Journal of the Operational Research Society

ISSN: 0160-5682 (Print) 1476-9360 (Online) Journal homepage: https://www.tandfonline.com/loi/tjor20

Graph-based rank aggregation method for high-dimensional and partial rankings

Yu Xiao, Hong-Zhong Deng, Xin Lu & Jun Wu

To cite this article: Yu Xiao, Hong-Zhong Deng, Xin Lu & Jun Wu (2021) Graph-based rank aggregation method for high-dimensional and partial rankings, Journal of the Operational Research Society, 72:1, 227-236, DOI: 10.1080/01605682.2019.1657365

To link to this article: https://doi.org/10.1080/01605682.2019.1657365

Published online: 12 Dec 2019.

JOURNAL OF THE OPERATIONAL RESEARCH SOCIETY
2021, VOL. 72, NO. 1, 227–236
https://doi.org/10.1080/01605682.2019.1657365

ORIGINAL ARTICLE

Graph-based rank aggregation method for high-dimensional and partial rankings

Yu Xiao[a], Hong-Zhong Deng[a], Xin Lu[a,b] and Jun Wu[a,c]

[a]College of Systems Engineering, National University of Defense Technology, Changsha, P. R. China; [b]School of Business, Central South University, Changsha, P. R. China; [c]International Academic Center of Complex Systems, Beijing Normal University, Zhuhai, P. R. China

ABSTRACT

Rank aggregation has recently become a common approach for combining individual rankings into a consensus and for quantifying and improving performance in various applications, such as elections, web page rankings, and sports. During the past few years, rankings from many sources have become increasingly high-dimensional and partial. In this study, we develop a rank aggregation method by constructing a directed weighted competition graph. We introduce the concept of "ratio of out- and in-degrees (ROID)" to transform high-dimensional partial rankings into a single consensus. Moreover, we provide a novel effectiveness measure for the aggregate ranking according to its deviations from the ground truth ranking. The proposed method is compared with four typical methods with synthetic rankings. The results indicate that our method outperforms the other four by a significant margin and can be particularly efficient in aggregating high-dimensional rankings. The empirical results validate the effectiveness and feasibility of our method.

ARTICLE HISTORY
Received 10 July 2018
Accepted 9 August 2019
Published online 10 December 2019

KEYWORDS
Decision analysis; group ranking; rank aggregation; competition graph; ratio of out- and in-degrees (ROID)

1. Introduction

As an obligatory task of various selection and evaluation boards, the problem of ranking has appeared under many guises, such as world university ranking (Thieme, Prior, Tortosa-Ausina, & Gempp, 2016), web page ranking (Page, Brin, Motwani, & Winograd, 1999), and sports ranking (Filippo, 2011). The ranking problem has been studied extensively in the past few decades (Langville & Meyer, 2012). If there is only a single criterion for ranking, the task is relatively easy. However, in many situations, one must consolidate a consensus ranking of alternatives, given the individual ranking preferences regarding several different criteria (Ahn, 2017; Dwork, Kumar, Naor, & Sivakumar, 2001; Read, Edwards, & Gear, 2000; Wang, Chin, & Yang, 2007). Aggregating individual rankings into a consensus is the essence of the group-ranking problem in social choice theory (Cook & Kress, 1990), where a group of voters rank the available alternatives. A voting rule is then applied to identify the best alternative or to aggregate each rank into a consensus regarding the alternatives (Aledo, Gámez, & Molina, 2016; Baucells & Sarin, 2003). This task is known as the "rank aggregation problem" (Aledo, Gámez, & Alejandro, 2017, 2018; Cook & Kress, 1990). This problem has penetrated many areas of decision-making and evaluation, such as meta-search engines (Dwork et al., 2001) and disease-related genes selection (Kolde, Laur, Adler, & Vilo, 2012). Many rank aggregation methods, including the Borda's method (Borda, 1781; Langville & Meyer, 2012), the Dowdall method (Reilly, 2002), the minimum violations ranking method (Ali, Cook, & Kress, 1986; Chartier, Kreutzer, Langville, Pedings, & Yamamoto, 2010; Park, 2005; Pedings, Langville, & Yamamoto, 2012), the FAST method (Amodio, D'Ambrosio, & Siciliano, 2016), the footrule method (Dwork et al., 2001) and the Markov chain method (Dwork et al., 2001), have been proposed over the past few centuries.

Rank aggregation methods can be classified into two categories: heuristic methods and optimization methods (Argentini & Blanzieri, 2012). Many classic methods belong to the heuristic group. They aim to assign an index to each alternative that can be sorted to define a consensus ranking. Alternatively, optimization methods aim to find a consensus ranking minimizing the distance to or violations of the input rankings using a specified ranking distance or violation measure, such as the Kendall tau distance or Spearman footrule distance. Rankings from many multiple-criteria decision-making problems arising from the rapid development of information technology have become increasingly high-dimensional and partial, as a large number of rankings are given by voters, but each voter can only rank a subset of the

CONTACT Jun Wu wujunpla@hotmail.com College of Systems Engineering, National University of Defense Technology, Changsha, Hunan, 410073, P. R. China

universal set of alternatives. This kind of input is typical in the ranking of commodities, movies or brands. It is also important to note that these rankings may be inaccurate due to the limited judgment ability of voters. Aggregating such high-dimensional, inaccurate and partial rankings presents immense challenges related to both effectiveness and efficiency. This article introduces a graph-based rank aggregation method and a new index called ROID for this problem. The theoretical underpinnings of stating new criteria in effective rank aggregation methods are provided, and our method is compared with several typical methods using a newly proposed experimental data generation method (Xiao, Deng, Wu, Deng, & Lu, 2017). The experimental results indicate that the proposed method is successful.

The remainder of this article is organized as follows. We first introduce four typical rank aggregation methods and describe the experimental data generation method in Section 2. The newly proposed method is introduced in Section 3 and compared with four typical methods in Section 4. In Section 5, the new method is applied to an empirical dataset. A summary of the contributions of this article and a discussion of future work are given in Section 6.

