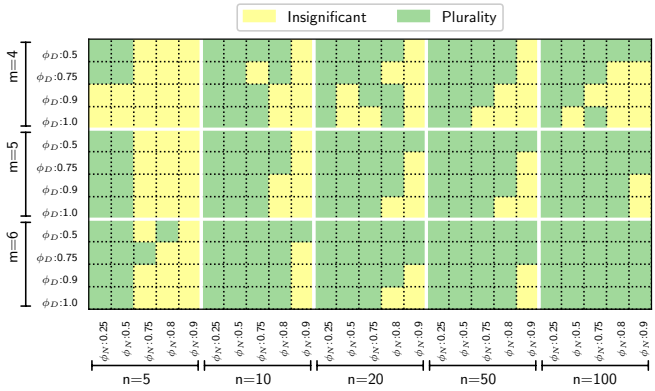
Figures 3(c) and 3(d) show the worst voting rules under SKTD and DD respectively. Plurality was the weakest rule in most settings with significant results.



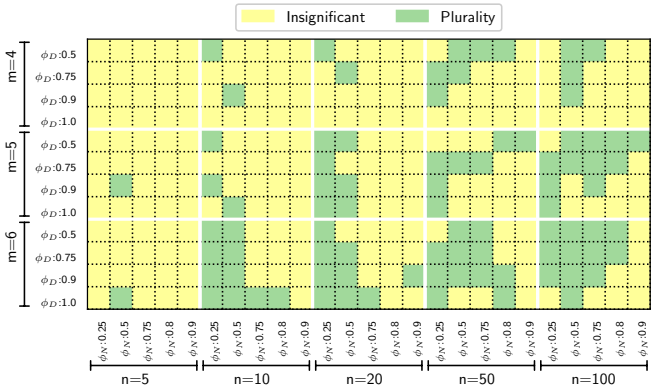
(a) Methods with min. avg. SKTD, unimodal dist.



(b) Methods with min. avg. DD, unimodal dist.



(c) Methods with max. avg. SKTD, unimodal dist.



(d) Methods with max avg. DD, unimodal dist.

FIGURE 3 The best (minimum avg. error) and worst (maximum avg. error) methods of each configuration, under our unimodal distribution, for DD and SKDT. Confidence level 0.05. Best viewed in colour.

