This article was downloaded by: [144.121.140.150] On: 02 October 2025, At: 03:52
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Spectral Ranking Inferences Based on General Multiway Comparisons

Jianqing Fan, Zhipeng Lou, Weichen Wang, Mengxin Yu

To cite this article:
Jianqing Fan, Zhipeng Lou, Weichen Wang, Mengxin Yu (2025) Spectral Ranking Inferences Based on General Multiway Comparisons. Operations Research

Published online in Articles in Advance 17 Jul 2025

. https://doi.org/10.1287/opre.2023.0439

Full terms and conditions of use: https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2025, INFORMS

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–20
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

Crosscutting Areas

Spectral Ranking Inferences Based on General Multiway Comparisons

Jianqing Fan,[a] Zhipeng Lou,[b] Weichen Wang,[c] Mengxin Yu[d,*]

[a] Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08544; [b] Department of Mathematics, University of California, San Diego, La Jolla, California 92093; [c] Innovation and Information Management, Faculty of Business and Economics, The University of Hong Kong, Hong Kong; [d] Statistics and Data Science, Washington University in St. Louis, St. Louis, Missouri 63130

*Corresponding author

Contact: jqfan@princeton.edu (JF); zlou@ucsd.edu (ZL); weichenw@hku.hk (WW); myu@wustl.edu, https://orcid.org/0000-0002-6818-4083 (MY)

Received: August 13, 2023
Revised: February 29, 2024; May 11, 2025
Accepted: May 22, 2025
Published Online in Articles in Advance: July 17, 2025

Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2023.0439

Copyright: © 2025 INFORMS

Abstract. This paper studies the performance of the spectral method in the estimation and uncertainty quantification of the unobserved preference scores of compared entities in a general and more realistic setup. Specifically, the comparison graph consists of hyper-edges of possible heterogeneous sizes, and the number of comparisons can be as low as one for a given hyper-edge. Such a setting is pervasive in real applications, circumventing the need to specify the graph randomness and the restrictive homogeneous sampling assumption imposed in the commonly used Bradley-Terry-Luce (BTL) or Plackett-Luce (PL) models. Furthermore, in scenarios where the BTL or PL models are appropriate, we unravel the relationship between the spectral estimator and the maximum likelihood estimator (MLE). We discover that a two-step spectral method, where we apply the optimal weighting estimated from the equal weighting vanilla spectral method, can achieve the same asymptotic efficiency as the MLE. Given the asymptotic distributions of the estimated preference scores, we also introduce a comprehensive framework to carry out both one-sample and two-sample ranking inferences, applicable to both fixed and random graph settings. It is noteworthy that this is the first time effective two-sample rank testing methods have been proposed. Finally, we substantiate our findings via comprehensive numerical simulations and subsequently apply our developed methodologies to perform statistical inferences for statistical journals and movie rankings.

Funding: This work was supported by the Office of Naval Research [Grant N00014-25-1-2317] and the National Science Foundation [Grants DMS-2052926, DMS-2053832, and DMS-2210833]. The work described in this paper was also partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 27307623).

Supplemental Material: All supplemental materials, including the code, data, and files required to reproduce the results, are available at https://doi.org/10.1287/opre.2023.0439.

Keywords: entity ranking • statistical inference • general comparison graph • spectral method

1. Introduction

Rank aggregation is crucial in various applications, including web search (Dwork et al. 2001, Wang et al. 2016), primate intelligence experiments (Johnson et al. 2002), assortment optimization (Aouad et al. 2018, Chen et al. 2020), recommendation systems (Baltrunas et al. 2010, Li et al. 2019), sports ranking (Massey 1997, Turner and Firth 2012), education (Avery et al. 2013, Caron et al. 2014), voting (Plackett 1975, Mattei and Walsh 2013), and instruction tuning used in the recent popular large language model ChatGPT (Ouyang et al. 2022). Therefore, it becomes an essential problem in many fields, such as psychology, econometrics, education, operations research, statistics, machine learning, artificial intelligence, etc.

Luce (1959) introduced the celebrated Luce's axiom of choice. Let $p(i|A)$ be the probability of selecting item i over all other items in the set of alternatives A. According to the axiom, when comparing two items i and j in any sets of alternatives A containing both i and j, the probability of choosing i over j is unaffected by the presence of other alternatives in the set. In specific, the axiom postulates that

$$\frac{\mathbb{P}(i \text{ is preferred in } A)}{\mathbb{P}(j \text{ is preferred in } A)} = \frac{\mathbb{P}(i \text{ is preferred in } \{i,j\})}{\mathbb{P}(j \text{ is preferred in } \{i,j\})}.$$

This assumption gives rise to a unique parametric choice model, the Bradley-Terry-Luce (BTL) model for pairwise comparisons, and the Plackett-Luce (PL) model for M-way rankings $M \geq 2$.

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

In this paper, we consider a collection of n items whose true ranking is determined by some unobserved preference scores θ_i^* for $i = 1, \ldots, n$. In this scenario, the BTL model assumes that an individual or a random event ranks item i over j with probability $\mathbb{P}(\text{item } i \text{ is preferred over item } j) = e^{\theta_i^*}/(e^{\theta_i^*} + e^{\theta_j^*})$. The Plackett-Luce model is an expanded version of pairwise comparison, which allows for a more comprehensive M-way full ranking, as initially described in Plackett (1975). This model takes individual preferences into account when ranking a selected subset of items with size $M < \infty$ (among all n items), which we represent as $i_1 \succ \cdots \succ i_M$. Think of this full ranking as $M-1$ distinct events where i_1 is favored over the set $i_1, \ldots, i_M$, followed by i_2 being favored over the set $i_2, \ldots, i_M$, and so on. The PL model calculates the probability of a full ranking $i_1 \succ \cdots \succ i_M$ using the formula

$$\mathbb{P}(i_1 \succ \cdots \succ i_M) = \prod_{j=1}^{M-1} \left[e^{\theta_{i_j}^*} \Big/ \sum_{k=j}^{M} e^{\theta_{i_k}^*} \right].$$

Next, we will give a brief introduction to the literature that has made progress on model estimation and uncertainty quantification for the BTL and the PL models over the parametric model.

1.1. Related Literature

A series of papers studied model estimation or inference based on the BTL or PL models. In the case of the Bradley-Terry-Luce model, its theoretical characteristics were solidified through a minorization-maximization algorithm, as outlined by Hunter (2004). Additionally, Negahban et al. (2012) developed an iterative rank aggregation algorithm called *Rank Centrality* (spectral method), which can recover the BTL model's underlying scores at an optimal ℓ_2-statistical rate. Subsequently, Chen and Suh (2015) used a two-step methodology (spectral method followed by maximum likelihood estimation (MLE)) to examine the BTL model in a context where the comparison graph is based on the Erdős-Rényi model, where every item pair is assumed to have a probability p of being compared, and once a pair is connected, it will be compared for L times. Subsequently, under a similar setting with Chen and Suh (2015), Chen et al. (2019) investigated the optimal statistical rates for recovering the underlying scores, demonstrating that regularized MLE and the spectral method are both optimal for retrieving top-K items when the conditional number is a constant. They derived ℓ_2- as well as ℓ_∞-rates for the unknown underlying preference scores in their study. Furthermore, Chen et al. (2022) extended the investigation of Chen et al. (2019) to the partial recovery scenarios and improved the analysis to unregularized MLE.

Expanding beyond simple pairwise comparisons, researchers also explored ranking issues through M-way comparisons, where $M \geq 2$. The Plackett-Luce model and its variations serve as prominent examples in this line of study, as evidenced by numerous references (Plackett 1975, Guiver and Snelson 2009, Cheng et al. 2010, Hajek et al. 2014, Maystre and Grossglauser 2015, Szörényi et al. 2015, Jang et al. 2018). In particular, Jang et al. (2018) investigated the Plackett-Luce model within the context of a uniform hyper-graph, where a tuple with size M is compared with probability p and once a tuple is connected or compared in the hyper-graph, it will be compared for L times. By dividing M-way comparison data into pairs, they employ the spectral method to obtain the ℓ_∞- statistical rate for the underlying scores. Additionally, they presented a lower bound for sample complexity necessary to identify the top-K items under the Plackett-Luce model. In a more recent development, under the same model setting, Fan et al. (2022b) enhanced the findings of Jang et al. (2018), focusing solely on the top choice. Rather than splitting the comparison data into pairwise comparisons, they applied the MLE and matched the sample complexity lower bound needed to recover the top-K items. This contrasts with the work by Jang et al. (2018), which requires a significantly denser comparison graph or a much larger number of comparisons for their results to hold.

The aforementioned literature primarily concentrated on the nonasymptotic statistical consistency for estimating item scores. However, the results of limiting distribution for ranking models remained largely unexplored. Only a limited number of findings on asymptotic distributions of estimated ranking scores exist under the Bradley-Terry-Luce model, whose comparison graphs are sampled from the Erdős-Rényi graph with connection probability p and for which each observed pair has the same number of comparisons L. For example, Simons and Yao (1999) established the asymptotic normality of the BTL model's maximum likelihood estimator when all comparison pairs are entirely conducted (i.e., $p = 1$). Han et al. (2020) expanded these findings for dense, but not fully connected, comparison graphs (Erdős-Rényi random graphs) where $p \gtrsim n^{-1/10}$. Recently, Liu et al. (2022) introduced a Lagrangian debiasing approach to derive asymptotic distributions for ranking scores, accommodating sparse comparison graphs with $p \asymp 1/n$ but necessitating comparison times $L \gtrsim n^2$. Furthermore, Gao et al. (2021) employed a "leave-two-out" technique to determine asymptotic distributions for ranking scores, achieving optimal sample complexity and allowing $L = O(1)$ in the sparse comparison graph setting (i.e., $p \asymp 1/n$ up to logarithm terms). Fan et al. (2022a) extended existing research by incorporating covariate information into the BTL model. By introducing an innovative proof, they presented the MLE's asymptotic variance with optimal sample complexity when $p \asymp 1/n$ and $L = O(1)$. In the sequel, Fan et al. (2022b)

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

broadened the asymptotic results of the BTL model to encompass PL models with $M \geq 2$, again with optimal sample complexity. They also developed a unified framework to construct rank confidence intervals, requiring the number of comparisons $L \gtrsim \text{poly}(\log n)$. Moreover, recently, Han and Xu (2023) further extended the settings in Fan et al. (2022b) by investigating the asymptotic distribution of the MLE, where the comparisons are generated from a mixture of Erdős-Rényi graphs with different sizes or a hyper-graph stochastic block model.

Finally, we discuss related literature of an important application of our framework: assortment optimization (Talluri and Van Ryzin 2004; Rusmevichientong and Topaloglu 2012; Vulcano et al. 2012; Davis et al. 2014; Chen et al. 2020, 2023; Zhang et al. 2020; Shen et al. 2023), which is of great importance in revenue management. Specifically, in their settings, each product (including the no-purchase alternative) is also associated with an unknown customer preference score, which can characterize the customers' choice behavior over a set of offered products. Based on the consistently estimated preference scores and the available profit information of each product, various efficient algorithms have been proposed to determine the optimal assortment under different kinds of practical constraints (Talluri and Van Ryzin 2004, Rusmevichientong et al. 2010, Gallego and Topaloglu 2014, Sumida et al. 2020). Moreover, uncertainty quantification of the estimated preference scores also enables statistical inference on the properties of the optimal assortment (Shen et al. 2023).

1.2. Motivations and Contributions

In this section, we discuss our motivation and problem settings and compare our results with previous literature in different aspects, namely, the comparison graph, the connection between the spectral method and MLE, and ranking inferences.

1.2.1. Comparison Graph. Previous studies on parametric ranking models mostly require comparison graphs derived from a specific random graph. For example, in the endeavor of understanding multiway comparisons (e.g., Jang et al. 2016, Fan et al. 2022b), it is typically assumed that the comparisons are generated explicitly from a homogeneous graph with a known distribution. This assumption may pose a challenge in certain contexts. Although practical applications with homogeneous comparisons exist, it is sometimes unrealistic to presume that all comparisons are generated from a known homogeneous distribution. In fact, there are more cases where we see heterogeneous comparisons, where some are compared more often than others and the comparison graph generation process is unknown. We will present Example 1 as our motivation.

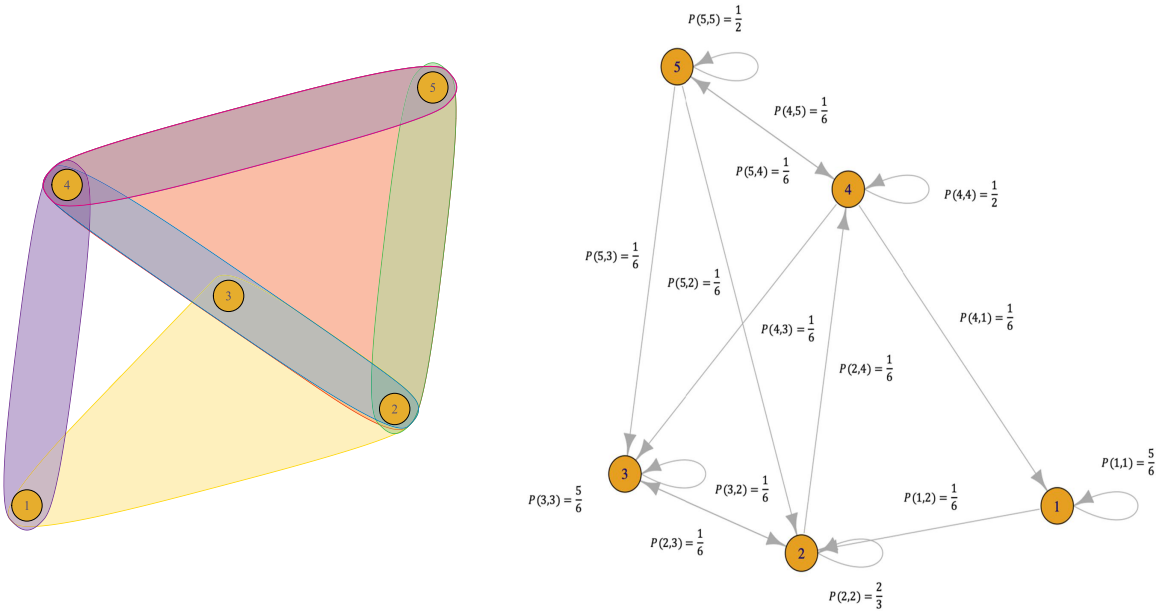
Example 1. There is a sequence of customers buying goods. For the l-th customer, according to her preference, her reviewed products are A_l (a choice set), and she finally chose product $c_l \in A_l$ (her top choice). Then the total data sets presented to us are $\{(A_l, c_l)\}_{l=1}^{D}$. If all choice sets are presented to the customers with the same probability and the reviewed number of items are of the same size (such as pairwise comparisons), we say the comparison graph is homogeneous. But if some choice sets, possibly with different sizes, are presented more often to the customers or are chosen arbitrarily based on the customers' preference profiles, then this heterogeneous comparison graph cannot be well approximated by a given random graph model. That is, the comparisons may not follow, say, the BTL or PL models with Erdős-Rényi types of uniform graphs.