2. Preliminaries

2.1. Borda's method

As one of the most typical rank aggregation methods (Borda, 1781), the Borda's method has been widely used in the past few centuries (Aledo et al., 2016). Given k rankings $R_1, R_2, \ldots, R_k$, for each alternative $a \in R_i$, the alternative a is first assigned a score $B_i(a)$ equal to the number of alternatives that a outranks in ranking R_i. Next, the Borda count $B(a)$ of alternative a is calculated as $\sum_{i=1}^{k} B_i(a)$. The alternatives are then sorted in the descending order based on their Borda counts to create a consensus ranking.

2.2. Dowdall method

As a "modified" form of the Borda's method, the Dowdall method has been widely used in political elections in many countries (Reilly, 2002). Given k rankings $R_1, R_2, \ldots, R_k$, for each alternative $a \in R_i$, alternative a is first assigned a score $D_i(a)$ equal to the reciprocal of its rank in ranking R_i. Next, the total score $D(a)$ of alternative a is calculated as $\sum_{i=1}^{k} D_i(a)$. The alternatives are then sorted in the descending order based on their total scores to create a consensus ranking.

2.3. Minimum violations ranking method

As its name suggests, the minimum violations ranking method searches specifically for consensus rankings with the minimum violations (Park, 2005). Typically, the binary integer linear program (BILP) formulation of the MVR problem is a preferred way of finding the optimal consensus ranking (Chartier et al., 2010; Langville & Meyer, 2012; Pedings et al., 2012). Denote by x_{ij} the decision variables that determine whether alternative a_i should be ranked above alternative a_j. Specifically,

$$x_{ij} = \begin{cases} 1, & \text{if } a_i \text{ is ranked above } a_j \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Several constraints must be added to force matrix X to have the properties that meet the basic needs for producing a unique ranking of the n alternatives:

$$x_{ij} + x_{ji} = 1; x_{ij} + x_{jk} + x_{ki} \leq 2 \tag{2}$$

Given k rankings $R_1, R_2, \ldots, R_k$ of the n alternatives, we define the following ranking scores for any pair of objects (Langville & Meyer, 2012):

$$\begin{aligned} c_{ij} = &(\#\, of \text{ rankings with } a_i \text{ above } a_j) \\ &-(\#\, of \text{ rankings with } a_i \text{ below } a_j). \end{aligned} \tag{3}$$

The objective of MVR is to find the consensus ranking, maximizing the conformity among input rankings. In terms of ranking scores c_{ij} and variables x_{ij}, this maximization problem becomes:

$$\max \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij}. \tag{4}$$

BILPs are typically solved with a technique called "branch and bound", that uses a series of linear programming (LP) relaxations of the problem to form a tree to narrow down the process of stepping through the discrete solution space (Langville & Meyer, 2012). When the branch and bound procedure terminates with an optimal solution X^*, we can obtain a MVR consensus ranking by sorting the column sums of X^* in the ascending order (Langville & Meyer, 2012).

2.4. Borda(M^r) method

Aledo et al. (Aledo et al., 2016) proposed using the concept of extension sets to manage the unobserved information, i.e., to deal with the uncertainty associated with the items not appearing in a given ranking by considering the positions of the ranking in which they could be placed.

Given a partial ranking π and a permutation σ, σ is consistent with π if for all alternatives a_i, a_j ranked in π, one of the two following conditions holds:

(1) a_i and a_j share same rank in π,
(2) if a_i outranks a_j in π, then a_i must outrank a_j in σ, and vice versa.

Then, the extension set of ranking π is

$$E(\pi) = \{\sigma | \sigma \text{ is consistent with } \pi\}.$$

Note that a non-ranked alternative of π can be ranked in σ between two alternatives that share the same rank in π. Given this possibility, the researchers proposed the concept of restricted consistent permutations.

σ is restricted consistently with π if for all alternatives a_i, a_j ranked in π, the two following conditions hold:

(1) σ is consistent with π,
(2) $\forall a_t$ that is not ranked in π cannot be ranked between a_i and a_j that share same rank in π.

The researchers call $E^r(\pi)$ the restricted extension set of partial ranking π,

$$E^r(\pi) = \{\sigma | \sigma \text{ is restricted consistent with } \pi\}.$$

To accomplish rank aggregation with extension sets and restricted extension sets, the researchers first introduce the concept of precedence extension value.

Given a ranking π and two alternatives a_i and a_j, the precedence extension value $V_{ij}(\pi)$ of π between a_i and a_j is defined by

$$V_{ij}(\pi) = \frac{1}{|E(\pi)|} \sum_{\sigma \in E(\pi)} 1(a_i \text{ outranks } a_j) \tag{5}$$

From the definition above, it follows immediately that $V_{ij}(\pi) = 1 - V_{ji}(\pi)$.

Given N rankings $\pi_1, \pi_2, \ldots \pi_k, \ldots \pi_N$, the researchers then define the precedence extension matrix $M = [M_{ij}]_{i,j=1:n}$ by

$$\begin{aligned} M_{ij} &= \frac{1}{N} \sum_{k=1}^{N} V_{ij}(\pi_k) \\ &= \frac{1}{N} \sum_{k=1}^{N} \frac{1}{|E(\pi_k)|} \sum_{\sigma \in E(\pi_k)} 1(a_i \text{ outranks } a_j) \end{aligned} \tag{6}$$

Similarly, the restricted precedence extension values $V^r_{ij}(\pi)$ of π between a_i and a_j are defined by

$$V^r_{ij}(\pi) = \frac{1}{|E^r(\pi)|} \sum_{\sigma \in E^r(\pi)} 1(a_i \text{ outranks } a_j) \tag{7}$$

The restricted precedence extension matrix $M^r = [M^r_{ij}]_{i,j=1:n}$ is defined by

$$\begin{aligned} M^r_{ij} &= \frac{1}{N} \sum_{k=1}^{N} V^r_{ij}(\pi_k) \\ &= \frac{1}{N} \sum_{k=1}^{N} \frac{1}{|E^r(\pi_k)|} \sum_{\sigma \in E^r(\pi_k)} 1(a_i \text{ outranks } a_j) \end{aligned} \tag{8}$$

Using matrices M and M^r computed in Equations 6 and 8, the researchers apply Borda(M) and Borda(M^r) methods to obtain the consensus rankings by sorting the column sums of M and M^r in the ascending order. It should be noted that we use the aggregated ranking obtained from the restricted precedence extension matrix M^r to perform experiments and name it BM(M^r) in this article.