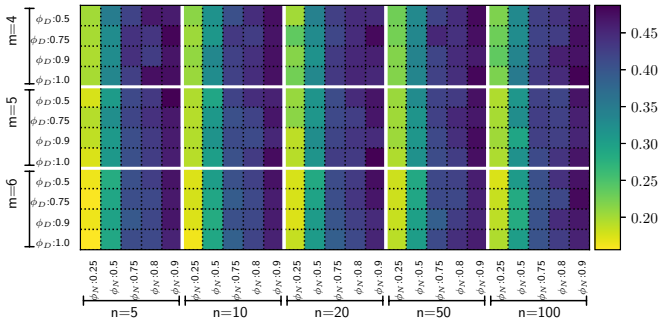
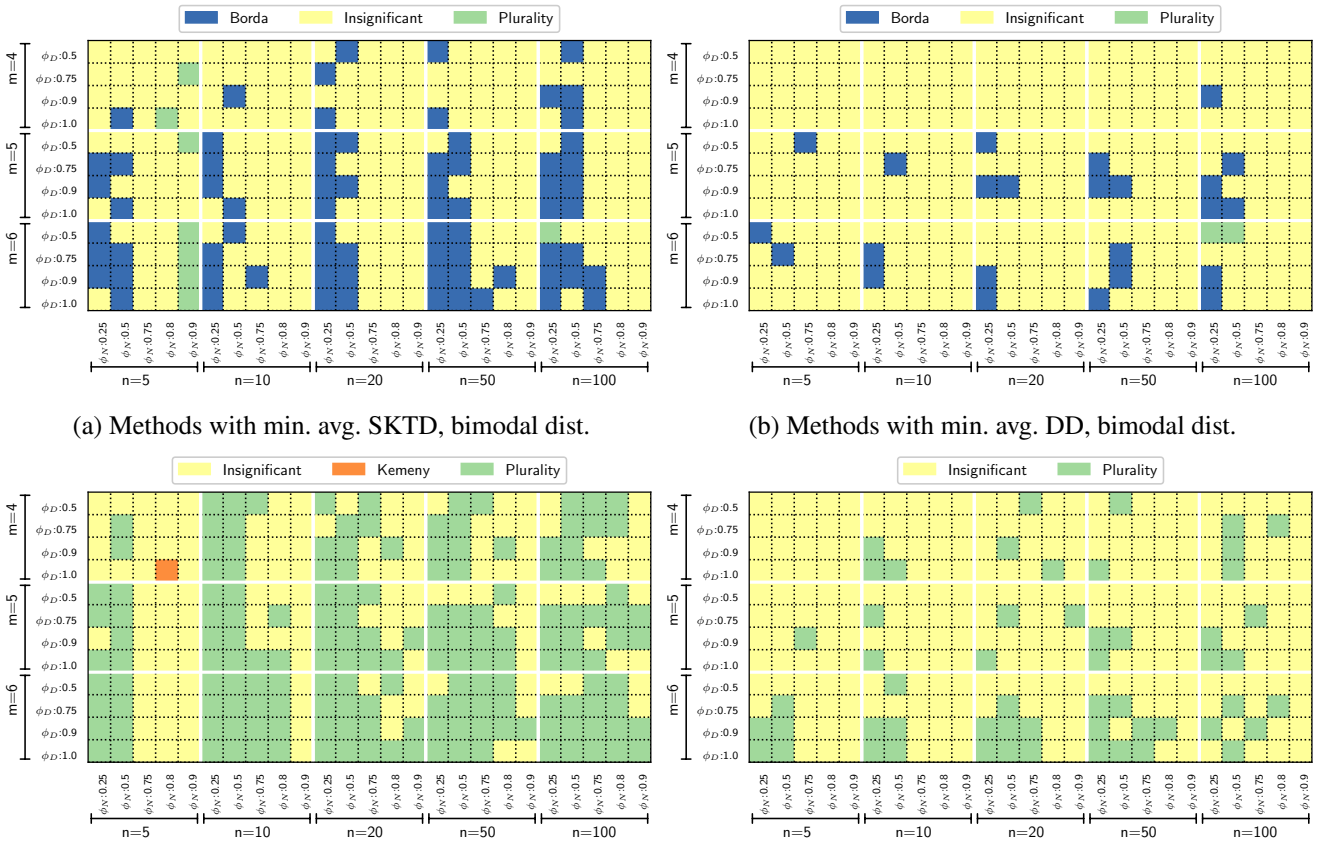


FIGURE 4 Average SKTD, Kemeny, bimodal distribution.

Bimodal Distributions for True Preferences. Figure 4 shows the average of SKTD for Kemeny under the bimodal distribution model.[4] Like with the unimodal distribution results, SKTD increases with ϕ_N. Interestingly, however, SKTD (and DD) are almost insensitive to ϕ_D and population size n, indicating that the structure of the underlying preference distribution is important.

Figure 5 shows the best (minimum avg. error) and worst (maximum avg. error) methods of each configuration, under the bimodal distribution, for SKTD and DD. The statistical test is the same as described for the unimodal setting. Figure 5(a) and

[4]Results for the other voting rules, as well as results for DD were qualitatively similar.



(a) Methods with min. avg. SKTD, bimodal dist.

(b) Methods with min. avg. DD, bimodal dist.

(c) Methods with max. avg. SKTD, bimodal dist.

