# A Margin-based MLE for Crowdsourced Partial Ranking

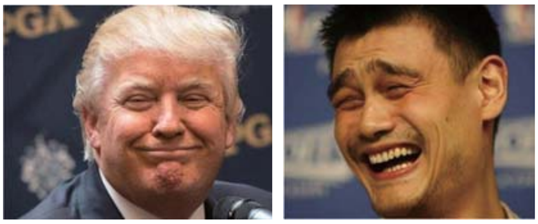
Qianqian Xu[1], Jiechao Xiong[2], Xinwei Sun[3,4],
Zhiyong Yang[5], Xiaochun Cao[5], Qingming Huang[1,6,7]*, Yuan Yao[8]*
[1] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, 100190, China
[2] Tencent AI Lab, Shenzhen, 518057, China
[3] School of Mathematical Sciences, Peking University, Beijing, 100871, China
[4] DeepWise AI Lab, Beijing, 100085, China
[5] State Key Laboratory of Info. Security (SKLOIS), Inst. of Info. Engin., CAS, Beijing, 100093, China
[6] University of Chinese Academy of Sciences, Beijing, 100049, China
[7] Key Lab of Big Data Mining and Knowledge Management, CAS, Beijing, 100190, China
[8] Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong
xuqianqian@ict.ac.cn,jcxiong@tencent.com,sxwxiaoxiaohehe@pku.edu.cn
{yangzhiyong,caoxiaochun}@iie.ac.cn,qmhuang@ucas.ac.cn,yuany@ust.hk

## ABSTRACT

A preference order or ranking aggregated from pairwise comparison data is commonly understood as a strict total order. However, in real-world scenarios, some items are intrinsically ambiguous in comparisons, which may very well be an inherent uncertainty of the data. In this case, the conventional total order ranking can not capture such uncertainty with mere global ranking or utility scores. In this paper, we are specifically interested in the recent surge in crowdsourcing applications to predict partial but more accurate (i.e., making less incorrect statements) orders rather than complete ones. To do so, we propose a novel framework to learn some probabilistic models of partial orders as a *margin-based Maximum Likelihood Estimate* (MLE) method. We prove that the induced MLE is a joint convex optimization problem with respect to all the parameters, including the global ranking scores and margin parameter. Moreover, three kinds of generalized linear models are studied, including the basic uniform model, Bradley-Terry model, and Thurstone-Mosteller model, equipped with some theoretical analysis on FDR and Power control for the proposed methods. The validity of these models are supported by experiments with both simulated and real-world datasets, which shows that the proposed models exhibit improvements compared with traditional state-of-the-art algorithms.

## CCS CONCEPTS

• **Information systems → Rank aggregation**;



(a) Trump (b) Ming Yao

**Figure 1: Smile as a relative attribute in paired comparisons.**

## KEYWORDS

Partial Ranking; Pairwise Comparison; Crowdsourcing; Margin-based MLE

**ACM Reference Format:**
Qianqian Xu, Jiechao Xiong, Xinwei Sun, Zhiyong Yang, Xiaochun Cao, Qingming Huang, Yuan Yao. 2018. A Margin-based MLE for Crowdsourced Partial Ranking. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240597

## 1 INTRODUCTION

Imagine you are given a pile of distorted images of the same content, and you are asked to sort or rank them according to their quality. Can you do it? In other tasks such as relative attribute ordering in computer vision, for example in Figure 1, can you rank the faces according to the "degree" of smiling? These are typical scenarios in crowdsourced ranking.

Nature imposes a limitation that humans are unable to make accurate preference judgement on even moderately large sets. As it has been argued that most people can rank only between 5 to 9 alternatives at a time [27]. This is probably why many rating scales (e.g. the ones used by Amazon, eBay, Netflix, YouTube) are based on a 5-star (level) scale. In a 5-star test, individuals are asked to give a rating from Bad to Excellent in 5 levels (e.g. Bad-1, Poor-2, Fair-3, Good-4, and Excellent-5) to grade the candidates. This leads to partial orders or ranking of the candidates where the items on the same level will be regarded as equivalent classes. There are some

*Corresponding author.