2.5. Experimental data generation method

Most previous studies evaluated and compared rank aggregation methods using real-world datasets that were limited not only because they were typically hard to obtain but also because they lacked ground truth ranking of alternatives to evaluate the effectiveness of rank aggregation, and used the totalized Kendall tau distance between the aggregated ranking and all the input rankings, $\sum_{i=1}^{N} K(\hat{R}, R_i)$, as the measure to evaluate the effectiveness. The essence of this traditional measure is to characterize the centrality of the aggregated ranking with respect to individual rankings, whereas the centrality is not equivalent to correctness. Therefore, a proper benchmark synthetic data generation method is needed to evaluate rank aggregation methods in rank aggregation research. Note that we use the experimental data generation method developed by Xiao et al. (Xiao et al., 2017) that can provide ground truth ranking of alternatives to perform experiments.

Consider a rank aggregation problem with M voters and N alternatives. The researchers first assume that there exists a ground truth ranking of the alternatives, which can be the latent ranking of the actual strengths of each alternative that individual voters and by extension, the rank aggregation itself are attempting to estimate given the displayed abilities of those alternatives. To acquire it, the researchers denote the inherent ability of alternative a_j by ϕ_j. It may be a certain attribute of a_j, such as the height of a person, or the quality of a product. The researchers assume that ϕ_j follows a uniform distribution in the region $[0, 1]$. Then, the ground truth rank r_j of a_j is acquired based on ϕ_j, and the ground truth ranking of alternatives is denoted by $R_0 = [r_1, r_2, ..., r_M]$. Intuitively, a larger inherent ability of an alternative corresponds to a higher rank. The researchers introduce $\tilde{\phi}_{ij}$, the displayed inherent ability of alternative a_j for voter b_i, and assume that voters rank alternatives based on it because voters may not be perfectly aware of ϕ_j in practice. Denote by $R_i = [\tilde{r}_{i1}, \tilde{r}_{i2}, ..., \tilde{r}_{iM}]$ the ranking given by voter b_i, where $\tilde{r}_{ij}$ is the rank of alternative a_j. As shown in Figure 1, $\tilde{\phi}_{ij}$ is a random variable following a uniform distribution in the region $[\phi_j - \phi_j(1 - \beta_{ij}), \phi_j + (1 - \phi_j)(1 - \beta_{ij})]$. Variable $\beta_{ij} \in [0, 1]$ represents the accuracy of the displayed inherent ability

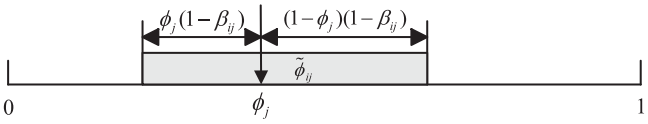

Figure 1. Displayed inherent ability of a_j for voter b_i.

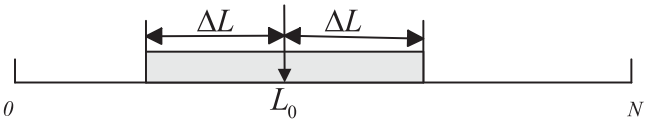

Figure 2. Length of an individual ranking R_i.

of alternative a_j for voter b_i; note that a larger β_{ij} results in a narrower distribution region and a more accurate displayed inherent ability $\tilde{\phi}_{ij}$. Similarly to the authors, in this article, we assume that the displayed accuracy β_{ij} is the same for all alternatives and voters, i.e., $\beta_{ij} = \beta$ for all $i \in [1, M]$ and $j \in [1, N]$.

In R_i, if voter b_i does not rank $a_j, \tilde{r}_{ij} = 0$. Then, the length of ranking R_i is $L_i = |\{\tilde{r}_{ij} | \tilde{r}_{ij} > 0, 1 \le j \le M\}|$, and $0 \le L_i \le M$. L_i is a random variable following a uniform distribution in the region $[L_0 - \Delta L, L_0 + \Delta L]$, as shown in Figure 2. Parameter $L_0 \in [0, M]$ represents the baseline length of individual rankings, and $0 \le \Delta L < L_0$ represents the variation in the individual ranking lengths.

Using this experimental data generation method, the researchers propose measuring the effectiveness of rank aggregation methods by the Kendall tau distance D between the aggregated ranking $\hat{R}$ and the ground truth ranking R_0 that counts the number of pairwise disagreements between two rankings and can be written as follows:

$$D = K(\hat{R}, R_0) = |\{(a_i, a_j) | i < j, \hat{R}(a_i) < \hat{R}(a_j), \; but \; R_0(a_i) > R_0(a_j)\}|. \tag{9}$$

Note that $\hat{R}(a)$ and $R_0(a)$ are the ranks of the alternatives. Intuitively, the smaller the value of D is, the more effective the rank aggregation method. Instead of centrality of the aggregated ranking, this measure D characterizes the correctness of the aggregated ranking.