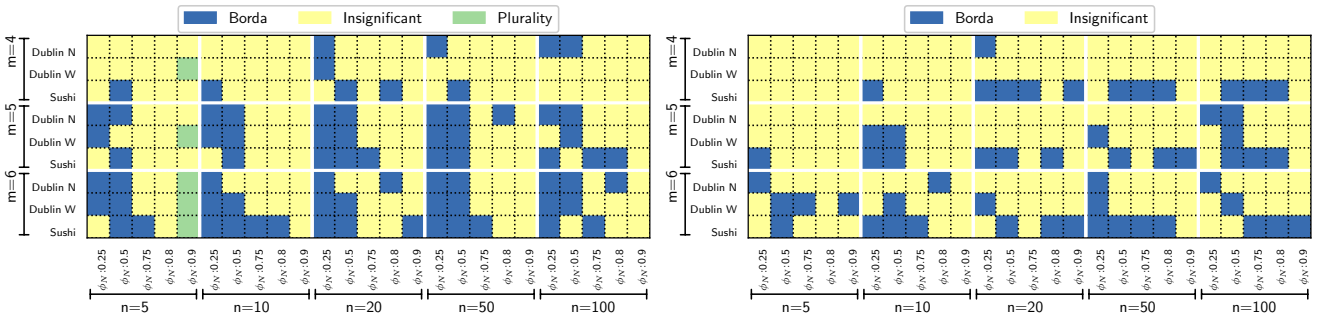
(d) Methods with max avg. DD, bimodal dist.

FIGURE 5 The best and worst methods, under bimodal distribution. Conf. level 0.05. Best viewed in colour.

Figure 5(b) show the best methods under SKTD and DD respectively. We observe that, in general, Borda outperforms other methods, unlike in the unimodal case where Kemeny was the dominant voting rule. We also note, again, that Plurality did well with small populations and high ϕ_D and ϕ_N. Figures 5(c) and 5(d) show the worst methods under SKTD and DD (resp.). Again, Plurality is the worst voting rule under either criteria for almost all settings where a worst rule could be determined. The cross-examination of Figures 3 and 5 suggests that when the dispersion of true preferences is high (i.e., large ϕ_D) and the noise dispersion is low (i.e., small ϕ_N), the Borda rule is a dominant rule with regard to both DD and SKTD regardless of the underlying distribution for true preferences.

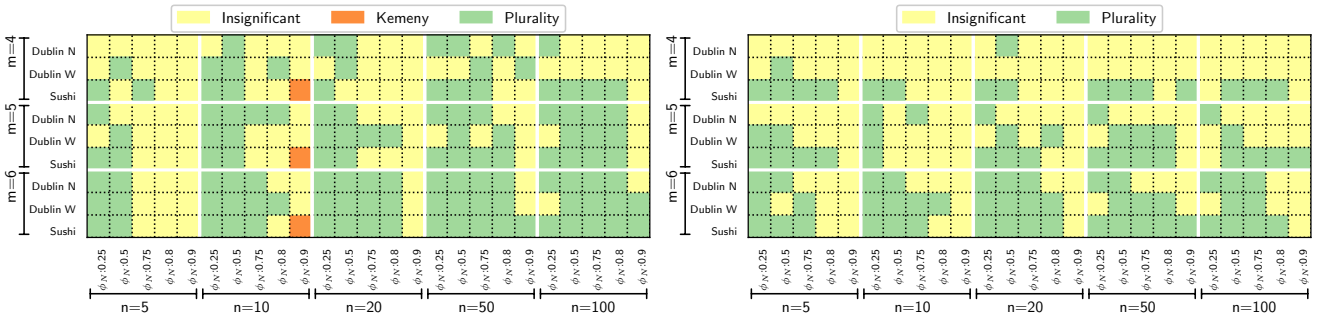
Real Preference Data for True Preferences. Figure 6 shows the best (minimum avg. error) and worst (maximum avg. error) voting methods of each configuration for various data sets and performance metrics. We observe that, in general, Borda outperforms other methods for all data sets when measured with both SKTD and DD; see Figure 6(a-b). This is similar to the bimodal distribution findings reported earlier. Plurality seems to do well again when noise is very high and group sizes are very small. Figures 6(c) and 6(d) show the worst methods under SKTD and DD (resp.). Again, Plurality is the worst for all settings (except three ones) under both SKTD and DD.