The heterogeneous comparison scheme in Example 1 above is applicable across a wide range of practical scenarios. For instance, it covers the typical setup of the assortment optimization, wherein the no-purchase alternative is also included in the choice set A_l. We give a toy example with five products in Figure 1, where sizes of A_l are among $\{2, 3, 4\}$. Because of the heterogeneity, it is unrealistic to assume A_l is of the same size or sampled from an explicit random graph. However, most of the previous works focused on statistical properties under certain ad hoc random graphs, most commonly the Erdős-Rényi type of uniform graphs, for example, Chen and Suh (2015), Chen et al. (2019), Jang et al. (2016), Gao et al. (2021), Fan et al. (2022b), Liu et al. (2022), and Han and Xu (2023). One interesting piece of research that indeed worked with the fixed comparison graph is Shah et al. (2015), who explored the optimality of MLE in ℓ_2 estimation with pairwise comparisons. Following this work, Li et al. (2022) further discussed the optimality of MLE in ℓ_∞ error. Still, little has been known about the inference results for general fixed comparison graphs.

In this paper, we focus on the setting of a general fixed comparison graph, where we circumvent the modeling for the complicated graph-generating process. Specifically, we study statistical estimation and uncertainty quantification of the preference scores over a general, heterogeneous choice set via the spectral method. In addition, we also study the theoretical performance of the spectral method when we do have a homogeneous random comparison graph, and compare the results. Our results require slightly stronger conditions to handle fixed graph settings because we need to make sure each item has enough information to be ranked.

For the general setting, we denote the choice preference for the l-th comparison as (c_l, A_l), wherein A_l signifies the set of choices with heterogeneous size, which

Figure 1. (Color online) Example with Five Toy Products



Notes. A simple illustration of a collection of $\{c_l, A_l\}_{\ell=1}^{D}$ with $(c_1, A_1) = (3, \{2,3,4,5\})$, $(c_2, A_2) = (2, \{1,2,3\})$, $(c_3, A_3) = (2, \{2,5\})$, $(c_4, A_4) = (4, \{4,5\})$, $(c_5, A_5) = (4, \{2,4\})$, $(c_6, A_6) = (1, \{1,4\})$, $(c_7, A_7) = (5, \{4,5\})$. The left panel illustrates the choice sets $\{A_l\}_{\ell=1}^{7}$, where all nodes inside each A_l are surrounded by an open area with the same color. The right panel presents the comparison-induced Markov transition matrix, whose computation is detailed in Section 2.2. A directed edge from i to $j, (j \neq i)$ exists if and only if $i, j, (i \neq j)$ are compared in some A_l and j is the winner $(c_l = j)$.

can be either fixed or random, and $c_l \in A_l$ represents the most preferred item in A_l. Hence, in the l-th comparison, we understand that c_l outranks all other elements within A_l. As such, our broadest comparison data are symbolized as $\mathcal{D} = \{l | (c_l, A_l)\}$. The associated collection of choice sets is denoted as $\mathcal{G} = \{A_l | l \in \mathcal{D}\}$. This framework also contains the Plackett-Luce model as a special case, if we treat the PL model as $M-1$ selections over $M-1$ choice sets. Under mild conditions, we manage to obtain optimal statistical rates for our spectral estimator conditional on the comparison graphs and specify the explicit form of its asymptotic distribution. This gives an efficient solution when one encounters heterogeneous comparison graphs. In addition, because the graph is fixed or conditioned upon, it is not necessary for us to repeat each comparison for $L \geq 1$ times. We can even accommodate situations where a choice set is chosen and compared for just a single time, which is true in many practical applications.

1.2.2. Connection Between the Spectral Estimator and the MLE. Our general setting, as introduced in Section 1.2.1, encompasses homogeneous random comparison graphs as a particular instance. The bulk of prior research has centered around the evaluation of the MLE or the spectral method when applied to homogeneous comparisons (Chen and Suh 2015; Chen et al. 2019, 2022; Fan et al. 2022a, 2022b). Both are effective methods for examining ranking issues within the context of the BTL or PL model. Hence, an interesting question arises: What is the relationship between the MLE and the spectral method?

A handful of studies have offered insights into this question. For instance, Maystre and Grossglauser (2015) identified a link between the MLE and spectral method in multiway comparisons, where the spectral estimator aligns with the MLE through the iterative updating of specific weighting functions in constructing the spectral estimator. This connection was only limited to the first order in the sense that the paper only concerns the convergence property. Furthermore, Gao et al. (2021) demonstrated that the asymptotic variance of the spectral method exceeds that of the MLE in pairwise comparisons using the BTL model. However, this discrepancy arises from their choice of a suboptimal weighting function for the spectral estimator.

In our paper, we leverage the homogeneous random comparison graph case (as it is the most popularly studied setting in many previous articles) to illustrate that by employing certain optimally estimated information weighting functions, the asymptotic variance of the spectral estimator matches that of the MLE with multiway comparisons in the PL model. Therefore, the MLE could be considered a "two-step" spectral method, where in the first step we consistently estimate the unobserved preference scores, and in the second step we use the proper weighting in the spectral method to achieve an efficient estimator. It is also noteworthy that we achieve the optimal sample complexity over the sparsest sampling graph up to logarithmic terms.

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

1.2.3. Ranking Inferences: One Sample vs. Two Samples. As another contribution, we also study several rank-related inference problems. We have the following motivating example:

Example 2. First, we consider the one-sample inference problem. Consider a group of candidate items $\{1,\dots,n\}$ and one observed data set on their comparisons; we are interested in

- Building the confidence intervals for the ranks of certain items $\{r_1,\dots,r_m\}$ of our interest.
- Testing whether a given item m is in the top-K set, which includes K best items.

Second, we consider the two-sample inference problem. For two groups of data sets of the same list of items $\{1,\dots,n\}$, we are interested in

- Testing whether the rank of a certain item m is preserved across these two samples (e.g., groups or time periods).
- Testing whether the top-K sets have changed or not.

Rankings are ubiquitous in real-world applications, for instance, in the evaluation of universities, sports teams, or web pages. However, most of these rankings provide only first-order information, presenting the results without offering any measure of uncertainty. For example, under the BTL model, when two items have equivalent underlying scores, there's a 50% probability of one being ranked higher than the other. Thus, rankings between these two items could be unreliable because of their indistinguishable underlying scores, emphasizing the necessity for confidence intervals in rankings.

Given these critical considerations, our study offers a comprehensive framework that efficiently solves the problems outlined in Example 1 and Example 2 over heterogeneous comparison graphs. Additionally, our approach enhances the sample complexity of several previous works. For instance, when restricting our general framework to homogeneous random comparison graphs, regarding the one-sample inference, Liu et al. (2022) required $L \gtrsim n^2$ to carry out the statistical inference, whereas Fan et al. (2022b) further improved this requirement to $L \gtrsim \text{poly}(\log n)$. In our paper, our framework can allow $L = \mathcal{O}(1)$ and even $L = 1$. Furthermore, two-sample ranking inference, which can be widely applied in real-world scenarios like policy evaluation, treatment effect comparison, change point detection, etc., has not been previously studied. Our paper also introduces a general framework for studying the two-sample ranking inference problems, again offering optimal sample complexity.

1.2.4. Theoretical Contributions. We build up our theoretical analyses based on some previously developed techniques from Gao et al. (2021) and Chen et al. (2019). However, our proofs have the following novelty: In those two papers and other papers with a random comparison graph, graph randomness and ranking outcome randomness are typically intertwined in the analysis. We will separate them and reveal the proper quantities of interest to summarize the information in a fixed comparison graph. We study the connection between these quantities and the spectral ranking performance and provide sufficient conditions under the fixed graph for valid ranking inference. This theoretical attempt has not previously been seen in the literature. In addition, all our analyses allow for varying comparison sizes and an arbitrary number of repetitions of each comparison set. This significantly broadens the applicability of our proposed methodology, as in practice, a lot of ranking problems contain nonrepeating comparisons of different numbers of items. We also work on the theory when we also have graph randomness. We realize that the homogeneity of sampling each comparison tuple can lead to more relaxed assumptions. With more relaxed conditions, we clearly show where and how we can achieve an improved performance guarantee (see the assumption and proof of Theorem 4). This part highlights the difference between fixed and random graphs and provides more theoretical insight into the role of graph randomness in spectral ranking.

1.3. Roadmap and Notations

In Section 2, we set up the model and introduce the spectral ranking method. Section 3 is dedicated to the examination of the asymptotic distribution of the spectral estimator, based on fixed comparison graphs and random graphs with the PL model, respectively. In the same section, we also introduce the framework designed for the construction of rank confidence intervals and rank testing statistics for both one-sample and two-sample analysis. Section 4 details the theoretical guarantees for all proposed methodologies. Sections 5 and 6 contain comprehensive numerical studies to verify theoretical results and two real data examples to illustrate the usefulness of our ranking inference methods. Finally, we conclude the paper with some discussion in Section 7. All the proofs are deferred to the Online Appendix. The code for reproducing both the simulation studies and real data analysis is available at https://github.com/MaxineYu/Spectral_Ranking.

Throughout this work, we use $[n]$ to denote the index set $\{1,2,\dots,n\}$. For any given vector $\mathbf{x} \in \mathbb{R}^n$ and $q \geq 0$, we use $\|\mathbf{x}\|_q = (\sum_{i=1}^n |x_i|^q)^{1/q}$ to represent the vector ℓ_q norm. For any given matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, we use $\|\cdot\|$ to denote the spectral norm of $\mathbf{X}$ and write $\mathbf{X} \succcurlyeq 0$ or $\mathbf{X} \preccurlyeq 0$ if $\mathbf{X}$ or $-\mathbf{X}$ is positive semidefinite. For event A, $1(A)$ denotes an indicator which equals one if A is true and zero otherwise. For two positive sequences $\{a_n\}_{n\geq 1}$, $\{b_n\}_{n\geq 1}$, we write $a_n = \mathcal{O}(b_n)$ or $a_n \lesssim b_n$ if there exists a positive constant C such that $a_n/b_n \leq C$ and we write $a_n = o(b_n)$ if $a_n/b_n \to 0$. In addition, $\mathcal{O}_p(\cdot)$ and $o_p(\cdot)$ share

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

similar meanings as $\mathcal{O}(\cdot)$ and $o(\cdot)$, respectively, but these relations hold asymptotically with probability tending to one. Similarly we have $a_n = \Omega(b_n)$ or $a_n \gtrsim b_n$ if $a_n/b_n \geq c$ with some constant $c > 0$. We use $a_n = \Theta(b_n)$ or $a_n \asymp b_n$ if $a_n = \mathcal{O}(b_n)$ and $a_n = \Omega(b_n)$. For two random variables A_n, B_n, if we write $A_n \approx B_n$, it holds that $A_n - B_n = o(1)$ with probability going to one. Given n items, we use θ_i^* to indicate the underlying preference score of the i-th item. Define $r : [n] \to [n]$ as the rank operator on the n items, which maps each item to its population rank based on the preference scores. We write the rank of the i-th item as r_i or $r(i)$. By default, we consider ranking from the largest score to the smallest score.

2. Multiway Comparison Model and Spectral Ranking

We first introduce a general discrete choice model, which encompasses the classical Plackett-Luce model as well as fixed comparison graph scenario.

2.1. Discrete Choice Model

We assume there are n items to be ranked. According to Luce's choice axiom (Luce 1959), the preference scores of a given group of n items can be parameterized as a vector $(\theta_1^*, \ldots, \theta_n^*)^\top$ such that $\mathbb{P}(i \text{ wins among } A) = e^{\theta_i^*}/(\sum_{k\in A} e^{\theta_k^*})$ for any choice set A and item $i \in A$. Because the parameters are only identifiable up to a location shift, without loss of generality, we assume $\sum_{i=1}^n \theta_i^* = 0$ for identification. We consider the general comparison model, where we are given a collection of comparisons and outcomes $\{(c_l, A_l)\}_{l\in\mathcal{D}}$. Here c_l denotes the selected item over the choice set A_l with probability $e^{\theta_{c_l}^*}/(\sum_{k\in A_l} e^{\theta_k^*})$.

Remark 1. This general comparison model contains many well-known special cases.

- For the BTL model, it is easy to set A_l as the pair being compared every time. If each pair is compared for L times independently, we just need to write the outcomes as (c_l, A_l) and reindex l.
- For the PL model, we have obtained the full ranking of a choice set $B = \{i_1, \ldots, i_B\}$. The probability of observing a certain ranking becomes

$$\begin{aligned}\mathbb{P}(i_1 > i_2 > \cdots > i_B) &= \mathbb{P}(i_1 \text{ wins among } C_1 \mid C_1 = B)\\ &\quad \cdot \mathbb{P}(i_2 \text{ wins among } C_2 \mid C_2 = B\{-i_1\}) \cdots\\ &\quad \cdot \mathbb{P}(i_{B-1} \text{ wins among } C_{B-1} \mid C_{B-1}\\ &\quad = B\{-i_1, \ldots, -i_{B-2}\})\\ &= \frac{e^{\theta_{i_1}^*}}{\sum_{j=1}^B e^{\theta_{i_j}^*}} \cdot \frac{e^{\theta_{i_2}^*}}{\sum_{j=2}^B e^{\theta_{i_j}^*}} \cdots \frac{e^{\theta_{i_{B-1}}^*}}{\sum_{j=B-1}^B e^{\theta_{i_j}^*}},\end{aligned}$$

where $C_i, i \geq 1$ is the i-th comparison set and $B\{-i_1, \ldots, -i_M\}$ denotes the set of remaining items after removing $\{i_1, \ldots, i_M\}$. These comparison results can also be decomposed into the comparisons

$$\{(i_1, B), (i_2, B\{-i_1\}), \ldots, (i_{B-1}, B\{-i_1, \ldots, -i_{B-2}\})\}.$$

- With fixed comparison graphs, $\{A_l, l \in \mathcal{D}\}$ are given and hence have no randomness, so the comparison results in c_l are assumed independent. In contrast, with a random comparison graph, such as in the PL model, A_l may be dependent. For instance, (θ_{i_1}, B) and $(\theta_{i_2}, B\{-i_1\})$ are dependent as $B\{-i_1\}$ depends on the winner of the first comparison i_1. Therefore, we have to explicitly lay out the random process assumption for comparison generation in order to study the theoretical properties in a case-by-case manner.

Recall that the general comparison data are denoted as $\{(c_l, A_l)\}_{l\in\mathcal{D}}$. The corresponding collection of choice sets is $\mathcal{G} = \{A_l \mid l \in \mathcal{D}\}$. When we only have pairwise comparisons, $|A_l| = 2$ and $\mathcal{G}$ represents the set of all edges we have compared. But in a general setting, A_l can have different sizes and we denote $M = \max_{l\in\mathcal{D}} |A_l| < \infty$ as the maximal size of the comparison hyper-graph edge. Also, if we have L independent comparisons of the same comparison hyper-edge A_l, we use different l to indicate the comparison. So in $\mathcal{G}$, the hyper-edge A_l may be compared for multiple times with different outcomes c_l.