3. Graph-based rank aggregation method

3.1. Competition graph of alternatives

Consider M rankings of N alternatives given by M voters. The ranking matrix is denoted by $R = (r_{ij})_{M \times N}$, where r_{ij} represents the rank of alternative a_j given by voter b_i. Note that voters may only rank a small number of alternatives; $r_{ij} = 0$ if voter b_i did not rank alternative a_j. A simple example of alternative rankings given by voters is shown in Table 1 for 8 voters and 5 alternatives. The corresponding ranking matrix is as follows:

$$R = \begin{bmatrix} 5 & 4 & 3 & 2 & 1 \\ 0 & 4 & 2 & 3 & 1 \\ 3 & 4 & 1 & 2 & 0 \\ 0 & 3 & 2 & 0 & 1 \\ 4 & 3 & 2 & 1 & 0 \\ 4 & 3 & 0 & 2 & 1 \\ 3 & 4 & 2 & 0 & 1 \\ 4 & 3 & 0 & 1 & 2 \end{bmatrix}. \tag{10}$$

The transition matrix for voter b_i is denoted by $P^i = (p^i_{st})_{N \times N}$, where $p^i_{st} = 1$ if alternative a_s outranks alternative a_t according to voter b_i $(0 < r_{is} < r_{it})$; otherwise, $p^i_{st} = 0$. If there is a tie, i.e., $r_{is} = r_{it}$, we assume that $p^i_{st} = p^i_{ts} = 1$. As an example, the transition matrix for voter b_8 shown in Table 1 is

$$P^8 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{11}$$

Based on the transition matrix, we define the competition matrix as $A = (a_{st})_{N \times N}$, where $a_{st} = \sum_i^M p^i_{st}$. In the example shown in Table 1, the competition matrix A is

$$A = \begin{bmatrix} 0 & 2 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 \\ 4 & 6 & 0 & 2 & 0 \\ 5 & 6 & 2 & 0 & 1 \\ 4 & 6 & 4 & 3 & 0 \end{bmatrix} \tag{12}$$

Based on the competition matrix, the competition graph is defined by $G = (V, E)$, where V is the set of nodes representing alternatives, and $E \subseteq V \times V$ is the set of edges representing the win-loss records among the alternatives. If $a_{st} > 0$, then there is a directed edge e_{st} with the weight a_{st} between alternative a_s and alternative a_t. The weight of the directed edge e_{st} represents the number of times alternative a_s is ranked ahead of alternative a_t across the

Table 1. Example of alternative rankings given by voters, where $N = 5$ and $M = 8$.

voter	ranking
b_1	$[a_5 \ a_4 \ a_3 \ a_2 \ a_1]$
b_2	$[a_5 \ a_3 \ a_4 \ a_2]$
b_3	$[a_3 \ a_4 \ a_1 \ a_2]$
b_4	$[a_5 \ a_3 \ a_2]$
b_5	$[a_4 \ a_3 \ a_2 \ a_1]$
b_6	$[a_5 \ a_4 \ a_2 \ a_1]$
b_7	$[a_5 \ a_3 \ a_1 \ a_2]$
b_8	$[a_4 \ a_5 \ a_2 \ a_1]$

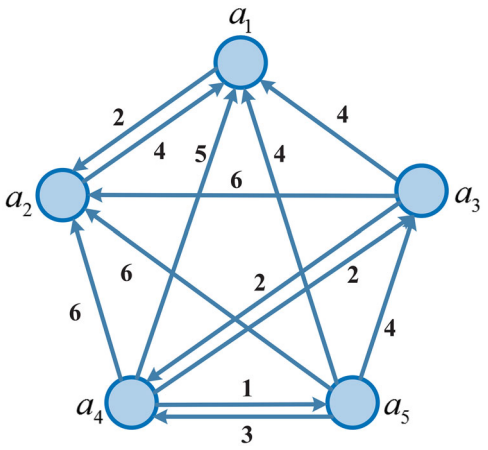

Figure 3. Competition graph for the example shown in Table 1.

spectrum of all rankings being aggregated. Figure 3 shows the competition graph for the above example.

3.2. Aggregated alternative rankings

To obtain the final alternative rankings by aggregating the rankings given by voters, let d_j^+ in the competition graph G be the out-degree of node v_j (the sum of weights of all edges departing from node v_j), and let d_j^- be the in-degree of node v_j (the sum of weights of all edges leading to node v_j). The out- and in-degrees of node v_j can thus be calculated as follows:

$$d_j^+ = \sum_{t=1}^{N} a_{jt}, \tag{13}$$

and

$$d_j^- = \sum_{s=1}^{N} a_{sj}. \tag{14}$$

Following intuition, we can define the ratio of out- and in-degrees (ROID)

$$\eta_j = d_j^+ / d_j^- \tag{15}$$

to quantify the strength of alternative a_j. An alternative with a larger d_j^+ and a smaller d_j^- might have a higher rank. However, in the two extreme cases $d_j^+ = 0$ or $d_j^- = 0$, the above approach may be problematic. To overcome this, we consider adding

Table 2. Aggregated rank of five alternatives based on the competition graph.

alternative	d_j^+	d_j^-	η_j	rank
a_1	2	17	0.17	5
a_2	4	20	0.24	4
a_3	12	6	1.86	3
a_4	14	5	2.50	2
a_5	17	1	9.00	1

to the competition graph a virtual "super" node connected to each node by two directed edges leading to and from it. That is, an extra "super" alternative is added that is superior to the other alternatives once and, at the same time, was inferior to the other alternatives once. Then, we modify the formula for the "ratio of out- and in-degree" as follows:

$$\eta_j = \frac{d_j^+ + 1}{d_j^- + 1} \tag{16}$$

Using the ratio of out- and in-degrees (ROID) as the criterion, we sort all alternatives and denote the aggregated ranking by $\hat{R}$. The ROIDs of 5 alternatives in the above example are shown in Table 2, and the aggregated ranking is $\hat{R} = [a_5, a_4, a_3, a_2, a_1]$. We call this method the "competition graph (CG) method".

3.3. Desirable properties

Most of the rank aggregation methods are inspired by results obtained in social choice theorywhich studies how to aggregate the individual preferences to reach a collective consensus. In social choice theory, there are several desirable properties and requirements for the voting systems (Arrow, 1950, 1952; Fishburn, 1977; Moritz, Reich, Schwarz, Bernt, & Middendorf, 2015), i.e. universality, monotonicity, transitivity, independence of irrelevant alternatives, non-imposition and non-dictatorship. It is worthy discussing those properties and requirements for the rank aggregation even though it has been proven that no system can have all these properties.