Different Noise Models. We now vary noise models while fixing $m = 6$. Figure 7 shows the best and worst methods for various data sets, noise models, group sizes, and performance metrics. For the *Entire* and *Top* noise models, Borda outperforms other methods for all data sets under both SKTD and DD; see Figure 7(a-b), while Plurality performs poorly. This is reasonable since plurality is highly sensitive to noise in the top ranked items, as captured by the *Top* noise model. For the *Bottom* noise model we observe that Plurality outperforms all other voting methods in terms of both SKTD and DD, as it is immune to noise in low-ranked items. When considering SKTD, both Kemeny and Copeland perform poorly, while for DD and small populations, Borda performs poorly. This might be explained by the observation that when the individuals have relatively diverse preferences, scores of noisy lower-ranked items are influential in the aggregated scores when the population is small.



(a) Methods with min. avg. SKTD, data sets

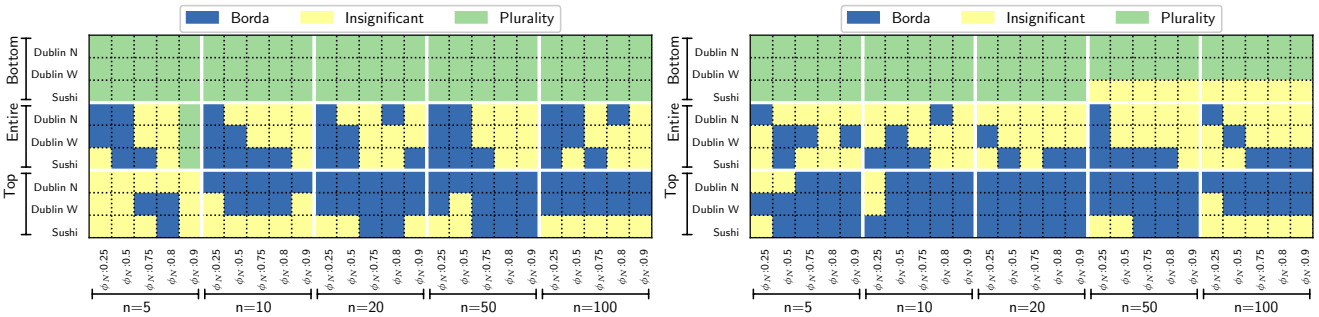
(b) Methods with min. avg. DD, data sets



(c) Methods with max. avg. SKTD, data sets

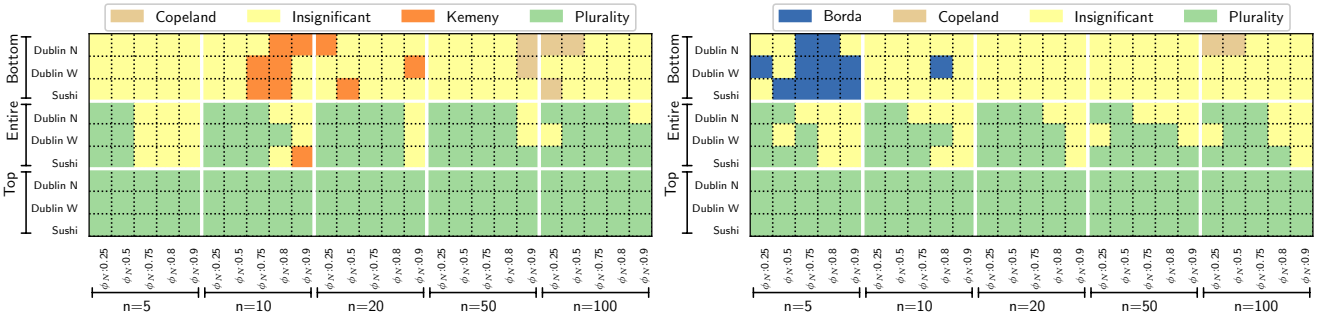
(d) Methods with max avg. DD, data sets

FIGURE 6 The best and worst methods for real-world data sets. Conf. level 0.05. Best viewed in colour.