Throughout this paper, we consider using the spectral method on the preference data based on multiway comparisons. We will first focus on the fixed comparison graph and then consider commonly used random comparison graph structures. Notice that no matter whether the comparison graph $\mathcal{G}$ is fixed or random, our spectral ranking methodology will be conditional on $\mathcal{G}$, which is observed in practice. The underlying model for generating $\mathcal{G}$ can be very general: it can be given, or random based on the Erdős-Rényi random graph with the same probability p, or more generally induced from some other comparison rules, which could even cause some A_l to be dependent. For example, if we view each comparison of the PL model as $M-1$ pairwise comparisons involving top-1 versus top-2, top-2 versus top-3, …, top-$M-1$ versus top-M, then the resulting comparison data, denoted as $(c_l = i_k, A_l = \{i_k, i_{k+1}\})$ for $k = 1, \ldots, M-1$, are dependent (even the definition of A_l depends on the complete comparison result).

2.2. Spectral Ranking

In the spectral method, we formally define a Markov chain, denoted as $M = (S, P)$. Here, S signifies the collection of n states corresponding to the n items to be compared, represented as vertices of a directed comparison graph, and P constitutes the transition matrix defined below. This matrix oversees transitions amongst the states by representing whether any two particular

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

states within S are capable of being connected via nonzero transition probabilities.

Define two index sets $\mathcal{W}_j, \mathcal{L}_i$ for comparisons, with j as the winner and i as the loser:

$$\mathcal{W}_j = \{l \in \mathcal{D} | j \in A_l, c_l = j\}, \qquad \mathcal{L}_i = \{l \in \mathcal{D} | i \in A_l, c_l \neq i\}.$$

So their intersection for $i \neq j$ gives all situations where i, j are compared and j wins, that is, $\mathcal{W}_j \cap \mathcal{L}_i = \{l \in \mathcal{D} | i, j \in A_l, c_l = j\}$. Define the transition matrix P with transition probability

$$P_{ij} = \begin{cases} \frac{1}{d} \sum_{l \in \mathcal{W}_j \cap \mathcal{L}_i} \frac{1}{f(A_l)}, & \text{if } i \neq j, \\ 1 - \sum_{k:k \neq i} P_{ik}, & \text{if } i = j. \end{cases}$$

Here, d is chosen to be large enough so that the diagonal element is nonnegative, but not too large to give enough transition probability. When the comparison graph is random, we choose d to make nonnegative diagonal elements with probability approaching one by studying the concentration inequality for $\sum_{k:k \neq i} P_{ik}$. Here, $f(A_l) > 0$ is a weighting function to encode the total information in the l-th comparison. A natural choice is $f(A_l) = |A_l|$, giving more weight to hyperedges with a smaller number of compared items. We will discuss later the optimal choice of $f(\cdot)$.

When $i \neq j$, P_{ij} can also be written as

$$P_{ij} = \frac{1}{d} \sum_{l \in \mathcal{D}} 1(i, j \in A_l) 1(c_l = j) \frac{1}{f(A_l)}.$$

Conditioning on $\mathcal{G}$, the population transition is

$$P^*_{ij} = E[P_{ij} | \mathcal{G}] = \begin{cases} \frac{1}{d} \sum_{l \in \mathcal{D}} 1(i, j \in A_l) \frac{e^{\theta^*_j}}{\sum_{u \in A_l} e^{\theta^*_u}} \frac{1}{f(A_l)}, & \text{if } i \neq j, \\ 1 - \sum_{k:k \neq i} P^*_{ik}, & \text{if } i = j. \end{cases}$$

Let

$$\pi^* = (e^{\theta^*_1}, \ldots, e^{\theta^*_n}) / \sum_{k=1}^n e^{\theta^*_k}.$$

Note that both $\sum_{u \in A_l} e^{\theta^*_u}$ and $f(A_l)$ in the denominator are symmetric with respect to θ^*_i, θ^*_j as long as both i, j belong to A_l. So we have $P^*_{ij} \pi^*_i = P^*_{ji} \pi^*_j$. This is the so-called detailed balance that leads to π^* being the stationary measure of the Markov chain with the above population transition for any $f(\cdot)$. That is, $\pi^{*\top} P^* = \pi^{*\top}$; namely, π^* is the top-left eigenvector of P^*.

The spectral method estimates π^* by using the stationary measure $\hat{\pi}$ of the empirical transition P, namely,

$$\hat{\pi}^\top P = \hat{\pi}^\top.$$

Note that if we consider the directed graph induced by P to be strongly connected, this implies that the Markov chain it generates will be ergodic, which ensures the existence of a unique stationary distribution $\hat{\pi}$ as defined above. Consider the toy example in Figure 1; if we naively choose $f(\cdot) = 1$ as a constant weighting function, it is not hard to count the times that j beats i and fill the value into the transition matrix P_{ij} (divided by d). In the right panel, the transition probabilities are calculated with $d = 6$ to guarantee that the self-loop transition happens with a positive probability. For this P, the stationary distribution is $\hat{\pi} = (0.199, 0.531, 0.796, 0.199, 0.066)^\top$, meaning that the estimated ranking of preference scores of the five products is $3 > 2 > 1 = 4 > 5$.

Finally, given the identifiability condition of $1^\top \theta^* = 0$, we can estimate θ^*_i by

$$\tilde{\theta}_i := \log \hat{\pi}_i - \frac{1}{n} \sum_{k=1}^n \log \hat{\pi}_k. \tag{1}$$

It is worth mentioning that the spectral estimator is easier to compute in practice, by only requiring one-step eigen-decomposition. Indeed, we need only the eigenvector that corresponds to the largest eigenvalue, which can even be computed very fast by the power method. In comparison, the MLE is typically computationally heavier in terms of data storage and step size determination during the implementation of the gradient descent algorithm.

3. Ranking Inference Methodology

In this section, we study the inference methodology for the spectral estimator for the underlying scores $\{\theta^*_i\}_{i \in [n]}$ of n items. To be specific, we need to establish the statistical convergence rates and asymptotic normality for $\tilde{\theta}_i$.

3.1. Uncertainty Quantification with Fixed Comparison Graph

For the estimation of π^*, we use the following two approximations, which we will justify later to be accurate enough so as not to affect the asymptotic variance. Let us first focus on our intuition. Firstly, we have

$$\hat{\pi}_i = \frac{\sum_{j:j \neq i} P_{ji} \hat{\pi}_j}{\sum_{j:j \neq i} P_{ij}} \approx \frac{\sum_{j:j \neq i} P_{ji} \pi^*_j}{\sum_{j:j \neq i} P_{ij}} =: \overline{\pi}_i.$$

Equivalently,

$$\frac{\hat{\pi}_i - \pi^*_i}{\pi^*_i} \approx \frac{\overline{\pi}_i - \pi^*_i}{\pi^*_i} = \frac{\sum_{j:j \neq i} (P_{ji} \pi^*_j - P_{ij} \pi^*_i)}{\pi^*_i \sum_{j:j \neq i} P_{ij}}. \tag{2}$$

Secondly, the denominator above can be approximated by its expected value so that (2) can further be approximated as

$$J^*_i := \frac{\sum_{j:j \neq i} (P_{ji} e^{\theta^*_j} - P_{ij} e^{\theta^*_i})}{\sum_{j:j \neq i} E[P_{ij} | \mathcal{G}] e^{\theta^*_i}}, \tag{3}$$

by using $\pi^*_i \propto e^{\theta^*_i}$. We will rigorously argue that the

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

asymptotic distributions of $\frac{\hat{\pi}_i-\pi_i^*}{\pi_i^*}$ and J_i^* are identical. For now, let us look at the asymptotic distribution of J_i^*. Obviously, it is mean zero because of the detailed balance: $E[P_{ji}|\mathcal{G}]\pi_j^* = E[P_{ij}|\mathcal{G}]\pi_i^*$. The denominator of J_i^* is a constant and can be explicitly written out as follows:

$$\begin{aligned}\tau_i(\theta^*) &:= \sum_{j:j\neq i} E[P_{ij}|\mathcal{G}]e^{\theta_i^*} \\ &= \frac{1}{d}\sum_{l\in\mathcal{D}} 1(i\in A_l)\left(1-\frac{e^{\theta_i^*}}{\sum_{u\in A_l} e^{\theta_u^*}}\right)\frac{e^{\theta_i^*}}{f(A_l)}. \end{aligned} \tag{4}$$

Thus, J_i^* can be expressed as

$$\begin{aligned} J_i^* &= \frac{1}{d\tau_i}\sum_{l\in\mathcal{D}}\frac{1(i\in A_l)}{f(A_l)}\left(1(c_l=i)\sum_{u\in A_l,u\neq i} e^{\theta_u^*} - e^{\theta_i^*}\sum_{u\in A_l,u\neq i} 1(c_l=u)\right) \\ &=: \frac{1}{d}\sum_{l\in\mathcal{D}} J_{il}(\theta^*), \end{aligned} \tag{5}$$

where τ_i is short for $\tau_i(\theta^*)$. Because each (c_l, A_l) is independent in the fixed graph setting (see Remark 1 for discussions), the variance of J_i^* is

$$\begin{aligned}\mathrm{Var}(J_i^*|\mathcal{G}) &= \frac{1}{d^2\tau_i^2}\sum_{l\in\mathcal{D}}\frac{1(i\in A_l)}{f^2(A_l)}\cdot \mathrm{Var}\left(1(c_l=i)\sum_{u\in A_l,u\neq i} e^{\theta_u^*} - e^{\theta_i^*}\sum_{u\in A_l,u\neq i} 1(c_l=u)\right) \\ &= \left(\sum_{l\in\mathcal{D}} 1(i\in A_l)\frac{(\sum_{u\in A_l} e^{\theta_u^*} - e^{\theta_i^*})e^{\theta_i^*}}{f(A_l)^2}\right) \\ &\quad \Big/ \left[\sum_{l\in\mathcal{D}} 1(i\in A_l)\left(\frac{\sum_{u\in A_l} e^{\theta_u^*} - e^{\theta_i^*}}{\sum_{u\in A_l} e^{\theta_u^*}}\right)\frac{e^{\theta_i^*}}{f(A_l)}\right]^2. \end{aligned} \tag{6}$$

A few important comments are in order. Firstly, the function f achieves the minimal variance in (6) when $f(A_l)\propto\sum_{u\in A_l} e^{\theta_u^*}$ because of simply applying the Cauchy-Schwarz inequality. Actually, Maystre and Grossglauser (2015) showed that when $f(A_l)=\sum_{u\in A_l} e^{\theta_u^*}$, the spectral method estimator converges to the MLE up to the first order. Secondly, under the situation of pairwise comparison in the BTL model, each (c_l, A_l) is independent and we assume in $\mathcal{D}$ each pair (i,j) is either compared for L times (denoted as $\tilde{A}_{ij}=1$) or never compared (denoted as $\tilde{A}_{ij}=0$). Further assuming $f(A_l)=|A_l|=2$, we have

$$\mathrm{Var}(J_i^*|\mathcal{G}) = \frac{1}{L}\left(\sum_{j:j\neq i}\tilde{A}_{ij}e^{\theta_i^*}e^{\theta_j^*}\right)\Big/\left[\sum_{j:j\neq i}\tilde{A}_{ij}\frac{e^{\theta_i^*}e^{\theta_j^*}}{e^{\theta_i^*}+e^{\theta_j^*}}\right]^2.$$

This exactly matches with proposition 4.2 of Gao et al. (2021). In addition, if we choose $f(A_l)=\sum_{u\in A_l} e^{\theta_u^*}$, we get the most efficient variance, just like the MLE variance given in proposition 4.1 of Gao et al. (2021), which is

$$\mathrm{Var}(J_i^*|\mathcal{G}) = \left(L\cdot\sum_{j:j\neq i}\tilde{A}_{ij}\frac{e^{\theta_i^*}e^{\theta_j^*}}{(e^{\theta_i^*}+e^{\theta_j^*})^2}\right)^{-1}.$$

With the above discussion and computation, after some additional derivations, we come to the conclusion that $\tilde{\theta}_i-\theta_i^*$ has the same asymptotic distribution as $\frac{\hat{\pi}_i-\pi_i^*}{\pi_i^*}$ and J_i^*. Therefore,

$$\mathrm{Var}(J_i^*|\mathcal{G})^{-1/2}(\tilde{\theta}_i-\theta_i^*)\Rightarrow N(0,1),$$

for all $i\in[n]$. Based on this result, we can make an inference for $\tilde{\theta}_i$ and additionally the rank of item i (see Section 3.3). The rigorous derivations for this conclusion will be provided in Section 4.

3.2. Uncertainty Quantification for the PL Model

In this section, we consider the case of a random comparison graph, which could potentially lead to dependent (c_l, A_l), unlike the case of the fixed graph in the previous section. Note that because the random comparison graph generation can be arbitrary, we cannot work with each case. As an illustrating example, we consider the classical PL model from the Erdős-Rényi graph here for two reasons. Firstly, this is the most popularly studied random comparison model in the ranking literature (Chen and Suh 2015; Han et al. 2020; Gao et al. 2021; Chen et al. 2022; Fan et al. 2022a, 2022b; Liu et al. 2022). Secondly, we can use the model to verify the uncertainty quantification of the spectral method and compare it with that of the MLE. It turns out the model is good enough to give us new insights. To further simplify the discussion and presentation, we will only focus on $M=3$ in the PL model. Results for general M can be similarly derived with the same conclusion.

With the PL model, we can write down the specific variance of J_i^*. Consider the most natural way to encode a three-way comparison. Say one ranks (i,j,k) as $i>j>k$ where $a>b$ means a is better than b. Motivated by the likelihood function, which multiplies the probability of selecting i as the best one from $\{i,j,k\}$ and the probability of selecting j next from $\{j,k\}$, we break this complete three-way comparison into two dependent comparison data: $(i,\{i,j,k\})$ and $(j,\{j,k\})$. We call this *multilevel breaking*, where in the first-level comparison of all three items, i is preferred, and in the second-level comparison of the remaining items, j is preferred. By doing this, we can naturally link and compare results with the MLE. Azari Soufiani et al. (2013) also proposed other ways of breaking an M-way comparison into pairwise

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

comparisons, but different breaking methods will lead to different dependent structures, which we do not intend to analyze one by one in this work. So in the sequel, we only consider multilevel breaking, motivated by the likelihood function, and leave the study of other possible breaking methods to the future. We use $\tilde{A}_{ijk}=1$ or 0 to denote whether (i,j,k) has been compared for L times or is never compared.

Let us work on the multilevel breaking. Now the key difference is that the induced comparison graph $\mathcal{G}$ cannot be treated as fixed. Instead, we condition on $\tilde{\mathcal{G}}=\{\tilde{A}_{ijk}\}$. Similar to (3), we have

$$\frac{\overline{\pi}_i-\pi_i^*}{\pi_i^*}=\frac{\sum_{j:j\neq i}P_{ji}\pi_j^*-P_{ij}\pi_i^*}{\pi_i^*\sum_{j:j\neq i}P_{ij}}\approx\frac{\sum_{j:j\neq i}P_{ji}e^{\theta_j^*}-P_{ij}e^{\theta_i^*}}{\sum_{j:j\neq i}E[P_{ij}|\tilde{\mathcal{G}}]e^{\theta_i^*}}=:J_i^*. \tag{7}$$

In the case of the random comparison graph, that is, conditioning on $\tilde{\mathcal{G}}$, we obtain

$$P_{ij}=\frac{1}{d}\sum_{\ell=1}^{L}\sum_{k:k\neq i,j}\tilde{A}_{ijk}Z_{ijk}^{\ell}, \tag{8}$$

where $Z_{ijk}^{\ell}=1(y_{k>j>i}^{(\ell)}=1)/f(\{i,j\})+1(y_{j>i>k}^{(\ell)}=1)/f(\{i,j,k\})+1(y_{j>k>i}^{(\ell)}=1)/f(\{i,j,k\})$. Here $y_{i_1>i_2>i_3}^{(\ell)}$ is a binary variable which equals to one when the event $i_1\succ i_2\succ i_3$ holds under the ℓ-th comparison among items $\{i_1,i_2,i_3\}$. Essentially, we need to collect all terms induced from the same comparison into one term Z_{ijk}^{ℓ} so that the summation is always over independent terms.