*MM '18, October 22–26, 2018, Seoul, Republic of Korea*

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240597

work in the literature studying how to organize information in partial orders of such tied subsets or equivalent classes (partitions, bucket orders) [14, 19]. Specifically, the authors in [19] address computational aspects that arise when working with empirical distributions on partially ranked data.

Yet in many crowdsourcing tasks, even the 5-star scale may suffer from various problems such as ambiguity in the definition of scales, dissimilar interpretations of the scale among users, and so on, e.g. argued in [3] and reference therein. To address this issue, the pairwise comparison method becomes a rising paradigm recently in many crowdsourcing platforms, as for most people, it is a harder task to rank or rate many candidates than to compare a pair of candidates at a time. In pairwise comparisons, frequently, the available data presented to us is in the following form: the quality of image A is better than image B, etc. A ranking aggregated from pairwise comparison data is commonly understood as a strict total order, i.e., an irreflexive, asymmetric, and transitive relation, specifying for all pairs whether $i$ precedes $j$, or $j$ precedes $i$. [14] attempts to discover an underlying bucket or partial order from pairwise precedence information between the items without any ties.

Although some items or candidates could be obvious to rank, the ambiguity in choosing the preference is ubiquitous that often imposes some difficulties in making the choice. For example, the following list describes such cases met in crowdsourcing experiments.

**Example** 1. *In relative attributes in computer vision [24], some attributes such as smile or age, are hard to judge absolutely, but accessible to human within a pair on choosing which one to be stronger in the attribute. Surely there might be some obvious images easy to judge. Yet there will be other images where the distinction is quite subtle, or hardly perceivable. Figure 1 gives an example. Who is smiling more, Trump or Yao? Some people may think the basketball star Ming Yao is more smiling than Trump; while some people may think Ming Yao looks crying, so they prefer to Trump as more smiling. Besides, others may think it is difficult to tell which one in the pair looks stronger in the smile attribute. Participants may choose to abstain from this judgement when they are too confusing to make a decision.*

**Example** 2. *In subjective multimedia quality assessment [34, 35], videos and images of the same content are to be evaluated for its quality. Some pairs are easy to distinguish, while others are not. In particular, there might be multi-criteria among heterogeneous raters. In these cases, annotators may declare these two are confusing thus difficult to judge.*

**Example** 3. *In crowdsourced pairwise ranking platforms such as Allourideas[1], an option that "I can't decide" is provided with further information such as "I like both ideas", "I think both ideas are the same", "I don't like either idea", or "I don't know enough about either idea", etc. For example, in world college ranking a participant is asked about "which university (of the following two) would you rather attend?". When a voter thinks the two colleges listed are incomparable and difficult to judge, he may click this button with possible further options. Such voters essentially provide some information on partial orders, which can be distinguished from those voters who click this button just because they don't know both of these two colleges or one of them well.*

This kind of pairwise comparison data, together with "I can't decide" type of decision, arises in a variety of crowdsourcing applications. In all these examples, if a rater is not sufficiently certain regarding the relative order of the two items, he may abstain from his choice decision and instead declare these two as being incomparable. In fact, partial ranking can be interpreted as a ranking with partial abstention. In this way, a dataset with abstention of this kind provides us information about possible ties or equivalent classes of items in partial orders.

Despite a considerable amount of work on ranking in general and pairwise ranking in particular, there lacks a systematic treatment on learning partial orders or rankings from such pairwise comparison data with abstentions, which are ubiquitous in crowdsourcing applications nowadays. Among the prior work on partial ranking up to our knowledge, the one that comes closest to our goal is [4]. The idea is that it produces predictions in the form of partial order by thresholding a (valued) pairwise preference relation, i.e., by a "$\alpha$-cut" of preference relation. However, it leaves the optimal choice of hyper-parameter $\alpha$ to various heuristics and needs to know in advance the preference relation between every pair of items (i.e., $n(n-1)/2$ pairs in total for $n$ items), which requires a large number of comparisons, being too prohibitive in modern applications.