(a) Best methods with minimum average SKTD

(b) Best methods with minimum average DD



(c) Worst methods with maximum average SKTD

(d) Worst methods with maximum average DD

FIGURE 7 The best and worst methods for various data sets, noise models. Confidence level 0.05. Best viewed in colour.

6 | CONCLUSION AND FUTURE WORK

Noise in reported subjective preferences can lead to poor outcomes once preferences are aggregated and used for joint decision making or group recommendations. In this work we propose a general probabilistic framework for modeling noise of subjective preferences along with empirical methodology for studying the robustness of aggregation methods against noise. Under our framework, we then rigorously analyze the accuracy and reliability of four well-recognized social choice methods—Plurality, Borda, Kemeny, and Copeland—for both single-option group recommendation (or voting) and aggregation tasks.

Our empirical results show that the underlying noise model influences the performance of the different group decision making methods. Borda generally performed well when noise is distributed over entire or bottom of rankings. This is consistent with empirical findings which showed that Borda tended to be a robust choice for aggregating objective rankings in human-computation domains[53,33]. However, Plurality outperforms others when the noise is more present in the low-ranked items. We argue that a contribution of this work is in highlighting the fact that the robustness of social choice methods varies depending on the underlying noise model and preference distributions. This also highlights the importance of modeling subjective noise for group decision making and recommendation.

There are many fascinating directions to explore in future work. Theoretical analyses of our probabilistic framework can shed light on how model parameters effect the reliability of any social choice method, and possibly answer what are the main characteristics of social choice methods which are more resistant to noise. Of practical importance is studying the presence and form of noise in subjective real-world preferences in various contexts (e.g., movie, food, political orientation, etc.). One can develop inference algorithms which can leverage learned true preference of an individual for more accurate prediction of other individuals' true preferences. Another interesting direction is to generalize our probabilistic framework to partial preferences.

ACKNOWLEDGMENTS

This research was supported by Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. McCarthy Kevin, Salamó Maria, Coyle Lorcan, McGinty Lorraine, Smyth Barry, Nixon Paddy. Cats: A synchronous approach to collaborative group recommendation. In: FLAIRS'06:86–91; 2006.

2. Rakesh Vineeth, Lee Wang-Chien, Reddy Chandan K. Probabilistic group recommendation model for crowdfunding domains. In: WSDM'16:257–266; 2016.

3. Crossen Andrew, Budzik Jay, Hammond Kristian J. Flytrap: Intelligent group music recommendation. In: IUI'02:184–185; 2002.

4. Pizzutilo Sebastiano, De Carolis Berardina, Cozzolongo Giovanni, Ambruoso Francesco. Group modeling in a public space: methods, techniques, experiences. In: AIC'05:175–180; 2005.

5. Yu Zhiwen, Zhou Xingshe, Hao Yanbin, Gu Jianhua. TV Program Recommendation for Multiple Viewers based on User Profile Merging. *User Modeling and User-Adapted Interaction.* 2006;16(1):63–82.

6. O'connor Mark, Cosley Dan, Konstan Joseph A, Riedl John. PolyLens: a recommender system for groups of users. In: ECSCW'01:199–218; 2001.

7. Pigozzi Gabriella, Tsoukiàs Alexis, Viappiani Paolo. Preferences in Artificial Intelligence. *Annals of Mathematics and Artificial Intelligence.* 2016;77(3–4):361–401.

8. Brandt Felix, Conitzer Vincent, Endriss Ulle, Lang Jerome, Procaccia Ariel, eds.*Handbook of Computational Social Choice*. Cambridge University Press; 2016.

9. Felfernig Alexander, Boratto Ludovico, Stettinger Martin, Tkalčič Marko. *Group Recommender Systems: An introduction*. Springer; 2018.