We lightly abuse the notation of J_i^*, although here the expectation is conditioning on $\tilde{\mathcal{G}}$ instead of $\mathcal{G}$ used in the fixed graph case. Note that

$$E[Z_{ijk}^{\ell}|\tilde{\mathcal{G}}]=\frac{e^{\theta_k^*}e^{\theta_j^*}}{(e^{\theta_i^*}+e^{\theta_j^*}+e^{\theta_k^*})(e^{\theta_i^*}+e^{\theta_j^*})f(\{i,j\})}+\frac{e^{\theta_j^*}}{(e^{\theta_i^*}+e^{\theta_j^*}+e^{\theta_k^*})f(\{i,j,k\})}.$$

Using $\sum_{j\neq k}a_{ijk}=\sum_{j<k}(a_{ijk}+a_{ikj})$, the denominator of J_i^* can be expressed as

$$\begin{aligned}\tau_i^{\diamond}(\theta^*)&:=\sum_{j:j\neq i}E[P_{ij}|\tilde{\mathcal{G}}]e^{\theta_i^*}\\&=\frac{L}{d}\sum_{j<k:j,k\neq i}\tilde{A}_{ijk}e^{\theta_i^*}\left(\frac{e^{\theta_j^*}e^{\theta_k^*}}{(e^{\theta_i^*}+e^{\theta_j^*}+e^{\theta_k^*})(e^{\theta_i^*}+e^{\theta_j^*})f(\{i,j\})}\right.\\&\quad+\frac{e^{\theta_j^*}e^{\theta_k^*}}{(e^{\theta_i^*}+e^{\theta_j^*}+e^{\theta_k^*})(e^{\theta_i^*}+e^{\theta_k^*})f(\{i,k\})}\\&\quad\left.+\frac{e^{\theta_j^*}+e^{\theta_k^*}}{(e^{\theta_i^*}+e^{\theta_j^*}+e^{\theta_k^*})f(\{i,j,k\})}\right).\end{aligned}$$

Hence, the expression of J_i^* is given as follows:

$$\begin{aligned}J_i^*&=\frac{1}{\tau_i^{\diamond}}\left(\sum_{j:j\neq i}P_{ji}e^{\theta_j^*}-P_{ij}e^{\theta_i^*}\right)\\&=\frac{1}{d\tau_i^{\diamond}}\left(\sum_{\ell=1}^{L}\sum_{j:j\neq i}\sum_{k:k\neq i,j}\tilde{A}_{ijk}(Z_{jik}^{\ell}e^{\theta_j^*}-Z_{ijk}^{\ell}e^{\theta_i^*})\right)\\&=\frac{1}{d\tau_i^{\diamond}}\sum_{\ell=1}^{L}\sum_{j<k:j,k\neq i}\tilde{A}_{ijk}(Z_{jik}^{\ell}e^{\theta_j^*}+Z_{kij}^{\ell}e^{\theta_k^*}-Z_{ijk}^{\ell}e^{\theta_i^*}-Z_{ikj}^{\ell}e^{\theta_i^*})\\&=:\frac{1}{d}\sum_{\ell=1}^{L}\sum_{j<k:j,k\neq i}J_{ijk\ell}(\theta^*),\end{aligned} \tag{9}$$

where $\tau_i^{\diamond}$ is short for $\tau_i^{\diamond}(\theta^*)$. Because each three-way comparison is independent, it can be shown that the variance of J_i^* is

$$\mathrm{Var}(J_i^*|\tilde{\mathcal{G}})=\frac{L}{d^2(\tau_i^{\diamond})^2}\sum_{j<k:j,k\neq i}\tilde{A}_{ijk}e^{\theta_i^*}\left(\frac{(e^{\theta_j^*}+e^{\theta_k^*})}{f^2(\{i,j,k\})}+\frac{e^{\theta_j^*}e^{\theta_k^*}}{e^{\theta_i^*}+e^{\theta_j^*}+e^{\theta_k^*}}\left(\frac{1}{f^2(\{i,k\})}+\frac{1}{f^2(\{i,j\})}\right)\right).$$

The essential component is to compute $EJ_{ijk\ell}(\theta^*)^2$ because of independence and zero-mean. For a given triplet (i,j,k), there are six possible preference outcomes with probabilities governed by the PL model. Averaging the squared random outcomes over six probabilities gives $EJ_{ijk\ell}(\theta^*)^2$, which results in the expression above. We omit the details of these calculations.

Let us consider the simple situation that all θ_i^* are equal. In this case, if we apply the most efficient weighting function $f(A_l)\propto\sum_{u\in A_l}e^{\theta_u^*}$, that is, $f(\{i,j,k\})=3$, $f(\{i,j\})=f(\{i,k\})=2$, we have $\mathrm{Var}(J_i^*|\tilde{\mathcal{G}})=18/(7L)$. However, if we naively choose f as a constant function, we get $\mathrm{Var}(J_i^*|\tilde{\mathcal{G}})=8/(3L)$, which is indeed larger. It is also worth noting that when we choose $f(A_l)\propto\sum_{u\in A_l}e^{\theta_u^*}$, the aforementioned variance matches with the variance of MLE in Fan et al. (2022b).

With the above formula of $\mathrm{Var}(J_i^*|\tilde{\mathcal{G}})$ for the PL model, we can also conclude that

$$\mathrm{Var}(J_i^*|\tilde{\mathcal{G}})^{-1/2}(\tilde{\theta}_i-\theta_i^*)\Rightarrow N(0,1),$$

for all $i\in[n]$. The rigorous arguments will be introduced in Section 4.

3.3. Ranking Inference: One-Sample Confidence Intervals

In numerous practical applications, individuals frequently interact with data and challenges related to rankings. The prevalent approach to utilizing rankings typically revolves around computing preference scores and then showcasing these scores in ranked order.

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

These only provide first-order information on ranks and cannot answer many questions, such as

- How do we ascertain with high confidence that an item's true rank is among the top-3 (or general $K, K \geq 1$) choices? And how can we establish a set of candidates with high confidence, guaranteeing that the true top-3 candidates are not overlooked?
- How do we analyze if the ranking preferences for a given array of products are consistent in two distinct communities (such as male and female) or the same community but at two different time periods?

In sum, there is a need for tools and methodologies that address these and other insightful queries in real-world applications involving rankings, especially when the comparisons are drawn from a general comparison graph.

Within this section, we first present a comprehensive framework designed for the construction of two-sided confidence intervals for ranks. In endeavoring to establish simultaneous confidence intervals for the ranks, an intuitive methodology entails deducing the asymptotic distribution of the empirical ranks, denoted as $\tilde{r}_m, m \in \mathcal{M}$, and subsequently determining the critical value. However, it is well known that this task poses substantive challenges, given that $\tilde{r}_m$ is an integer and is intrinsically dependent on all estimated scores, making its asymptotic behavior daunting to analyze.

By capitalizing on the inherent interdependence between the scores and their corresponding ranks, we discern that the task of formulating confidence intervals for the ranks can be effectively converted to the construction of simultaneous confidence intervals for the pairwise differences among the population scores. It is notable that the distribution of these empirical score differences is more amenable to characterization. Consequently, we focus on the statistical properties of the estimated scores $\tilde{\theta}_m, m \in [n]$ and present our methodology for constructing two-sided (simultaneous) confidence intervals for ranks through estimated score differences.

Example 3. We let $\mathcal{M} = \{m\}$, where $1 \leq m \leq n$, to represent the item under consideration. We are interested in the construction of the $(1-\alpha) \times 100\%$ confidence interval for the true population rank r_m, where $\alpha \in (0,1)$ denotes a prespecified significance level. Suppose that we are able to construct the simultaneous confidence intervals $[\mathcal{C}_L(k,m), \mathcal{C}_U(k,m)], k \neq m, (k \in [n])$ for the pairwise differences $\theta^*_k - \theta^*_m, k \neq m \ \ (k \in [n])$, with the following property:

$$\mathbb{P}(\mathcal{C}_L(k,m) \leq \theta^*_k - \theta^*_m \leq \mathcal{C}_U(k,m) \text{ for all } k \neq m) \geq 1-\alpha. \tag{10}$$

One observes that if $\mathcal{C}_U(k,m) < 0$ (respectively, $\mathcal{C}_L(k,m) > 0$), it implies that $\theta^*_k < \theta^*_m$ (respectively, $\theta^*_k > \theta^*_m$). Enumerating the number of items whose scores are higher than item m gives a lower bound for rank r_m, and vice versa. In other words, we deduce from (10) that

$$\mathbb{P}\left(1 + \sum_{k \neq m} 1\{\mathcal{C}_L(k,m) > 0\} \leq r_m \leq n - \sum_{k \neq m} 1\{\mathcal{C}_U(k,m) < 0\}\right) \geq 1-\alpha. \tag{11}$$

This yields a $(1-\alpha) \times 100\%$ confidence interval for r_m, and our task reduces to constructing simultaneous confidence intervals for the pairwise differences (10).

We now formally introduce the procedure to construct the confidence intervals for multiple ranks $\{r_m\}_{m \in \mathcal{M}}$ simultaneously. Motivated by Example 3, the key step is to construct the simultaneous confidence intervals for the pairwise score differences $\{\theta^*_k - \theta^*_m\}_{m \in \mathcal{M}, k \neq m}$ such that (10) holds. To this end, we let

$$T_{\mathcal{M}} = \max_{m \in \mathcal{M}} \max_{k \neq m} \left| \frac{\tilde{\theta}_k - \tilde{\theta}_m - (\theta^*_k - \theta^*_m)}{\tilde{\sigma}_{km}} \right|, \tag{12}$$

where $\{\tilde{\sigma}_{km}\}_{k \neq m}$ is a sequence of positive normalization given by (14) below. For any $\alpha \in (0,1)$, let $Q_{1-\alpha}$ be a critical value such that $\mathbb{P}(T_{\mathcal{M}} \leq Q_{1-\alpha}) \geq 1-\alpha$. Then, as in Example 3, our $(1-\alpha) \times 100\%$ simultaneous confidence intervals for $\{r_m\}_{m \in \mathcal{M}}$ are given by $\{[R_{mL}, R_{mU}]\}_{m \in \mathcal{M}}$, where

$$R_{mL} = 1 + \sum_{k \neq m} 1(\tilde{\theta}_k - \tilde{\theta}_m > \tilde{\sigma}_{km} \times Q_{1-\alpha}),$$
$$R_{mU} = n - \sum_{k \neq m} 1(\tilde{\theta}_k - \tilde{\theta}_m < -\tilde{\sigma}_{km} \times Q_{1-\alpha}).$$

3.4. Multiplier Bootstrap Procedure

The key step for constructing the confidence interval of ranks of interest is to pick the critical value $Q_{1-\alpha}$. To calculate the critical value above, we propose to use the wild bootstrap procedure. The uncertainty quantification for the spectral estimator in (3) reveals that $\tilde{\theta}_i - \theta^*_i \approx J_i(\theta^*)$ uniformly over $i \in [n]$ (see details in Section 4), which further implies that asymptotically

$$T_{\mathcal{M}} \approx \max_{m \in \mathcal{M}} \max_{k \neq m} \left| \frac{J_k(\theta^*) - J_m(\theta^*)}{\tilde{\sigma}_{km}} \right|. \tag{13}$$

We focus on the fixed graph setting and leave the random graph setting to Remark 3 below. Practically, the empirical version of $J_i(\theta^*)$ can be obtained via plugging in the spectral estimator $\tilde{\theta}$, namely, from (5),

$$J_i(\tilde{\theta}) = \frac{1}{d} \sum_{l \in \mathcal{D}} J_{il}(\tilde{\theta}), \ i \in [n].$$

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

Let $\sigma_{km}^2 = \mathrm{Var}\{J_k(\theta^*) - J_m(\theta^*)|\mathcal{G}\}$ for each $k \neq m$. Then our estimator for σ_{km}^2 is defined by

$$\tilde{\sigma}_{km}^2 = \frac{e^{\tilde{\theta}_k}}{d^2\tau_k^2(\tilde{\theta})}\sum_{l\in\mathcal{D}}\frac{1(k\in A_l)}{f^2(A_l)}\left(\sum_{j\in A_l}e^{\tilde{\theta}_j} - e^{\tilde{\theta}_k}\right) + \frac{e^{\tilde{\theta}_m}}{d^2\tau_m^2(\tilde{\theta})}\sum_{l\in\mathcal{D}}\frac{1(m\in A_l)}{f^2(A_l)}\left(\sum_{j\in A_l}e^{\tilde{\theta}_j} - e^{\tilde{\theta}_m}\right), \quad (14)$$

where $\tau_k(\tilde{\theta})$ and $\tau_m(\tilde{\theta})$ also plug in $\tilde{\theta}$; see (6). Let $\omega_1,\ldots,\omega_{|\mathcal{D}|} \in \mathbb{R}$ be i.i.d. $N(0,1)$ random variables. The Gaussian multiplier bootstrap statistic is then defined by

$$G_{\mathcal{M}} = \max_{m\in\mathcal{M}}\max_{k\neq m}\left|\frac{1}{d\tilde{\sigma}_{km}}\sum_{l\in\mathcal{D}}\{J_{kl}(\tilde{\theta}) - J_{ml}(\tilde{\theta})\}\omega_l\right|. \quad (15)$$

Let $\mathbb{P}^*(\cdot) = \mathbb{P}(\cdot|\{(c_l, A_l)\}_{l\in\mathcal{D}})$ denote the conditional probability. Then, for $\alpha \in (0,1)$, our estimator for $Q_{1-\alpha}$ is defined by the $(1-\alpha)$-th conditional quantile of $G_{\mathcal{M}}$, namely,

$$\mathcal{Q}_{1-\alpha} = \inf\{z : \mathbb{P}^*(G_{\mathcal{M}} \le z) \ge 1-\alpha\},$$

which can be computed by the Monte Carlo simulation. Then, our simultaneous confidence intervals $\{[\mathcal{R}_{mL}, \mathcal{R}_{mU}]\}_{m\in\mathcal{M}}$ are given by

$$\mathcal{R}_{mL} = 1 + \sum_{k\neq m}1(\tilde{\theta}_k - \tilde{\theta}_m > \tilde{\sigma}_{km}\times\mathcal{Q}_{1-\alpha}),$$
$$\mathcal{R}_{mU} = n - \sum_{k\neq m}1(\tilde{\theta}_k - \tilde{\theta}_m < -\tilde{\sigma}_{km}\times\mathcal{Q}_{1-\alpha}). \quad (16)$$