In our method, universality is agreed because voters are free to rank alternatives. Monotonicity imposes that if an alternative a_i rises or does not fall in the ranking of each voter without any other change in those rankings and if a_i was preferred to another alternative a_j before the changes in individual rankings, then a_i is still preferred to a_j. In our method, this corresponds that the winning record of a_i at least once more than it of a_j while the losing record of a_i at least once less than it of a_j in each individual ranking, which means that the alternative a_i receives a bigger d_i^+ but a smaller d_i^- compared with a_j ($d_i^+ \geq d_j^+ + M$ and $d_i^- \leq d_j^- - M$, where M is the number of the input rankings). Consequently,

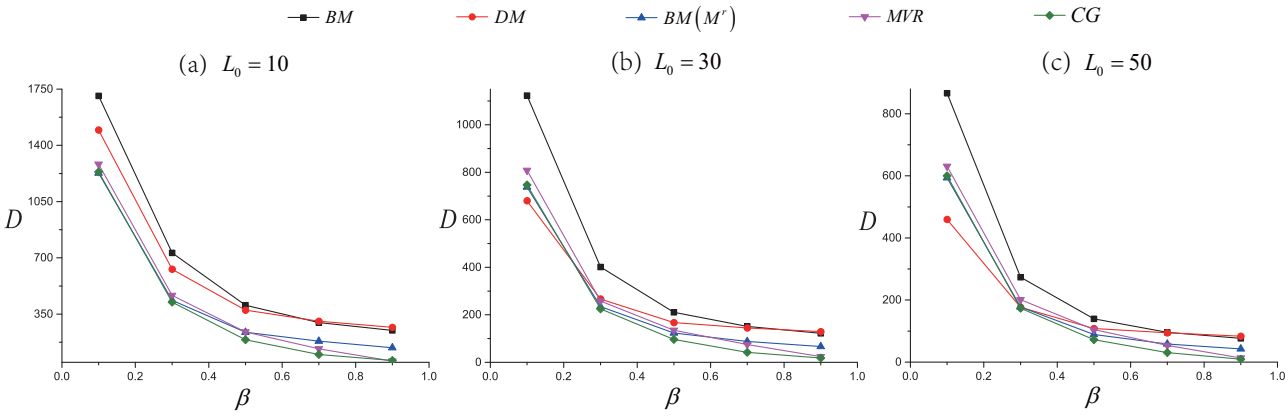

Figure 4. Effectiveness measure D vs. ranking accuracy β for various L_0, where $N = 100$, $M = 1000$ and $\Delta L = 0.3L_0$. The results were averaged over 100 independent trials.

the $\eta_i = \frac{d_i^+ + 1}{d_i^- + 1}$ will become bigger than it of a_j, which means that the alternative a_i will be ranked above a_j in the final consensus ranking. In this sense, this property is agreed with our method.

Transitivity states that if an alternative a_i is preferred to another alternative a_j which is preferred to a third alternative a_k, then surely, a_i is preferred to a_k. In our method, if a_i is preferred to a_j, there must be $\eta_i > \eta_j$, and similarly, $\eta_j > \eta_k$. Given this, we have $\eta_i > \eta_k$, which means that a_i is preferred to a_k. Thus, we can believe that our method agrees with this property.

Independence of irrelevant alternatives (IIA) states that the relation of preference between two alternatives cannot depend on others alternatives being present or absent. While for our method, in the competition graph, the removal or addition of nodes can change the structure of the graph, and our method focuses primarily on the macroscopic and statistical index of all alternatives in the competition graph, as a result, the final aggregated ranking can be thus changed to some extent. On the other hand, if a rank aggregation method can give stable results despite changes in the input rankings, then this method can be considered to be robust, which is a desirable property. In this sense, IIA can be connected with the robustness research in this field, and this is what we have been studying recently.

In addition, we treat all input rankings equally, which means that non-imposition and non-dictatorship are also agreed in our rank aggregation method.

4. Comparison with typical rank aggregation methods

To test the effectiveness of the proposed method (CG), in this section, we compare it with four typical rank aggregation methods introduced in Section 2: the Borda's method (BM), the Dowdall method (DM), Borda(M^r) and the minimum violations ranking method (MVR). We use the benchmark model introduced in Section 2.5 to generate synthetic data with different ranking accuracies and lengths.

4.1. Comparison of effectiveness for different ranking accuracies

To compare the effectiveness of the five methods in aggregating rankings with different accuracies, we perform numerical experiments and present the effectiveness measures of rank aggregation methods D with various β and L_0 in Figure 4. The specific data is also presented in Table 3. It can be observed that as the value of the ranking accuracy β increases, the value of the effectiveness measure D decreases rapidly, indicating that rank aggregation methods with high ranking accuracy β are more effective; this agrees with our intuition.

In particular, a threshold for ranking accuracy β is found, i.e., the value, above which the ranking accuracy rises, and the proposed method (CG) outperforms the other four methods. It should be noted that the threshold depends on the baseline length of the ranking L_0 and the variation of ranking length ΔL. For instance, the threshold is 0.3 if $L_0 = 50$ and $\Delta L = 15$, while the threshold could be slightly less than 0.3 if $L_0 = 30$ and $\Delta L = 9$.

In addition, we observed that when the accuracy β is high (equal to 0.9), MVR can be as effective as our CG method, whereas in the scenarios of low ranking accuracies, our CG method can perform much better than MVR because, as an optimization method, MVR is not robust to the noise or errors in the input rankings. In contrast, the graph-based CG method focuses not on the relationship of a single pair of alternatives but primarily on the macroscopic and statistical index of alternatives. In this sense, our CG method can be robust to the noise or errors in the input rankings. As a result, CG can aggregate the inaccurate rankings more effectively.

Table 3. Effectiveness measure D vs. ranking accuracy β for various L_0, where $N = 100$, $M = 1000$ and $\Delta L = 0.3L_0$. Values in bold and italic represent the best effectiveness measures among the five methods. The results were averaged over 100 independent trials.