10. Binmore Ken. *Natural Justice*. Oxford University Press; 2005.

11. Hill Will, Stead Larry, Rosenstein Mark, Furnas George. Recommending and evaluating choices in a virtual community of use. In: CHI'95:194-201; 1995; Denver, CO.

12. Cosley Dan, Lam Shyong, Albert Istvan, Konstan Jospeh, Riedl John. Is seeing believing?: How recommender system interfaces affect users' opinions. In: CHI'03:585-592; 2003; Fort Lauderdale, USA.

13. Amatriain Xavier, Pujol Josep M., Oliver Nuria. I Like It... I Like It Not: Evaluating User Ratings Noise in Recommender Systems. In: UMAP'09:247–258; 2009; Trento, Italy.

14. McCarthy Joseph F, Anagnost Theodore D. MusicFX: an Arbiter of Group Preferences for Computer Supported Collaborative Workouts. In: CSCW'98:363–372; 1998.

15. Berkovsky Shlomo, Freyne Jill. Group-based Recipe Recommendations: Analysis of Data Aggregation Strategies. In: RecSys'10:111–118; 2010.

16. Masthoff Judith. Group Modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers. *User Modeling and User-Adapted Interaction.* 2004;14(1):37–85.

17. Baltrunas Linas, Makcinskas Tadas, Ricci Francesco. Group Recommendations with Rank Aggregation and Collaborative Filtering. In: RecSys'10:119–126; 2010.

18. Amer-Yahia Sihem, Roy Senjuti Basu, Chawlat Ashish, Das Gautam, Yu Cong. Group Recommendation: Semantics and Efficiency. *Proc. VLDB Endow.*. 2009;2(1):754–765.

19. Seko Shunichi, Yagi Takashi, Motegi Manabu, Muto Shinyo. Group Recommendation Using Feature Space Representing Behavioral Tendency and Power Balance Among Members. In: RecSys'11:101–108; 2011.

20. Gartrell Mike, Xing Xinyu, Lv Qin, et al. Enhancing Group Recommendation by Incorporating Social Relationship Interactions. In: GROUP'10:97–106; 2010.

21. Salehi-Abari Amirali, Boutilier Craig. Preference-oriented Social Networks: Group Recommendation and Inference. In: RecSys'15; 2015.

22. Xiao Lin, Min Zhang, Yongfeng Zhang, Zhaoquan Gu, Yiqun Liu, Shaoping Ma. Fairness-Aware Group Recommendation with Pareto-Efficiency. In: RecSys'17:107–115; 2017.

23. Salehi-Abari Amirali, Boutilier Craig, Larson Kate. Empathetic decision making in social networks. *Artificial Intelligence.* 2019;275:174–203.

24. Procaccia Ariel D., Rosenschein Jeffrey S., Kaminka Gal A.. On the Robustness of Preference Aggregation in Noisy Environments. In: AAMAS'07:661–667; 2007.

25. Shiryaev Dmitry, Yu Lan, Elkind Edith. On Elections with Robust Winners. In: AAMAS'13:415–422; 2013.

26. Procaccia Ariel D., Shah Nisarg. Optimal Aggregation of Uncertain Preferences. In: AAAI'16:608–614; 2016.

27. Lu Tyler, Boutilier Craig. Robust Approximation and Incremental Elicitation in Voting Protocols. In: IJCAI'11:287–293; 2011; Barcelona.

28. Lu Tyler, Boutilier Craig. Multi-winner Social Choice with Incomplete Preferences. In: IJCAI'13; 2013; Beijing.

29. Young H Peyton. Condorcet's theory of voting. *The American Political Science Review.* 1988;:1231–1244.

30. Conitzer Vincent, Sandholm Tuomas. Common Voting Rules As Maximum Likelihood Estimators. In: UAI'05:145–152; 2005.

31. Xia Lirong, Conitzer Vincent. A Maximum Likelihood Approach Towards Aggregating Partial Orders. In: IJCAI'11:446–451; 2011.