Remark 2 (One-Sample One-Sided Confidence Intervals). Now we provide details on constructing simultaneous one-sided intervals for population ranks. For one-sided intervals, the overall procedure is similar to constructing two-sided confidence intervals. Specifically, let

$$G_{\mathcal{M}}^{\circ} = \max_{m\in\mathcal{M}}\max_{k\neq m}\frac{1}{d\tilde{\sigma}_{km}}\sum_{l\in\mathcal{D}}\{J_{kl}(\tilde{\theta}) - J_{ml}(\tilde{\theta})\}\omega_l, \quad (17)$$

where $\omega_1,\ldots,\omega_{|\mathcal{D}|}$ are as before i.i.d. $N(0,1)$ random variables. Correspondingly, let $\mathcal{Q}_{1-\alpha}^{\circ}$ be its $(1-\alpha)$-th quantile. Then the $(1-\alpha)\times 100\%$ simultaneous lower confidence bounds for $\{r_m\}_{m\in\mathcal{M}}$ are given by $\{[\mathcal{R}_{mL}^{\circ}, n]\}_{m\in\mathcal{M}}$, where

$$\mathcal{R}_{mL}^{\circ} = 1 + \sum_{k\neq m}1(\tilde{\theta}_k - \tilde{\theta}_m > \tilde{\sigma}_{km}\times\mathcal{Q}_{1-\alpha}^{\circ}). \quad (18)$$

Remark 3 (Ranking Inference for the PL Model with Random Comparison Graph). Section 3.2 reveals that $\tilde{\theta}_i - \theta_i^* \approx J_i(\theta^*)$ uniformly over $i \in [n]$, where following (9),

$$J_i(\theta^*) = \frac{1}{d}\sum_{\ell=1}^{L}\sum_{j<s:j,s\neq i}J_{ijs\ell}(\theta^*).$$

In order to carry out ranking inference for the PL model, we need to rewrite this equation in a slightly different format. Let $\mathcal{N} = \sum_{i<j<k}\tilde{A}_{ijk}$ denote the total number of connected components on the random graph $\tilde{\mathcal{G}}$ and write $\{(i,j,k) : i<j<k \text{ and } \tilde{A}_{ijk}=1\} =: \{\tilde{A}_q\}_{q=1,\ldots,\mathcal{N}}$. Let $y_q^{(\ell)}$ denote the ℓ-th full-ranking comparison result for $\tilde{A}_q$. Then we can rewrite P_{ij} as

$$P_{ij} = \frac{1}{d}\sum_{\ell=1}^{L}\sum_{q=1}^{\mathcal{N}}\sum_{k:k\neq i,j}1\{(i,j,k) = \tilde{A}_q\}Z_{ijkq}^{(\ell)}, i\neq j,$$

where for $i\neq j\neq k$ and $q\in[\mathcal{N}]$, and $Z_{ijkq}^{(\ell)} = 1\{y_q^{(\ell)} = (k>j>i)\}/f(\{i,j\}) + (1\{y_q^{(\ell)} = (j>i>k)\} + 1\{y_q^{(\ell)} = (j>k>i)\})/f(\{i,j,k\})$. It is straightforward to verify that this P_{ij} is exactly the same with (8). Therefore, we rewrite $J_i(\theta^*) = d^{-1}\sum_{\ell=1}^{L}\sum_{q=1}^{\mathcal{N}}J_{iq\ell}^{\diamond}(\theta^*)$, where

$$J_{iq\ell}^{\diamond}(\theta^*) = \sum_{j<s:j,s\neq i}1\{(i,j,s) = \tilde{A}_q\}J_{ijs\ell}(\theta^*). \quad (19)$$

As is assumed, $\{J_{iq\ell}^{\diamond}(\theta^*)\}_{\ell\in[L],q\in[\mathcal{N}]}$ are independent for each $i\in[n]$ conditioning on the comparison graph $\tilde{\mathcal{G}}$. Let $\{\omega_{q\ell}\}_{q,\ell\in\mathbb{N}}$ be i.i.d. $N(0,1)$ random variables. Then, following (15), the corresponding bootstrap test statistic is given by

$$G_{\mathcal{M}}^{\diamond} = \max_{m\in\mathcal{M}}\max_{k\neq m}\left|\frac{1}{d\tilde{\sigma}_{km}^{\diamond}}\sum_{\ell=1}^{L}\sum_{q=1}^{\mathcal{N}}\{J_{kq\ell}^{\diamond}(\tilde{\theta}) - J_{mq\ell}^{\diamond}(\tilde{\theta})\}\omega_{q\ell}\right|,$$

where $\{\tilde{\sigma}_{km}^{\diamond}\}_{k\neq m}$ are as before the sequence of positive normalization, calculated as the sum of the variance of $J_k(\theta^*)$ and $J_m(\theta^*)$ similar to (14). Consequently, the simultaneous confidence intervals for the ranks can be similarly constructed.

3.5. Ranking Inference: Two-Sample and One-Sample Testing Applications

In this section, we further illustrate how we may apply our inference methodology to a few salient testing applications, in both one-sample and two-sample testing.

Example 4 (Testing Top-K Placement). Let $\mathcal{M} = \{m\}$ for some $m\in[n]$ and let $K\ge 1$ be a prescribed positive integer. Our objective is to ascertain if the item m is a member of the top-K ranked items. Consequently, we shall examine the following hypotheses:

$$H_0 : r_m \le K \text{ versus } H_1 : r_m > K. \quad (20)$$

Based on the one-sided confidence interval $[\mathcal{R}_{mL}^{\circ}, n]$ in (18), for any $\alpha\in(0,1)$, a level α test for (20) is simply given by $\phi_{m,K} = 1\{\mathcal{R}_{mL}^{\circ} > K\}$. Under the conditions of Theorem EC.1 in the Online Appendix, we have $\mathbb{P}(\phi_{m,K} = 1|H_0) \le \alpha + o(1)$; that is, the effective control of the Type I error can be achieved below the significant level α when the null hypothesis is true.

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

Example 5 (Top-K Sure Screening Set). Another example is on constructing a screened candidate set that contains the top-K items with high probability. This is particularly useful in college candidate admission or company hiring decisions. Oftentimes, a university or a company would like to design a certain admission or hiring policy with the high-probability guarantee of the sure screening of true top-K candidates.

Let $\mathcal{K} = \{r^{-1}(1), \ldots, r^{-1}(K)\}$ denote the top-K ranked items of the rank operator $r : [n] \to [n]$. We aim at selecting a set of candidates $\hat{\mathcal{I}}_K$ which contains the top-K candidates with a prescribed probability. Mathematically, this requirement can be expressed as $\mathbb{P}(\mathcal{K} \subseteq \hat{\mathcal{I}}_K) \geq 1 - \alpha$, where $\alpha \in (0,1)$. Herein, we define $\mathcal{M} = [n]$, and let $\{[\mathcal{R}^{\circ}_{mL}, n], m \in [n]\}$ represent the set of $(1-\alpha)\times 100\%$ simultaneous left-sided confidence intervals, as given in (18). It is easy to observe that the inequality $\mathcal{R}^{\circ}_{mL} > K$ infers that $r_m > K$. Consequently, a selection for $\hat{\mathcal{I}}_K$, which satisfies the probability constraint $\mathbb{P}(\mathcal{K} \subseteq \hat{\mathcal{I}}_K) \geq 1 - \alpha$, is given by

$$\hat{\mathcal{I}}_K = \{m \in [n] : \mathcal{R}^{\circ}_{mL} \leq K\}.$$

Example 6 (Testing Ranks of Two Samples). In many applications, we are concerned with the question of whether the ranks of certain items using two samples have been changed or preserved. For example, we may care about whether

- Ranking allocation differs before and after a treatment or policy change.
- Different communities such as males versus females can have different ranking preferences over the same set of products.
- People's perceived preferences over the same things have changed in two time periods.

Suppose we observe two independent data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ with preference scores $\theta^*_{[1]} = (\theta^*_{11}, \ldots, \theta^*_{1n})^\top$ and $\theta^*_{[2]} = (\theta^*_{21}, \ldots, \theta^*_{2n})^\top$. The associated true rankings are respectively denoted by

$$r_{[1]} = (r_{11}, \ldots, r_{1n})^\top \text{ and } r_{[2]} = (r_{21}, \ldots, r_{2n})^\top.$$

Given any $m \in [n]$, we are interested in testing whether the same rank is preserved for item m across these two samples, that is, testing the hypotheses

$$H_0 : r_{1m} = r_{2m} \text{ versus } H_1 : r_{1m} \neq r_{2m}. \tag{21}$$

To this end, firstly we construct simultaneous confidence intervals $[R_{1mL}, R_{1mU}]$ and $[R_{2mL}, R_{2mU}]$ such that

$$\mathbb{P}(r_{1m} \in [R_{1mL}, R_{1mU}] \text{ and } r_{2m} \in [R_{2mL}, R_{2mU}]) \geq 1 - \alpha. \tag{22}$$

Then our α-level test for (21) is defined by

$$\phi_m = 1\{|[R_{1mL}, R_{1mU}] \cap [R_{2mL}, R_{2mU}]| = 0\}.$$

It is straightforward to verify that $\mathbb{P}(\phi_m = 1 | H_0) \geq 1 - \alpha$.

Example 7 (Testing Top-K Sets of Two Samples). Besides testing for a single or a few ranks, one may want to evaluate whether two top-K sets are identical or not, between two groups of people, two periods of time, or before and after a significant event or change. Let $\mathcal{S}_{1K} = \{r^{-1}_{[1]}(1), \ldots, r^{-1}_{[1]}(K)\}$ and $\mathcal{S}_{2K} = \{r^{-1}_{[2]}(1), \ldots, r^{-1}_{[2]}(K)\}$ denote the sets of top-K ranked items, respectively. We consider testing the hypotheses

$$H_0 : \mathcal{S}_{1K} = \mathcal{S}_{2K} \text{ versus } H_1 : \mathcal{S}_{1K} \neq \mathcal{S}_{2K}. \tag{23}$$

For $\alpha \in (0,1)$, we begin with constructing $(1-\alpha)\times 100\%$ simultaneous confidence sets $\hat{\mathcal{I}}_{1K}$ and $\hat{\mathcal{I}}_{2K}$ for $\mathcal{S}_{1K}$ and $\mathcal{S}_{2K}$ such that

$$\mathbb{P}(\mathcal{S}_{1K} \subset \hat{\mathcal{I}}_{1K} \text{ and } \mathcal{S}_{2K} \subset \hat{\mathcal{I}}_{2K}) \geq 1 - \alpha. \tag{24}$$

Then our α-level test for (23) is defined by

$$\tilde{\phi}_K = 1\{|\hat{\mathcal{I}}_{1K} \cap \hat{\mathcal{I}}_{2K}| < K\}.$$

Remark 4. Several methodologies, including the Bonferroni adjustment (which constructs a $(1-\alpha/2)\times 100\%$ confidence interval for each source) and Gaussian approximation (achieved by taking the maximum of the test statistics of each source), enable us to establish simultaneous confidence intervals as illustrated in Equations (22) and (24). To maintain clarity and simplicity in the subsequent context, we simply employ the Bonferroni adjustment for two samples. Moreover, the framework outlined in Examples 6 and 7 can be extended in a straightforward way to evaluate whether the ranks of items or sets are identical across three or more sources.

4. Theoretical Justifications

In this section, we rigorously justify the conclusions in Section 3 and explicitly lay out the necessary assumptions to arrive at those conclusions. The first assumption is to make sure we are comparing θ^*_i's in the same order in a meaningful way. Otherwise, we can always group items into categories with similar qualities and then work on each subgroup or screen some extreme items. In addition, as we have discussed, we need an identifiability condition for θ^*.

Assumption 1. *There exists some positive constant* $\overline{\kappa} < \infty$ *such that*

$$\max_{i \in [n]} \theta^*_i - \min_{i \in [n]} \theta^*_i \leq \overline{\kappa}.$$

In addition, for identifiability, assume $1^\top \theta^* = 0$.

In Assumption 1, we assume $\overline{\kappa}$ is finite, indicating we only rank items with preference scores on the same scale. If $\overline{\kappa}$ is diverging, some items will be trivially more or less favorable than others. In this case, it is typically easy in practice to separate the items into subgroups with similar preference scores, and then we can

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

conduct ranking inference within each group. Although we assume bounded $\overline{\kappa}$, it serves in the role of a condition number whose effect has been made explicit in all our results for interested readers. However, we do not claim this dependency is optimal as our nontrivial analysis can easily encounter powers of $e^{\overline{\kappa}}$, say, in bounding the ratio of π_i^*/π_j^*.

4.1. Estimation Accuracy and Asymptotic Normality with Fixed Comparisons

To derive the asymptotic distribution of the spectral estimator, we need to rigorously justify the approximations (2) and (3). We first take care of approximating (2) using (3), where all comparisons in $\mathcal{G}$ are assumed to be fixed. Note that

$$P_{ij} - E[P_{ij}|\mathcal{G}] = \frac{1}{d}\sum_{l\in\mathcal{D}} 1(i,j\in A_l)\left[1(c_l=j) - \frac{\pi_j^*}{\sum_{u\in A_l}\pi_u^*}\right]\frac{1}{f(A_l)}.$$

Let $Z^j_{A_l} = 1(c_l = j)/f(A_l)$, which is bounded from above and below as long as f is bounded from above and below. Furthermore, each $Z^j_{A_l}$ is independent. Therefore, $P_{ij} - E[P_{ij}|\mathcal{G}] = d^{-1}\sum_{l\in\mathcal{D}} 1(i,j\in A_l)[Z^j_{A_l} - E(Z^j_{A_l})]$. By Hoeffding's inequality, conditioning on $\mathcal{G}$, we have, with a large probability $1 - o(1)$,

$$\max_{i\neq j} |P_{ij} - E[P_{ij}|\mathcal{G}]| \lesssim \frac{1}{d}\sqrt{(\log n)n^{\ddagger}},$$

where $n^{\ddagger} = \max_{i\neq j}\sum_{l\in\mathcal{D}} 1(i,j\in A_l)$ is the maximum number of times that each pair is compared. Similarly, we can get the concentration bound for $\sum_{j:j\neq i} P_{ij}$. Because $Z^j_{A_l}$'s are independent, another level of summation over j will lead to the following. Again, by Hoeffding's inequality, with a large probability tending to one, we obtain

$$\max_i \left|\sum_{j:j\neq i} P_{ij} - \sum_{j:j\neq i} E[P_{ij}|\mathcal{G}]\right| \lesssim \frac{1}{d}\sqrt{(\log n)n^{\dagger}},$$

where $n^{\dagger} = \max_i \sum_{l\in\mathcal{D}} 1(i\in A_l)$ is the maximum number of times that each item is compared. In addition, we assume

$$\sum_{j:j\neq i} E[P_{ij}|\mathcal{G}] = \tau_i e^{-\theta_i^*} \asymp \frac{1}{d}n^{\dagger},$$

where τ_i defined in (4) is the denominator of J_i^*. This assumption makes sense as $\tau_i e^{-\theta_i^*} \lesssim \sum_{l\in\mathcal{D}} 1(i\in A_l)/d$, and it states for each i the comparison graph cannot be too asymmetric. Note that $\sum_{l\in\mathcal{D}} 1(i,j\in A_l)$ can still be widely different from $n^{\ddagger}$ for a different pair (i,j). Because the expectation term dominates the deviation if $n^{\dagger} \gtrsim \log n$, it is not hard to show that in (3), changing the denominator by its expectation will only cause a small-order difference, which does not affect the asymptotic distribution.