	L_0=10					L_0=30					L_0=50				
β	BM	DM	BM(M^r)	MVR	CG	BM	DM	BM(M^r)	MVR	CG	BM	DM	BM(M^r)	MVR	CG
0.1	1707	1494	***1227***	1283	1236	1123	***680***	738	808	747	866	***459***	594	631	600
0.3	731	629	435	466	***424***	401	266	234	258	***225***	273	177	177	201	***174***
0.5	405	375	237	240	***191***	211	167	123	135	***96***	139	108	89	105	***72***
0.7	297	306	181	134	***99***	151	145	88	75	***42***	96	94	58	54	***30***
0.9	249	268	141	***58***	61	122	129	67	24	***18***	77	83	42	13	***9***

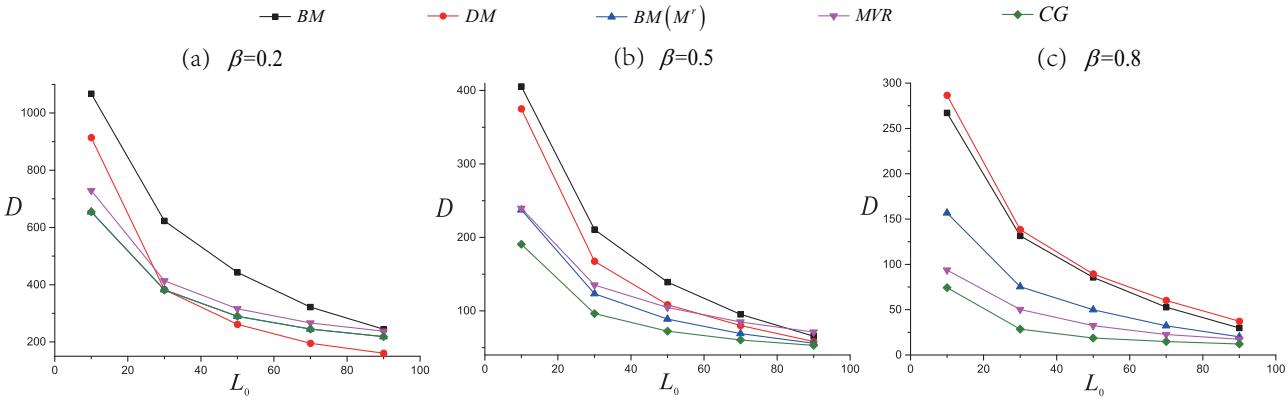

Figure 5. Effectiveness measure D vs. the baseline length L_0 for various β, where $N = 100$, $M = 1000$ and $\Delta L = 0.3L_0$. The results were averaged over 100 independent trials.

Table 4. Effectiveness measure D vs. the baseline length L_0 for various β, where $N = 100$, $M = 1000$ and $\Delta L = 0.3L_0$. Values in bold and italic represent the best effectiveness measures among the five methods. The results were averaged over 100 independent trials.

	$\beta = 0.2$					$\beta = 0.5$					$\beta = 0.8$				
L_0	BM	DM	BM(M^r)	MVR	CG	BM	DM	BM(M^r)	MVR	CG	BM	DM	BM(M^r)	MVR	CG
10	1067	914	***654***	729	665	405	375	237	240	***191***	267	287	157	94	***74***
30	623	383	382	414	***381***	211	167	123	135	***96***	132	138	75	50	***28***
50	443	***262***	289	316	290	139	108	89	105	***72***	86	89	50	32	***19***
70	322	***196***	245	266	245	95	80	69	85	***60***	53	60	32	23	***15***
90	244	***161***	218	237	219	66	58	56	71	***53***	30	37	20	17	***12***

4.2. Comparison of effectiveness for different ranking lengths

To compare the effectiveness of the five methods in aggregating rankings with different ranking lengths, we perform a host of experiments and display the effectiveness measure D as a function of the baseline length L_0 at various β in Figure 5. The specific data is shown in Table 4. It is observed that increased values of L_0 were associated with decreased values of measure D, indicating that the effectiveness of rank aggregation methods was better for larger baseline ranking lengths L_0. This finding was expected because larger baseline ranking lengths L_0 were associated with more complete evaluation information.

We observe that, in most cases, the proposed method produces much better results than those of the other four methods, especially when the baseline length is small. Moreover, we observe that BM and DM perform poorly in general, especially when the ranking length is small. The reason is that BM and DM only use part of the ranking information to complete the aggregation. In fact, the Borda count of the alternative a_j is merely the out-degree d_j^+ of v_j in our method, which means that BM only considers the occurrences of winning but ignores those of losing. Thus, BM cannot evaluate the alternatives comprehensively. It should be noted that the rankings in many early applications were mostly the complete lists; thus, BM has become one of the most widely used rank aggregation methods. Moreover, we can observe from Figure 5 that if the ranking length is small, the difference in effectiveness between BM (DM) and CG is distinct, while it becomes very small if the rankings are nearly complete, which further strengthens the analysis above.

4.3. Comparison of computational efficiency

To demonstrate the computational efficiency of our methods, we have performed numerical experiments with several large values of M, where N was fixed at 100, and presented the running time T using the

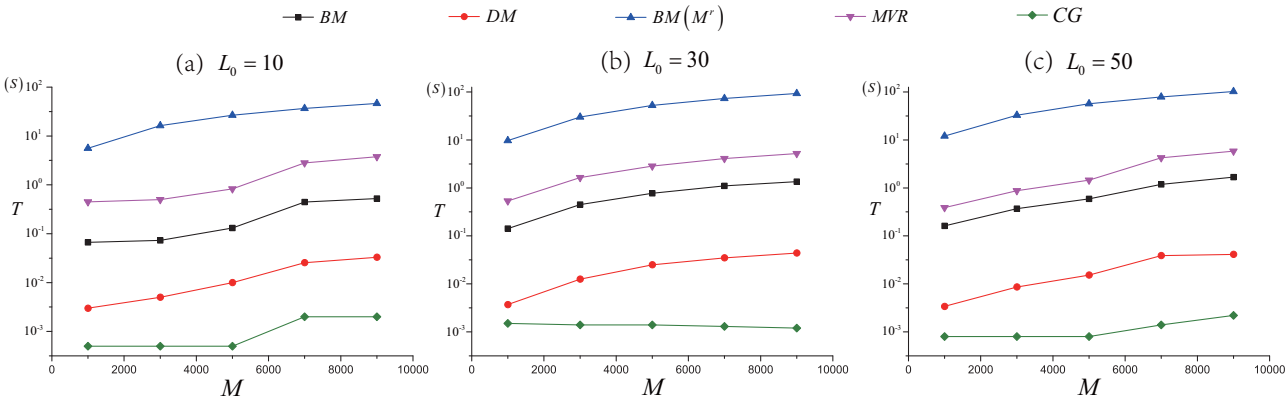

Figure 6. Running time T vs. number of voters M for various L_0, where $\beta = 0.5$ and $\Delta L = 0.3L_0$. The results were averaged over 100 independent trials.