32. Caragiannis Ioannis, Procaccia Ariel D., Shah Nisarg. When Do Noisy Votes Reveal the Truth?. In: EC '13:143–160; 2013.

33. Caragiannis Ioannis, Chatzigeorgiou Xenophon, Krimpas George A., Voudouris Alexandros A.. Optimizing Positional Scoring Rules for Rank Aggregation. In: :430–436; 2017.

34. Caragiannis Ioannis, Micha Evi. Learning a Ground Truth Ranking Using Noisy Approval Votes. In: :149–155; 2017.

35. Weerdt Mathijs M., Gerding Enrico H., Stein Sebastian. Minimising the Rank Aggregation Error: (Extended Abstract). In: AAMAS'16:1375–1376; 2016.

36. Young Peyton. Optimal voting rules. *The Journal of Economic Perspectives.* 1995;9(1):51–64.

37. Xia Lirong, Conitzer Vincent, Lang Jérôme. Aggregating Preferences in Multi-issue Domains by Using Maximum Likelihood Estimators. In: AAMAS'10:399–408; 2010.

38. Procaccia Ariel D., Reddi Sashank J., Shah Nisarg. A Maximum Likelihood Approach for Selecting Sets of Alternatives. In: UAI'12:695–704; 2012.

39. Procaccia Ariel D, Shah Nisarg, Zick Yair. Voting rules as error-correcting codes. *Artificial Intelligence.* 2016;231:1–16.

40. Gormley Isobel Claire, Murphy Thomas Brendan. Exploring Voting Blocs Within the Irish Electorate. *Journal of the American Statistical Association.* 2008;103(483):1014–1027.

41. Busse Ludwig M., Orbanz Peter, Buhmann Joachim M.. Cluster Analysis of Heterogeneous Rank Data. In: ICML'07:113–120; 2007.

42. Lu Tyler, Boutilier Craig. Learning Mallows Models with Pairwise Preferences. In: ICML'12:145–152; 2011.

43. Azari Hossein, Parkes David, Xia Lirong. Random Utility Theory for Social Choice. In: NIPS'12:126–134; 2012.

44. Benter W.F.. In: Hausch Donald B., Lo Victor S.Y., Ziemba William T., eds. *Efficiency of Race Track Betting Markets*, Academic Press 1994.

45. Gormley Isobel Claire, Murphy Thomas Brendan. A Latent Space Model for Rank Data. In: Airoldi Edoardo M., Blei David M., Fienberg Stephen E., Goldenberg Anna, Xing Eric P., Zheng Alice X., eds. *Statistical Network Analysis: Models, Issues, and New Directions*, Springer 2007 (pp. 90–102).

46. Murphy Thomas Brendan, Martin Donal. Mixtures of Distance-based Models for Ranking Data. *Computational Statistics & Data Analysis.* 2003;41(3):645–655.

47. Lee Paul H., Yu Philip L.H.. Mixtures of Weighted Distance-based Models for Ranking Data with Applications in Political Studies. *Computational Statistics & Data Analysis.* 2012;56(8):2486 – 2500.

48. Salehi-Abari Amirali, Boutilier Craig. Ranking Networks. In: FNAMMA'13; 2013.

49. Marden John I.. *Analyzing and Modeling Rank Data*. London: Chapman and Hall; 1995.

50. Fligner Michael A, Verducci Joseph S. Distance Based Ranking Models. *Journal of the Royal Statistical Society. Series B (Methodological).* 1986;:359–369.

51. Officer Dublin County Returning. Irish Election Datasets http://www.dublincountyreturningofficer.comRetrieved: January 2012; 2002.

52. Kamishima Toshihiro, Kazawa Hideto, Akaho Shotaro. Supervised Ordering — An Empirical Survey. In: ICDM'05:673–676; 2005.

53. Mao Andrew, Procaccia Ariel, Chen Yiling. Better Human Computation Through Principled Voting. In: AAAI'13:1142–1148; 2013; Bellevue, WA, USA.