Based on the above discussion, we impose the following assumption.

Assumption 2. *In the case of a fixed comparison graph, we assume the graph is connected,* $\tau_i e^{-\theta_i^*} \asymp n^{\dagger}/d$ *for all* $i\in[n]$, $e^{2\overline{\kappa}}\log n = o(n)$, *and* $e^{3\overline{\kappa}} n^{\ddagger} n^{1/2}(\log n)^{1/2} = o(n^{\dagger})$.

The assumption is reasonable for a fixed comparison graph. If each pair (i,j) must be compared at least once, then every $\sum_{l\in\mathcal{D}} 1(i,j\in A_l) \geq 1$. If they are all in the same order, then $\sum_{l\in\mathcal{D}} 1(i\in A_l) = \sum_{j:j\neq i}\sum_{l\in\mathcal{D}} 1(i,j\in A_l)$ should be indeed in the order of $n^{\ddagger}n$. Assumption 2 allows some pair (i,j) to be never compared directly, so we need to leverage the information from comparing i and j with other items separately. Moreover, we also do not require $\sum_{l\in\mathcal{D}} 1(i,j\in A_l)$ to be in the same order for any $i,j: i\neq j$ because we only require the maximum pairwise degree $n^{\ddagger}$ to satisfy Assumption 2. However, in the case of a fixed graph, we do not have the randomness from the graph, and the graph must be relatively dense to make sure we have enough information to rank every item. This condition will be relaxed to $n^{\dagger} \gtrsim n^{\ddagger}\log n$ when we have a homogeneous random comparison graph in Section 4.2.

We need another technical condition on the structure of the comparison graph. Define $\Omega = \{\Omega_{ij}\}_{i\leq n, j\leq n}$ where $\Omega_{ij} = -P_{ji}\pi_j^*$ for $i\neq j$ and $\Omega_{ii} = \sum_{j:j\neq i} P_{ij}\pi_i^*$. Note that as we derived above, $E[\Omega_{ii}|\mathcal{G}]$ is in the order of $n^{\dagger}/(dn)$. We hope to understand the order of its eigenvalues. Because Ω has the minimal eigenvalue equal to zero, with the corresponding eigenvector $\mathbf{1}$, we only focus on the space orthogonal to $\mathbf{1}$. Following the notation of Gao et al. (2021),

$$\lambda_{\min,\perp}(A) = \min_{\|v\|=1, v^{\top}\mathbf{1}=0} v^{\top}Av.$$

Assumption 3. *There exist* $C_1, C_2 > 0$ *such that*

$$C_1 e^{-\overline{\kappa}}\frac{n^{\dagger}}{dn} \leq \lambda_{\min,\perp}(E[\Omega|\mathcal{G}]) \leq \lambda_{\max}(E[\Omega|\mathcal{G}]) \leq C_2 e^{\overline{\kappa}}\frac{n^{\dagger}}{dn}, \tag{25}$$

$$\|\Omega - E[\Omega|\mathcal{G}]\| = o_P\left(\frac{n^{\dagger}}{dn}\right). \tag{26}$$

When $\overline{\kappa} = O(1)$, Assumption 3 requires that all eigenvalues (except the minimal one) of $E[\Omega|\mathcal{G}]$ are in the order of $n^{\dagger}/(dn)$ and Ω also shares this same eigenvalue scale as $E[\Omega|\mathcal{G}]$. This assumption is intuitively correct, as we have seen that $E[\Omega_{ij}] \lesssim n^{\ddagger}/(dn)$ for $i\neq j$ and $E[\Omega_{ii}] \asymp n^{\dagger}/(dn)$. We will also rigorously show that this condition can be satisfied if we consider the PL model (Theorem 3).

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

Theorem 1. *Under Assumptions 1–3, the spectral estimator $\tilde{\theta}_i$ has the following uniform approximation: $\tilde{\theta}_i - \theta_i^* = J_i^* + \delta_i$, uniformly for all $i \in [n]$, where $\|\delta := (\delta_1, \ldots, \delta_n)\|_\infty = o(1/\sqrt{n^\dagger})$ with probability $1 - o(1)$.*

To prove Theorem 1, we need to verify (2). We leave the detailed proof to the Online Appendix. Given Theorem 1, we can easily conclude the next theorem following the properties of J_i^*, which lead to the rate of convergence for $\tilde{\theta}$ as well as its asymptotic normality.

Remark 5. The results of Theorem 1 and the following theorems are proved via Bernstein and Hoeffding type inequalities with union bound over n items. Therefore, all of the high-probability terms that hold with probability $(1 - o(1))$ (similarly for $o_p(\cdot)$ and $O_p(\cdot)$) mentioned in the main text equivalently hold with probability in the form of $1 - \mathcal{O}(n^{-\zeta})$ where $\zeta \geq 2$ is a positive integer (different choice of ζ will only affect constant terms in the involved concentration inequalities).

Theorem 2. *Under Assumptions 1–3, the spectral estimator (1) satisfies that*

$$\|\tilde{\theta} - \theta^*\|_\infty \asymp \|J^*\|_\infty \lesssim e^{\overline{\kappa}} \sqrt{\frac{\log n}{n^\dagger}}, \tag{27}$$

with probability $1 - o(1)$, where $J^* = (J_1^*, \ldots, J_n^*)$ with $J_i^*, i \in [n]$ being defined in (5). In addition,

$$\rho_i(\theta)(\tilde{\theta}_i - \theta_i^*) \Rightarrow N(0,1),$$

for all $i \in [n]$ with

$$\rho_i(\theta) = \left[\sum_{l \in \mathcal{D}} 1(i \in A_l) \left(\frac{\sum_{u \in A_l} e^{\theta_u} - e^{\theta_i}}{\sum_{u \in A_l} e^{\theta_u}}\right) \frac{e^{\theta_i}}{f(A_l)}\right] \Bigg/ \left[\sum_{l \in \mathcal{D}} 1(i \in A_l) \left(\frac{\sum_{u \in A_l} e^{\theta_u} - e^{\theta_i}}{f(A_l)}\right) \frac{e^{\theta_i}}{f(A_l)}\right]^{1/2}, \tag{28}$$

for both $\theta = \theta^*$ and $\theta =$ any consistent estimator of θ^*.

Note that Theorem 2 indicates that the choice of $f(\cdot) > 0$ does not affect the rate of convergence, but it affects the estimation efficiency. As we argued in Section 3.1, the optimal weighting to minimize the asymptotic variance is $f(A_l) \propto \sum_{u \in A_l} e^{\theta_u^*}$ in the class of spectral estimators. In practice, however, we do not know θ_u^* beforehand. Therefore, we could implement a two-step procedure to improve the efficiency of the spectral estimator: in the first step, we obtain our initial consistent estimator $\hat{\theta}_u^{(\text{initial})}$ with weighting, say, $f(A_l) = |A_l|$, and in the second step, we estimate $f(A_l) = \sum_{u \in A_l} e^{\theta_u^*}$ by plugging in $\hat{\theta}_u^{(\text{initial})}$ and run the spectral method again with this optimal weighting to get the final asymptotically efficient estimator $\hat{\theta}_u^{(\text{final})}$. Note that we do not intend to prove the theoretical properties of this two-step estimator, as the data dependency in the optional weighting of the second step makes the uniform approximation analysis highly nontrivial because of non-i.i.d. ranking outcomes. Nonetheless, we could circumvent this theoretical difficulty by splitting data into a very small part ($o(|\mathcal{D}|)$ samples) for step 1, to achieve consistency with a worse convergence rate, and using the remaining majority ($|\mathcal{D}| - o(|\mathcal{D}|)$ samples for step 2, to maintain the same asymptotic behavior. In addition, empirically, we found that directly using the same whole data in both steps achieves decent performance given a large sample size. We refer interested readers to our numerical studies.

4.2. Estimation Accuracy and Asymptotic Normality for the PL Model

In the random graph case, we have to specify the graph generation process in order to study the theoretical properties. We consider the commonly used PL model, where we sample each M-way comparison with probability p and compare this set for L times. Furthermore, we will only work with $M = 3$ because we plan to focus on a transparent and intuitive discussion. We can easily generalize all the discussions to general M, but derivations and formulas can be more tedious.

The PL model with three-way comparisons has been studied in Fan et al. (2022b) by using MLE, where they explicitly write down the likelihood function. The proposed spectral method can work for any fixed graph, including the one generated from the PL model. In this section, we would like to compare the performance of the spectral method with that of the MLE. To make sure the spectral method works for the PL model, we need to prove the approximations (2) and (7).

We first take care of (7). Consider conditioning on $\tilde{\mathcal{G}}$, where all comparisons in $\tilde{\mathcal{G}}$ are independent; each $\tilde{A}_{ijk}$ is compared for L times if $\tilde{A}_{ijk} = 1$. Now c_l and A_l are induced from $\tilde{\mathcal{G}}$, and can be dependent. In this case, we can write

$$P_{ij} - E[P_{ij}|\tilde{\mathcal{G}}] = \frac{1}{d} \sum_{\ell=1}^{L} \sum_{k:k \neq j,i} \tilde{A}_{ijk}[Z_{ijk}^l - EZ_{ijk}^l],$$

where $Z_{ijk}^l = 1(A_l = \{i,j\}, c_l = j)/f(\{i,j\}) + 1(A_l = \{i,j,k\}, c_l = j)/f(\{i,j,k\})$, which is again bounded from above and below and independent for any given $\tilde{A}_{ijk}$. In this case, with a little abuse of notations, we redefine

$$n^\ddagger = L \max_{i \neq j} \sum_{k:k \neq j,i} \tilde{A}_{ijk}, \quad n^\dagger = L \max_i \sum_{j<k:j,k \neq i} \tilde{A}_{ijk}.$$

Similar to Section 4.1, conditional on $\tilde{\mathcal{G}}$, we have

$$\max_i \left| \sum_{j:j \neq i} P_{ij} - \sum_{j:j \neq i} E[P_{ij}|\tilde{\mathcal{G}}] \right| = \mathcal{O}_P(d^{-1}\sqrt{n^\dagger \log n}),$$

$$\max_{i \neq j} |P_{ij} - E[P_{ij}|\tilde{\mathcal{G}}]| = \mathcal{O}_P\left(d^{-1}\sqrt{n^\ddagger \log n}\right),$$

$$\sum_{j:j \neq i} E[P_{ij}|\tilde{\mathcal{G}}] = \tau_i^\diamond e^{-\theta_i^*} \asymp \frac{1}{d} n^\dagger \ (\text{assumption}).$$

We adapt Assumption 2 to the following assumption. Note that we have no assumption on L, so L can be as low as one.

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

Assumption 4. *In the PL model with M-way complete comparisons, choose* $d \asymp n^\dagger$ *in the spectral ranking, and assume* $\tau_i^\diamond e^{-\theta_i^*} \asymp n^\dagger/d$ *for all* $i \in [n]$, $e^{4\overline{\kappa}} = o(n)$, *and* $p \gtrsim e^{6\overline{\kappa}} \mathrm{poly}(\log n)/\binom{n-1}{M-1}$.

Under Assumption 4, we can prove

$$n^\dagger \asymp \binom{n-1}{M-1} pL, \qquad \max\left\{\binom{n-2}{M-2}p - \log n, 0\right\} L \lesssim n^\ddagger \lesssim \left[\binom{n-2}{M-2}p + \log n\right] L,$$

with probability $1 - o(1)$. Note that in $n^\dagger$, by Assumption 4, we know the dominating term is $\binom{n-1}{M-1}pL$. However, in $n^\ddagger$, we have the additional term $\log n$, which comes from the subexponential tail decay in Bernstein inequality, and if p is really small, it could happen that $\log n$ dominates $n^\ddagger$. When p is large, that is $\binom{n-2}{M-2}p \gtrsim \log n$, then $n^\dagger \asymp nn^\ddagger$ and Assumption 2 holds. Therefore, we have a dense comparison graph, and the proof for this part follows in a similar vein as Theorem 2. When p is small, that is, $\binom{n-2}{M-2}p \lesssim \log n$, $n^\ddagger \lesssim \log n$ if L is bounded. In this case, we will modify the proof of Theorem 2 to the random graph case in order to show Theorem 4 below. In addition, because $\sum_{j:j\neq i} P_{ij} = \mathcal{O}_P(n^\dagger/d)$, it makes sense to choose $d \asymp n^\dagger$ in Assumption 4 to make the diagonal elements of the transition matrix a constant order. Note that in the fixed graph case, we do not need to impose rate assumptions on d as the comparison graph has no randomness.

Next, we verify that under the PL model, Assumption 3 holds with high probability.

Theorem 3. *Under the PL model and Assumption 4, with probability* $1 - o(1)$, *Assumption 3 holds when we condition on* $\tilde{\mathcal{G}}$ *instead of* $\mathcal{G}$.

We next hope to show that under Assumption 4, the spectral estimator $\tilde{\theta}_i$ has the uniform approximation: the differences between $\tilde{\theta}_i - \theta_i^*$ and J_i^* for all $i \in [n]$ are $o_P(1/\sqrt{n^\dagger})$. The key step is still the verification of (2) under this weaker Assumption 4 for a random comparison graph.

Theorem 4. *Under the PL model and Assumptions 1 and 4, the spectral estimator* $\tilde{\theta}_i$ *has the uniform approximation* $\tilde{\theta}_i - \theta_i^* = J_i^* + o_P(1/\sqrt{n^\dagger})$, *uniformly for all* $i \in [n]$. *Therefore, the spectral estimator (1) satisfies*

$$\|\tilde{\theta} - \theta^*\|_\infty \lesssim e^{\overline{\kappa}} \sqrt{\frac{\log n}{\binom{n-1}{M-1}pL}}, \tag{29}$$

with probability $1 - o(1)$. *In addition,*

$$\rho_i(\theta)(\tilde{\theta}_i - \theta_i^*) \Rightarrow N(0,1),$$

for all $i \in [n]$ *with* $\rho_i(\theta) = \mathrm{Var}(J_i^* | \tilde{\mathcal{G}})^{-1/2}$, *where in the formula of* $\mathrm{Var}(J_i^* | \tilde{\mathcal{G}})$ *we can choose both* $\theta = \theta^*$ *and* $\theta =$ *any consistent estimator of* θ^*.

Remark 6. The two-step estimator under optimal weight $f(A_l) = \sum_{u\in A_l} e^{\theta_u^*}$ (can be consistently estimated with a small proportion of a separate data set) achieves the same variance as the MLE, which matches the Crámer Rao lower bound among all estimators (Fan et al. 2022a, 2022b).