Table 5. Running time T of CG vs. number of voters M for various L_0, where $N = 100$, $\beta = 0.5$ and $\Delta L = 0.3L_0$. The results were averaged over 100 independent trials.

M	$L_0 = 10$	$L_0 = 30$	$L_0 = 50$
1000	0.001	0.002	0.002
5000	0.001	0.002	0.002
10000	0.002	0.002	0.002
50000	0.003	0.003	0.003
100000	0.004	0.004	0.004

semi-logarithmic scale for various M and L_0 in Figure 6. We observe that the CG method can always be more efficient than the four other methods in different situations. To show the efficiency of our method more clearly, in Table 5, we have listed the running time T of the CG method for several large values of M. We observe that CG can aggregate the high-dimensional rankings in a very short time (less than 0.01 second) even if there are 10,000 rankings. Note that due to the "for" loops running very slowly in MATLAB, BM(M^r) did not perform as well as other methods. All experiments were performed using a computer with Intel dual-core i5 CPU and 4 GB of RAM, running Windows 7. All methods were implemented in MATLAB-R2013b.

5. Empirical analysis

Although the proposed rank aggregation method performed well in experiments using synthetic datasets, its practical application remains unconfirmed. Therefore, it is imperative to apply the method to an empirical dataset. We choose the teacher rankings given by students as the empirical data. There are two reasons teacher rankings are suitable for this study. In the context of student evaluation of teachers where students rank teachers rather than rate them, a large number of rankings are first given by students describing their interactions as well as the win-loss records of teachers. Second, it is conceivable that these rankings are extremely partial due to each student being taught by a small subset of teachers. In this sense, student evaluations of teachers represent an optimal source of data. A real-world dataset as of 2015 from the National University of Defense Technology, China, was used for this purpose, with the total of 7199 students and 1139 teachers. It should be noted that the large number of students and the diversity of the students' curricula ensured that the competition graph of teachers was connected.

In this dataset, the lengths of rankings varied from 5 to 28 with an average of 13. These individual rankings were remarkably partial relative to the total number of teachers (1139). The largest ROID in the dataset was 238, with out-degree of 237 and in-degree of 0, i.e., all of 237 students taught by the respective teacher ranked that teacher as the first. Conversely, the smallest ROID was 0, with in-degree of 15 and out-degree of 0, i.e., all of 15 students taught by the respective teacher ranked that teacher as the last. We analysed the final result and determined that teachers of required courses performed better than those of elective courses (p-value = 0.01438). The average ROID for teachers of required courses was 2.15, whereas it was only 1.06 for teachers of elective courses.

To assess the effectiveness and feasibility of our method, we have performed an extensive practical survey following the final teacher ranking. The results of the survey show that the final aggregated ranking of teachers is a good reflection of the actual situation and fits the expectations of evaluation experts well. We observe that most of the top-10 teachers in the final ranking have won the national award for teaching. For instance, Wei Chen, a teacher from the College of Computer Science who placed second in the final ranking with ROID of 17.27 (out-degree of 776 and in-degree of 44), has won the first prize in the National Teaching Competition 2016. Another example, Daquan Liu, a teacher at the College of Space Science and Engineering who placed fourth in the final ranking with ROID of 14.93 (out-degree of 208 and in-

degree of 13), was the winner of the Award for Bringing up Talents in Universities.

6. Conclusions and discussion

With the increasing emphasis on the problem of multiple-criteria decision-making, rank aggregation has emerged as an effective approach to this scenario, attracting significant attention from various fields for several decades. However, the expected roles and tasks of rank aggregation methods arising from the rapid development of information technology are increasingly comprehensive and necessary for decision-making. The task of aggregation of high-dimensional and partial rankings is encountered in many situations, posing challenges to existing rank aggregation methods.

In this article, a graph-based rank aggregation method has been proposed for aggregating large numbers of partial rankings into a consensus. In this method, we first defined the competition graph based on the competition matrix, in which the nodes represented alternatives, the directed edges represented the win-loss records among the alternatives, and the weights of a directed edge represented the number of times an alternative was ranked ahead of another alternative. Afterwards, we introduced the concept of the "ratio of out- and in-degrees (ROID)" as a new index for evaluating alternatives. The proposed method was compared with four typical rank aggregation methods using various synthetic data. Results suggested that the proposed method could significantly outperform other methods, especially in cases of low accuracy and incomplete information. Compared with rank aggregation methods based on optimization, such as MVR, our graph-based method is more robust to the noise or errors in the input rankings. Furthermore, because BM and DM use only part of the ranking information to complete the aggregation, these methods always perform poorly when the rankings are particularly partial. In contrast, the proposed method overcomes this drawback by considering the losing records and developing a novel approach to evaluating the alternatives. In addition, the computational efficiency of our method is particularly high in aggregating high-dimensional rankings.

Real-world experimental results based on the teacher rankings given by students were used to demonstrate the applicability of the novel method to aggregating high-dimensional partial rankings. It should be noted that this method could also be applied to other evaluation problems, including political elections, world university rankings, movie recommendations, or brand evaluations. In the future, we will extend our method to specific application contexts. We must note that we do not consider the strength of the opponents when calculating the ROID of the alternatives. However, this problem could be solved if we performed the competition matrix operation iteratively. In conclusion, the proposed method provides insights into the rank aggregation research and can be an effective and efficient tool for aggregating high-dimensional and partial rankings.