Corollary 1. *Under the conditions of Theorem 4, if we have* $\theta_{(K)}^* - \theta_{(K+1)}^* \geq \Delta$, *with* $\theta_{(i)}^*$ *denoting the underlying score of the item with true rank i for* $i \in [n]$, *and when the sample complexity satisfies*

$$e^{2\overline{\kappa}} \Delta^{-2} \cdot \log n = \mathcal{O}\left(\binom{n-1}{M-1} pL\right),$$

we have $\{i \in [n], \hat{r}_i \leq K\} = \{i \in [n], r_i^* \leq K\}$ *(the selected top-K set is identical to the true top-K set) with probability* $1 - o_p(1)$, *where* $\hat{r}_i, r_i^*$ *denote the empirical rank of* $\hat{\theta}_i$ *among* $\{\hat{\theta}_i, i \in [n]\}$ *and true rank of the i-th item, respectively.*

We remark that when $M = 2$, and $\overline{\kappa} = \mathcal{O}(1)$, our conclusion from Corollary 1 reduces to the conclusion of theorem 1 in Chen et al. (2019).

4.3. Validity Justification for Bootstrap Procedure

The primary goal of this section is to justify the validity of the proposed bootstrap procedure in Section 3.4. Recall that the targeted quantity $T_{\mathcal{M}}$ is the maximum modulus of the random vector

$$\Delta_{\mathcal{M}} := \left\{\frac{\tilde{\theta}_k - \tilde{\theta}_m - (\theta_k^* - \theta_m^*)}{\tilde{\sigma}_{km}}\right\}_{m\in\mathcal{M}, k\neq m}.$$

For each marginal of $\Delta_{\mathcal{M}}$, the asymptotic normality can be similarly established following Theorem 2. However, studying the asymptotic distribution of $\|\Delta_{\mathcal{M}}\|_\infty$ becomes quite challenging as its dimension $(n-1)|\mathcal{M}|$ can increase with the number of items. In particular, the traditional multivariate central limit theorem for $\Delta_{\mathcal{M}}$ may no longer be valid asymptotically (Portnoy 1986). To handle the high dimensionality, we shall invoke the modern Gaussian approximation theory (Chernozhukov et al. 2017, 2019) in order to derive the asymptotic distribution of $\|\Delta_{\mathcal{M}}\|_\infty$, shown in Theorem EC.1 in the Online Appendix. Moreover, the validity of our multiplier bootstrap procedure is justified in the following theorem:

Theorem 5. *Assume* $e^{3\overline{\kappa}}(\log n)^2 = o(n)$ *and* $e^{5\overline{\kappa}} n^\ddagger n^{1/2} (\log n)^3 = o(n^\dagger)$. *Then, under the conditions of Theorem 1, we have*

$$|\mathbb{P}(T_{\mathcal{M}} > \mathcal{Q}_{1-\alpha}) - \alpha| \to 0.$$

Remark 7. Theorem 5 indicates that the estimated critical value $\mathcal{Q}_{1-\alpha}$ from the Gaussian multiplier

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

bootstrap indeed controls the significance level of the simultaneous confidence intervals (16) for $\{r_m\}_{m\in\mathcal{M}}$ to the prespecified level α, that is,

$$\mathbb{P}(r_m \in [\mathcal{R}_{mU}, \mathcal{R}_{mR}] \text{ for all } m \in \mathcal{M}) \geq 1 - \alpha + o(1).$$

Recently Fan et al. (2022b) proposed a similar approach to construct simultaneous confidence intervals for ranks in the context of the PL model with only the top choice observed for each comparison, which, however, requires the number of comparisons L for each connected item to be sufficiently large such that $L \gtrsim \text{poly}(\log n)$. In contrast, our procedure in Section 3.4 works without any constraints on the number of comparisons for each A_l (i.e., we even allow $L = 1$ for all comparisons) and is thus much more widely applicable, because in many real problems, sets of size M can be compared at different times and sometimes only once.

5. Numerical Studies

In this section, we validate the methodology and examine the theoretical results introduced in Sections 3 and 4. We conducted comprehensive simulation studies, but because of the page limit, we relegate the results to Online Appendix EC.1 and only briefly summarize our key findings here. We first validate the consistency and asymptotic distribution of the spectral estimator, and check the efficacy of the Gaussian bootstrap. All simulations match with our theoretical results perfectly. We also provide examples for constructing one-sample and two-sample confidence intervals for ranks and carrying out hypothesis testing of Examples 4–7. Finally, we empirically investigate the connections between the spectral method and MLE. In particular, we find the two-step or oracle weight spectral method behaves almost identically to MLE.

6. Real Data Analysis

We present the spectral ranking inferences for two real data sets in this section. The first one is about the ranking of statistics journals, which is based on pairwise comparisons (Journal B citing Journal A means A is preferred over B, which is consistent with the fact that good papers usually have higher citations). In particular, we can test, for two periods of time, whether the journal ranking has changed significantly. The second application is about movie ranking, which is based on multiway comparisons of different sizes (three or four movies are given to people to rank). The movie comparison graph shows strong heterogeneity in node degrees. Therefore, this comparison graph should be better modeled as a fixed graph without homogeneous sampling. In both cases, we will report the results from the two-step spectral estimator and the vanilla spectral estimator in Online Appendix EC.2.

6.1. Ranking of Statistics Journals

In this section, we study the Multi-Attribute Data set on Statisticians (MADStat), which contains citation information from 83,331 papers published in 36 journals between 1975 and 2015. The data set was collected and studied by Ji et al. (2022, 2023).

We follow the convention of Ji et al. (2023) to establish our pairwise comparison data. We will use journals' abbreviations given in the data. We refer interested readers to the complete journal names on the data website, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/V7VUFO. Firstly, we excluded three probability journals—AOP, PTRF, and AIHPP—because of their fewer citation exchanges with other statistics publications. Hence, our study comprises a total of 33 journals. Secondly, we only examine papers published between 2006 and 2015. For instance, if we treat 2010 as our reference year, we only count comparison results indicating "Journal A is superior to Journal B" if, and only if, a paper published in Journal B in 2010 has cited another paper that was published in Journal A between the years 2001 and 2010. This approach favors more recent citations, thus allowing the journal rankings to better reflect the most current trends and information. Finally, we chose to divide our study into two periods, 2006–2010 and 2011–2015, to detect the possible rank changes of journals.

Utilizing the data, we showcase the ranking inference results, summarized in Table 1. These results include two-sided, one-sided, and uniform one-sided confidence intervals for ranks within each of the two time periods (2006–2010 and 2011–2015). We calculate these intervals using the two-step spectral method over the fixed comparison graph (the results based on the one-step vanilla spectral method are presented in Online Appendix EC.2.1) and using the bootstrap method detailed in Sections 3.3–3.5.

From Table 1, we can easily get answers to the following questions on ranking inference. For example, is each journal's rank maintained unchanged across the two time periods? At a significance level of $\alpha = 10\%$, we find that the ranks of the following journals in alphabetical order demonstrate significant differences between the two time frames (Example 6):

$$AISM, AoAS, Biost, CSTM, EJS, JMLR, JoAS, JSPI.$$

This aligns with real-world observations. For instance, JMLR, EJS, and AOAS are newer journals that emerged after 2000. As a result, these journals received fewer citations in the earlier period and got recognized better in the more recent period.

We then turn our attention to the stability of the highest-ranked journals. Referring to Table 1, we observe that the top four journals (AoS, Bka, JASA, and JRSSB, known as the Big-Four in statistics) maintain their positions strongly across different time periods.

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

Table 1. Ranking Inference Results for 33 Journals in 2006–2010 and 2011–2015 Based on the Two-Step Spectral Estimator

	2006–2010						2011–2015					
Journal	$\tilde{\theta}$	$\tilde{r}$	TCI	OCI	UOCI	Count	$\tilde{\theta}$	$\tilde{r}$	TCI	OCI	UOCI	Count
JRSSB	1.654	1	[1, 1]	[1, n]	[1, n]	5,282	1.553	1	[1, 2]	[1, n]	[1, n]	5,513
AoS	1.206	3	[2, 4]	[2, n]	[2, n]	7,674	1.522	2	[1, 2]	[1, n]	[1, n]	11,316
Bka	1.316	2	[2, 3]	[2, n]	[2, n]	5,579	1.202	3	[3, 3]	[3, n]	[3, n]	6,399
JASA	1.165	4	[3, 4]	[3, n]	[3, n]	9,652	1.064	4	[4, 4]	[4, n]	[4, n]	10,862
JMLR	−0.053	20	[14, 25]	[15, n]	[13, n]	1,100	0.721	5	[5, 7]	[5, n]	[5, n]	2,551
Biost	0.288	13	[10, 18]	[10, n]	[9, n]	2,175	0.591	6	[5, 9]	[5, n]	[5, n]	2,727
Bcs	0.820	5	[5, 7]	[5, n]	[5, n]	6,614	0.571	7	[5, 9]	[6, n]	[5, n]	6,450
StSci	0.668	7	[5, 9]	[5, n]	[5, n]	1,796	0.437	8	[6, 13]	[6, n]	[6, n]	2,461
Sini	0.416	10	[9, 14]	[9, n]	[8, n]	3,701	0.374	9	[8, 13]	[8, n]	[8, n]	4,915
JRSSA	0.239	14	[10, 20]	[10, n]	[9, n]	893	0.370	10	[6, 13]	[8, n]	[6, n]	865
JCGS	0.605	8	[6, 9]	[6, n]	[6, n]	2,493	0.338	11	[8, 13]	[8, n]	[8, n]	3,105
Bern	0.793	6	[5, 8]	[5, n]	[5, n]	1,575	0.336	12	[8, 13]	[8, n]	[8, n]	2,613
ScaJS	0.528	9	[7, 12]	[7, n]	[6, n]	2,442	0.258	13	[8, 13]	[9, n]	[8, n]	2,573
JRSSC	0.113	15	[11, 22]	[11, n]	[11, n]	1,401	0.020	14	[14, 19]	[14, n]	[12, n]	1,492
AoAS	−1.463	30	[30, 33]	[30, n]	[30, n]	1,258	−0.017	15	[14, 20]	[14, n]	[14, n]	3,768
CanJS	0.101	17	[11, 22]	[11, n]	[11, n]	1,694	−0.033	16	[14, 20]	[14, n]	[14, n]	1,702
JSPI	−0.327	26	[24, 26]	[24, n]	[22, n]	6,505	−0.046	17	[14, 20]	[14, n]	[14, n]	6,732
JTSA	0.289	12	[9, 18]	[10, n]	[8, n]	751	−0.101	18	[14, 22]	[14, n]	[14, n]	1,026
JMVA	−0.126	22	[17, 25]	[17, n]	[15, n]	3,833	−0.148	19	[14, 22]	[15, n]	[14, n]	6,454
SMed	−0.131	23	[17, 25]	[18, n]	[17, n]	6,626	−0.242	20	[18, 25]	[18, n]	[17, n]	6,857
Extrem	−2.099	33	[30, 33]	[31, n]	[30, n]	173	−0.312	21	[16, 30]	[18, n]	[14, n]	487
AISM	0.317	11	[9, 18]	[10, n]	[9, n]	1,313	−0.359	22	[19, 30]	[20, n]	[18, n]	1,605
EJS	−1.717	32	[30, 33]	[30, n]	[30, n]	1,366	−0.367	23	[20, 29]	[20, n]	[19, n]	4,112
SPLet	−0.033	19	[15, 25]	[15, n]	[13, n]	3,651	−0.384	24	[21, 29]	[21, n]	[19, n]	4,439
CSDA	−0.975	29	[27, 30]	[27, n]	[27, n]	6,732	−0.467	25	[21, 30]	[21, n]	[21, n]	8,717
JNS	−0.255	25	[19, 26]	[20, n]	[17, n]	1,286	−0.484	26	[21, 30]	[21, n]	[21, n]	1,895
ISRe	0.082	18	[11, 25]	[11, n]	[10, n]	511	−0.491	27	[21, 30]	[21, n]	[20, n]	905
AuNZ	0.108	16	[11, 23]	[11, n]	[10, n]	862	−0.504	28	[21, 30]	[21, n]	[20, n]	816
JClas	−0.185	24	[15, 26]	[15, n]	[11, n]	260	−0.535	29	[18, 30]	[20, n]	[14, n]	224
SCmp	−0.096	21	[15, 25]	[15, n]	[14, n]	1,309	−0.561	30	[23, 30]	[24, n]	[21, n]	2,650
Bay	−1.494	31	[30, 33]	[30, n]	[27, n]	279	−1.102	31	[31, 32]	[31, n]	[30, n]	842
CSTM	−0.843	27	[27, 29]	[27, n]	[27, n]	2,975	−1.296	32	[31, 32]	[31, n]	[31, n]	4,057
JoAS	−0.912	28	[27, 30]	[27, n]	[27, n]	1,055	−1.904	33	[33, 33]	[33, n]	[33, n]	2,780

Notes. For each time period, there are six columns. The first column $\tilde{\theta}$ denotes the estimated underlying scores. The second through fifth columns denote their relative ranks and two-sided, one-sided, and uniform one-sided confidence intervals for the ranks with 95% coverage level, respectively. The sixth column denotes the number of comparisons in which each journal is involved.

Furthermore, with a significance level of $\alpha = 10\%$, we reject the hypothesis that the top-7 ranked items remain constant across the two time periods (Example 7). Specifically, for 2006–2010, the 95% confidence set for the top-7 items includes

$$AoS, Bern, Bcs, Bka, JASA, JCGS, JRSSB, ScaJS, StSci.$$

And for 2011–2015, the 95% confidence set for the top-7 items includes

$$AoS, Bcs, Biost, Bka, JASA, JMLR, JRSSA, JRSSB, StSci.$$

Clearly, these sets intersect only at six items, smaller than seven, reflecting a shift in the rankings over the two periods.

6.2. Ranking of Movies

In this section, we construct confidence intervals for the ranks of movies or television series featured within the Netflix Prize competition (Bennett and Lanning 2007), which aims to enhance the precision of the Netflix recommendation algorithm. The data set we examine corresponds to 100 random three- and four-candidate elections drawn from Data Set 1 of Mattei et al. (2012), which was extracted from the original Netflix Prize data set, devoid of any ties. The data set contains 196 movies in total and 163,759 three-way or four-way comparisons. For simplicity, we only use the top ranked movie, although it is straightforward to apply the multilevel breaking to use the complete ranking data. This data set can be accessed at the website https://www.preflib.org/dataset/00004.

We compute two-sided, one-sided, and uniform one-sided confidence intervals employing the bootstrap method as described in Sections 3.3–3.5, based on the two-step spectral method. The results are shown in Table 2. Additionally, the results from the one-step vanilla spectral method are detailed in Table EC.9 in Online Appendix EC.2.2.