Acknowledgements

We thank Yapeng Li, Mingze Qi and Ye Deng for their helpful insights.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Jun Wu acknowledges the National Natural Science Foundation of China under Grant Nos. 71871217, 71690233, 71371185 and the Natural Science Foundation of Hunan Province under Grant No. 2019JJ20019. Hongzhong Deng acknowledges the National Natural Science Foundation of China under Grant No. 71771214. Xin Lu acknowledges the National Natural Science Foundation of China under Grant No. 71522014, 71771213, 71790615, and 91846301.

References

Ahn, B. S. (2017). Aggregation of ranked votes considering different relative gaps between rank positions. *Journal of the Operational Research Society*, *68*, 1307–1311. doi:10.1057/s41274-016-0153-8

Aledo, J. A., Gámez, J. A., & Molina, D. (2016). Using extension sets to aggregate partial rankings in a flexible setting. *Applied Mathematics and Computation*, *290*, 208–223. doi:10.1016/j.amc.2016.06.005

Aledo, J. A., Gámez, J. A., & Alejandro, R. (2017). Utopia in the solution of the Bucket Order Problem. *Decision Support Systems*, *97*, 69–80. doi:10.1016/j.dss.2017.03.006

Aledo, J. A., Gámez, J. A., & Alejandro, R. (2018). Approaching rank aggregation problems by using evolution strategies: The case of the optimal bucket order problem. *European Journal of Operational Research*, *270*, 982–998. doi:10.1016/j.ejor.2018.04.031

Ali, I., Cook, W. D., & Kress, M. (1986). On the minimum violations ranking of a tournament. *Management Science*, *32*, 660–672. doi:10.1287/mnsc.32.6.660

Amodio, S., D'Ambrosio, A., & Siciliano, R. (2016). Accurate algorithms for identifying the median ranking when dealing with weak and partial rankings under the Kemeny axiomatic approach. *European Journal of Operational Research*, *249*, 667–676. doi:10.1016/j.ejor.2015.08.048

Argentini, A., & Blanzieri, E. (2012). *Ranking aggregation based on belief function*. IPMU 2012, Proceedings of the 14th International Conference on Information

Processing and Management of Uncertainty in Knowledge-Based Systems. *299*, 511–520.

Arrow, K. J. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy*, *58*, 328–346. doi:10.1086/256963

Arrow, K. J. (1952). Social choice and individual values. *Journal of Political Economy*, *60*, 422–432.

Baucells, M., & Sarin, R. K. (2003). Effect of Three Voting Rules on Resource Allocation Decisions. *Management Science*, *49*, 1105–1118. doi:10.1287/mnsc.49.8.1105.16400

Borda, J. C. (1781). Mémoire sur les élections au scrutin. *Histoire de L'Académie Royale Des Sciences*,

Chartier, T. P., Kreutzer, E., Langville, A. N., Pedings, K., & Yamamoto, Y. (2010). *Minimum violations sports ranking using evolutionary optimization and binary integer linear program approaches*. Proceedings of the Tenth Australian Conference on Mathematics and Computers in Sport, A. Bedford and M. Ovens, eds., MathSport (ANZIAM), New South Wales, Australia. 13–20.

Cook, W. D., & Kress, M. (1990). A data envelopment model for aggregating preference rankings. *Management Science*, *36*, 1302–1310. doi:10.1287/mnsc.36.11.1302

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). *Rank aggregation methods for the web*. WWW 2001, Proceedings of the 10th international conference on World Wide Web. New York, 613–622.

Filippo, R. (2011). Who is the best player ever? A complex network analysis of the history of professional tennis. *Plos One*, *6*, e17249. doi:10.1371/journal.pone.0017249

Fishburn, P. C. (1977). Condorcet social choice functions. *SIAM Journal on Applied Mathematics*, *33*, 469–489. doi:10.1137/0133030

Kolde, R., Laur, S., Adler, P., & Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics (Oxford, England)*, *28*, 573–580. doi:10.1093/bioinformatics/btr709

Langville, A. N., & Meyer, C. D. (2012). *Who is #1? The science of rating and ranking*. Princeton, NJ: Princeton University Press.

Moritz, R. L. V., Reich, E., Schwarz, M., Bernt, M., & Middendorf, M. (2015). Refined ranking relations for selection of solutions in multi objective metaheuristics. *European Journal of Operational Research*, *243*, 454–464. doi:10.1016/j.ejor.2014.10.044

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-0120, Computer Science Department, Stanford University.

Park, J. (2005). On minimum violations ranking in paired comparisons. *arXiv Preprint Physics*, */0510242*.

Pedings, K. E., Langville, A. N., & Yamamoto, Y. (2012). A minimum violations ranking method. *Optimization and Engineering*, *13*, 349–370. doi:10.1007/s11081-011-9135-5

Read, M. J., Edwards, J. S., & Gear, A. E. (2000). Group preference aggregation rules: The results of a comparative analysis with in situ data. *Journal of the Operational Research Society*, *51*, 557–563. doi:10.1057/palgrave.jors.2600925

Reilly, B. (2002). Social Choice in the South Seas: Electoral Innovation and the Borda Count in the Pacific Island Countries. *International Political Science Review*, *23*, 355–372. doi:10.1177/0192512102023004002

Thieme, C., Prior, D., Tortosa-Ausina, E., & Gempp, R. (2016). Value added, educational accountability approaches and their effects on schools' rankings: Evidence from Chile. *European Journal of Operational Research*, *253*, 456–471. doi:10.1016/j.ejor.2016.01.023

Wang, Y. M., Chin, K. S., & Yang, J. B. (2007). Three new models for preference voting and aggregation. *Journal of the Operational Research Society*, *58*, 1389–1393. doi:10.1057/palgrave.jors.2602295

Xiao, Y., Deng, Y., Wu, J., Deng, H. Z., & Lu, X. (2017). Comparison of rank aggregation methods based on inherent ability. *Naval Research Logistics (NRL)*, *64*(7), 556–565. doi:10.1002/nav.21771