Note that the heterogeneity in the number of comparisons (see the "Count" column in Table 2) is more

Table 2. Ranking Inference Results for Top-20 Netflix Movies or TV Series Based on the Two-Step Spectral Estimator

Movie or series	$\tilde{\theta}$	$\tilde{r}$	TCI	OCI	UOCI	Count
The Silence of the Lambs	3.002	1	[1, 1]	[1, n]	[1, n]	19,589
The Green Mile	2.649	2	[2, 4]	[2, n]	[2, n]	5,391
Shrek (full-screen)	2.626	3	[2, 4]	[2, n]	[2, n]	19,447
The X-Files: Season 2	2.524	4	[2, 7]	[2, n]	[2, n]	1,114
Ray	2.426	5	[4, 7]	[4, n]	[4, n]	7,905
The X-Files: Season 3	2.357	6	[4, 10]	[4, n]	[2, n]	1,442
The West Wing: Season 1	2.278	7	[4, 10]	[4, n]	[4, n]	3,263
National Lampoon's Animal House	2.196	8	[6, 10]	[6, n]	[5, n]	10,074
Aladdin: Platinum Edition	2.154	9	[6, 13]	[6, n]	[5, n]	3,355
Seven	2.143	10	[6, 11]	[7, n]	[6, n]	16,305
Back to the Future	2.030	11	[9, 15]	[9, n]	[8, n]	6,428
Blade Runner	1.968	12	[10, 16]	[10, n]	[9, n]	5,597
Harry Potter and the Sorcerer's Stone	1.842	13	[12, 22]	[12, n]	[11, n]	7,976
High Noon	1.821	14	[11, 25]	[11, n]	[10, n]	1,902
Sex and the City: Season 6: Part 2	1.770	15	[11, 30]	[11, n]	[8, n]	532
Jaws	1.749	16	[13, 25]	[13, n]	[13, n]	8,383
The Ten Commandments	1.735	17	[13, 28]	[13, n]	[12, n]	2,186
Willy Wonka & the Chocolate Factory	1.714	18	[13, 26]	[13, n]	[13, n]	9,188
Stalag 17	1.697	19	[12, 34]	[12, n]	[11, n]	806
Unforgiven	1.633	20	[14, 29]	[14, n]	[14, n]	9,422

Notes. The first column $\tilde{\theta}$ denotes the estimated underlying scores. The second through the fifth columns denote their relative ranks and two-sided, one-sided, and uniform one-sided confidence intervals for ranks with 95% coverage level, respectively. The sixth column denotes the number of comparisons in which each movie is involved.

pronounced than in the journal ranking data, resulting in adaptive lengths of the rank confidence intervals. Using the "OCI" column of Table 2, we test whether each individual movie belongs to the top-5 rated group ($K = 5$ in Example 4). We fail to reject this hypothesis for the first 10 films listed in Table 2. Additionally, the uniform one-sided confidence intervals (shown in the "UOCI" column) can be used to construct a candidate confidence set for the true top-5 movies ($K = 5$ in Example 5). The results suggest that, beyond the top-11 ranked films, the show *Sex and the City: Season 6: Part 2* should also be included in the top-5 confidence set. This inclusion, despite its lower ranking compared with *High Noon* (which is excluded from the set), is due to the smaller number of comparisons involving this show, which yields a wider confidence interval.

7. Conclusion and Discussion

In this work, we studied the performance of the spectral method in preference score estimation, quantified the asymptotic distribution of the estimated scores, and explored one-sample and two-sample inference on ranks. In particular, we worked with general multiway comparisons with fixed comparison graphs, where the size of each comparison can vary and can be as low as only one. This is much closer to real applications than the homogeneous random sampling assumption imposed in the BTL or PL models. The applications of journal ranking and movie ranking have demonstrated the clear usefulness of our proposed methodologies. Furthermore, we studied the relationship between the spectral method and the MLE in terms of estimation efficiency and revealed that with a carefully chosen weighting scheme, the spectral method can approximately achieve the same efficiency as the MLE, which is also verified using numerical simulations. Finally, to the best of our knowledge, it is the first time that effective two-sample rank testing methods have been proposed in the literature.

Although we have made significant improvements in relaxing conditions, the role of general comparison graphs is still not fully understood, especially in the setting of multiway comparisons. Questions like how to design a better sampling regime, either online or offline, remain open. In addition, the spectral method essentially encodes multiway comparisons into pairwise comparisons, where the encoding will break data independence. The best encoding or breaking method should be further investigated. Finally, a set of recent works on ranking inferences opens the door to many possibilities of theoretical studies on ranking inferences and related problems such as assortment optimization, under the setting of, say, rank time series, rank change point detection, rank panel data, recommendation based on rank inferences, uncertainty quantification, and inference for properties of the optimal assortment. These may find potential application in numerous management settings.

References

Aouad A, Farias V, Levi R, Segev D (2018) The approximability of assortment optimization under ranking preferences. *Oper. Res.* 66(6):1661–1669.

Avery CN, Glickman ME, Hoxby CM, Metrick A (2013) A revealed preference ranking of U.S. colleges and universities. *Quart. J. Econom.* 128(1):425–467.

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.

Azari Soufiani H, Chen W, Parkes DC, Xia L (2013) Generalized method-of-moments for rank aggregation. Burges CJ, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems (NIPS 2013)*, vol. 26 (MIT Press, Cambridge, MA), 2706–2714.

Baltrunas L, Makcinskas T, Ricci F (2010) Group recommendations with rank aggregation and collaborative filtering. Amatriain X, Torrens M, Resnick P, Zanker M, eds. *RecSys '10 Proc. 4th ACM Conf. Recommender Systems* (Association for Computing Machinery, New York), 119–126.

Bennett J, Lanning S (2007) The Netflix prize. *Proc. KDD Cup Workshop* (Association for Computing Machinery, New York).

Caron F, Teh YW, Murphy TB (2014) Bayesian nonparametric Plackett-Luce models for the analysis of preferences for college degree programmes. *Ann. Appl. Statist.* 8(2):1145–1181.

Chen Y, Suh C (2015) Spectral MLE: Top-*K* rank aggregation from pairwise comparisons. Bach F, Blei D, eds. *Proc. 32nd Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 37 (PMLR, New York), 371–380.

Chen P, Gao C, Zhang AY (2022) Partial recovery for top-*k* ranking: Optimality of MLE and suboptimality of the spectral method. *Ann. Statist.* 50(3):1618–1652.

Chen X, Krishnamurthy A, Wang Y (2023) Robust dynamic assortment optimization in the presence of outlier customers. *Oper. Res.* 72(3):999–1015.

Chen X, Wang Y, Zhou Y (2020) Dynamic assortment optimization with changing contextual information. *J. Machine Learn. Res.* 21(216):8918–8961.

Chen Y, Fan J, Ma C, Wang K (2019) Spectral method and regularized MLE are both optimal for top-*K* ranking. *Ann. Statist.* 47(4):2204–2235.

Cheng W, Dembczynski K, Hüllermeier E (2010) Label ranking methods based on the Plackett-Luce model. Fürnkranz J, Joachims T, eds. *Proc. 27th Internat. Conf. Machine Learn.* (Omnipress, Madison, WI), 215–222.

Chernozhukov V, Chetverikov D, Kato K (2017) Central limit theorems and bootstrap in high dimensions. *Ann. Probab.* 45(4): 2309–2352.

Chernozhukov V, Chetverikov D, Kato K, Koike Y (2019) Improved central limit theorem and bootstrap approximations in high dimensions. Preprint, submitted December 22, https://arxiv.org/abs/1912.10529.

Davis JM, Gallego G, Topaloglu H (2014) Assortment optimization under variants of the nested logit model. *Oper. Res.* 62(2):250–273.

Dwork C, Kumar R, Naor M, Sivakumar D (2001) Rank aggregation methods for the web. Shen V, Saito N, Lyu MR, Zurko ME, eds. *Proc. 10th Internat. Conf. World Wide Web* (Association for Computing Machinery, New York), 613–622.

Fan J, Hou J, Yu M (2022a) Uncertainty quantification of MLE for entity ranking with covariates. Preprint, submitted December 20, https://arxiv.org/abs/2212.09961.

Fan J, Lou Z, Wang W, Yu M (2022b) Ranking inferences based on the top choice of multiway comparisons. Preprint, submitted November 22, https://arxiv.org/abs/2211.11957.

Gallego G, Topaloglu H (2014) Constrained assortment optimization for the nested logit model. *Management Sci.* 60(10):2583–2601.

Gao C, Shen Y, Zhang AY (2021) Uncertainty quantification in the Bradley-Terry-Luce model. Preprint, submitted October 8, https://arxiv.org/abs/2110.03874.

Guiver J, Snelson E (2009) Bayesian inference for Plackett-Luce ranking models. Bottou L, Littman M, eds. *Proc. 26th Annual Internat. Conf. Machine Learn.* (Association for Computing Machinery, New York), 377–384.

Hajek B, Oh S, Xu J (2014) Minimax-optimal inference from partial rankings. Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems (NIPS 2014)*, vol. 27 (MIT Press, Cambridge, MA), 1475–1483.

Han R, Xu Y (2023) A unified analysis of likelihood-based estimators in the Plackett–Luce model. Preprint, submitted June 5, https://arxiv.org/abs/2306.02821.

Han R, Ye R, Tan C, Chen K (2020) Asymptotic theory of sparse Bradley-Terry model. *Ann. Appl. Probab.* 30(5):2491–2515.

Hunter DR (2004) MM algorithms for generalized Bradley-Terry models. *Ann. Statist.* 32(1):384–406.

Jang M, Kim S, Suh C (2018) Top-*K* rank aggregation from *M*-wise comparisons. *IEEE J. Selected Topics Signal Processing* 12(5):989–1004.

Jang M, Kim S, Suh C, Oh S (2016) Top-K ranking from pairwise comparisons: When spectral ranking is optimal. Preprint, submitted March 14, https://arxiv.org/abs/1603.04153.

Ji P, Jin J, Ke ZT, Li W (2022) Co-citation and co-authorship networks of statisticians. *J. Bus. Econom. Statist.* 40(2):469–485.

Ji P, Jin J, Ke ZT, Li W (2023) Meta-analysis on citations for statisticians. Working paper, University of Georgia, Athens.

Johnson VE, Deaner RO, Van Schaik CP (2002) Bayesian analysis of rank data with application to primate intelligence experiments. *J. Amer. Statist. Assoc.* 97(457):8–17.

Li W, Shrotriya S, Rinaldo A (2022) $\ell\infty$-bounds of the MLE in the BTL model under general comparison graphs. Cussens J, Zhang K, eds. *Proc. 38th Conf. Uncertainty Artificial Intelligence*, Proceedings of Machine Learning Research, vol. 180 (PMLR, New York), 1178–1187.

Li H, Simchi-Levi D, Wu MX, Zhu W (2019) Estimating and exploiting the impact of photo layout: A structural approach. Preprint, submitted October 16, https://doi.org/10.2139/ssrn.3470877.

Liu Y, Fang EX, Lu J (2022) Lagrangian inference for ranking problems. *Oper. Res.* 71(1):202–223.

Luce RD (1959) *Individual Choice Behavior: A Theoretical Analysis* (John Wiley & Sons, Inc., New York).

Massey K (1997) Statistical models applied to the rating of sports teams. Undergraduate thesis, Bluefield College, Bluefield, VA.

Mattei N, Walsh T (2013) Preflib: A library for preferences http://www.preflib.org. Perny P, Pirlot M, Tsoukiàs A, eds. *Algorithmic Decision Theory ADT 2013* (Springer, Berlin), 259–270.

Mattei N, Forshee J, Goldsmith J (2012) An empirical study of voting rules and manipulation with large datasets. Brandt F, Faliszewski P, eds. *Workshop Notes 4th Internat. Workshop Comput. Soc. Choice COMSOC 2012 (Krakow, Poland)*, 299–310.

Maystre L, Grossglauser M (2015) Fast and accurate inference of Plackett-Luce models. Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, eds. *Adv. Neural Inform. Processing Systems (NIPS 2015)*, vol. 28 (MIT Press, Cambridge, MA), 172–180.

Negahban S, Oh S, Shah D (2012) Iterative ranking from pair-wise comparisons. Pereira F, Burges CK, Bottou L, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems (NIPS 2012)*, vol. 25 (MIT Press, Cambridge, MA), 2474–2482.

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, et al. (2022) Training language models to follow instructions with human feedback. Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, eds. *Adv. Neural Inform. Processing Systems (NeurIPS 2022)*, vol. 35 (MIT Press, Cambridge, MA), 27730–27744.

Plackett RL (1975) The analysis of permutations. *J. Roy. Statist. Soc. Ser. C Appl. Statist.* 24(2):193–202.

Portnoy S (1986) On the central limit theorem in *Rp* when $p\to\infty$. *Probab. Theory Related Fields* 73(4):571–583.

Rusmevichientong P, Topaloglu H (2012) Robust assortment optimization in revenue management under the multinomial logit choice model. *Oper. Res.* 60(4):865–882.

Rusmevichientong P, Shen Z-JM, Shmoys DB (2010) Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Oper. Res.* 58(6):1666–1680.

Shah N, Balakrishnan S, Bradley J, Parekh A, Ramchandran K, Wainwright M (2015) Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. Lebanon G,

Vishwanathan SVN, eds. *Proc. 18th Internat. Conf. Artificial Intelligence Statist.*, Proceedings of Machine Learning Research, vol. 38 (PMLR, New York), 856–865.

Shen S, Chen X, Fang E, Lu J (2023) Combinatorial inference on the optimal assortment in multinomial logit models. Preprint, submitted February 27, https://doi.org/10.2139/ssrn.4371919.

Simons G, Yao Y-C (1999) Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons. *Ann. Statist.* 27(3):1041–1060.

Sumida M, Gallego G, Rusmevichientong P, Topaloglu H, Davis J (2020) Revenue-utility tradeoff in assortment optimization under the multinomial logit model with totally unimodular constraints. *Management Sci.* 67(5):2845–2869.

Szörényi B, Busa-Fekete R, Paul A, Hüllermeier E (2015) Online rank elicitation for Plackett-Luce: A dueling bandits approach. Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, eds. *Adv. Neural Inform. Processing Systems (NIPS 2015)*, vol. 28 (MIT Press, Cambridge, MA), 604–612.

Talluri K, Van Ryzin G (2004) Revenue management under a general discrete choice model of consumer behavior. *Management Sci.* 50(1):15–33.

Turner H, Firth D (2012) Bradley-Terry models in R: The BradleyTerry2 package. *J. Statist. Software* 48(9):1–21.

Vulcano G, Van Ryzin G, Ratliff R (2012) Estimating primary demand for substitutable products from sales transaction data. *Oper. Res.* 60(2):313–334.

Wang X, Bendersky M, Metzler D, Najork M (2016) Learning to rank with selection bias in personal search. Raffaele P, Fabrizio S, eds. *Proc. 39th Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (Association for Computing Machinery, New York), 115–124.

Zhang H, Rusmevichientong P, Topaloglu H (2020) Assortment optimization under the paired combinatorial logit model. *Oper. Res.* 68(3):741–761.

Jianqing Fan is Frederick L. Moore '18 Professor of Finance, Professor of Statistics, and Professor of Operations Research and Financial Engineering at Princeton University. He is the winner of The COPSS Presidents' Award, Morningside Gold Medal for Applied Mathematics, Pao-Lu Hsu Prize, Guy Medal in Silver, Noether Distinguished Scholar. His research lies in the developments of statistical machine learning theory and methods and their applications in finance, economics, genomics, and health.

Zhipeng Lou is an assistant professor in the Department of Mathematics at the University of California, San Diego. His research interests include high-dimensional statistical inference and time series analysis.

Weichen Wang is an assistant professor in the area of Innovation and Information Management at HKU Business School at the University of Hong Kong. His research interests include econometric factor analysis, high-dimensional statistical inference, robust methodologies, and machine learning.

Mengxin Yu is an assistant professor in the Department of Statistics and Data Science at Washington University in St. Louis. Her research interests include human preference learning, causal inference, and machine learning safety.

Downloaded from informs.org by [144.121.140.150] on 02 October 2025, at 03:52 . For personal use only, all rights reserved.