To fill in this gap, in this paper, we propose a novel framework to learn partial ranking probabilistic models as a margin-based Maximum Likelihood Estimate (MLE) method. In contrast to [4], all the parameters, including the global ranking score and the hyper parameter as threshold (called margin parameter here), can be automatically learned from pairwise comparison data with abstentions via a convex optimization. Our framework can deal with incomplete and imbalanced data, as an extension of the HodgeRank [16] from total orders to partial orders with generalized linear models.

As a summary, our main contributions in this new framework are highlighted as follows:

(A) We propose a framework of learning partial rankings from pairwise comparison data with abstentions, based on a margin-based Maximum Likelihood Estimate (MLE) for probabilistic models. We prove that for a general class of models, the induced MLE is a convex optimization problem with respect to all the parameters, including the global ranking scores and threshold/margin parameter.
(B) In this unified framework, three kinds of generalized linear models are particularly studied, including the basic uniform model, Bradley-Terry model, and Thurstone-Mosteller model, equipped with theoretical analysis on FDR and Power control of our proposed method.
(C) Experiments on simulated and crowdsourcing real-world datasets together show that our algorithm works effectively in practice.

The remainder of this paper is organized as follows. Sec.2 contains a review of related work. We systematically introduce the methodology for partial ranking in Sec.3. Detailed experiments with simulated and real-world datasets are presented in Sec.4. Finally, Sec.5 presents the conclusive remarks.

[1] http://www.allourideas.org/

## 2 RELATED WORK

**Pairwise Ranking**. Statistical preference aggregation, in particular ranking or rating from pairwise comparisons, is a classical problem that can be traced back to the $18^{th}$ century. This subject area has been widely studied in various fields including the social choice and voting theory in economics [1, 12], statistics [10, 22], multimedia [34, 35], computer vision [20, 37, 38], and others [2, 7, 17, 23, 25, 26, 30–32].

Various algorithms have been studied to solve this problem. They include maximum likelihood under a Bradley-Terry model assumption, rank centrality (PageRank/MC3) [9, 21], HodgeRank [16], and a pairwise variant of Borda count [11], etc. However, all of these methods have a major drawback: they aim to find one global common consensus, that assumes all users' choices are stochastic revelation of a common global preference function on candidates. To capture the discrepancies among users, lately [36] proposes a parsimonious mixed effect HodgeRank, which considers that a majority of users may follow the common social preferences while some users may exhibit distinct personalized preferences. However, all these methods above do not consider the inherent characteristic of real-world data: some pairs are intrinsically ambiguous, thus may be difficult to derive a strict global ranking. In this paper, we will focus on this kind of setting, allowing a model to make predictions in the form of partial instead of total orders.

**Partial Ranking**. Despite a considerable amount of work on ranking in general and pairwise ranking, the literature on partial rankings is relatively sparse. Pairwise comparisons with abstentions are governed by partial orders or rankings. But the notion of abstention is actually originated from classification community [5]. In classification with a reject option, for example, a classifier may abstain from a class prediction if making no decision is considered less harmful than making an unreliable and hence potentially false decision. Recently, worth mentioning is the work on a specific type of partial orders, namely linear orders of unsorted or tied subsets (partitions, bucket orders) [14, 19]. However, the problems addressed in these work are different from our goals. Among the existing work in the literature, [4] is the one that comes closest to our goal, which produces predictions in the form of partial order by thresholding a (valued) pairwise preference relation, i.e., by a "$\alpha$-cut" of preference relation. It lacks a solid principle to decide the hyper parameter $\alpha$ as the threshold. Moreover, it needs to know in advance the preference relation between every pair of items. In this paper, we propose a margin-based MLE for partial order ranking based on probability model which could solve these problems in [4].

## 3 METHODOLOGY

### 3.1 Pairwise Ranking on Graphs

Suppose there are $n$ alternatives or items to be ranked. The pairwise comparison labels collected from users can be naturally represented as a directed comparison graph $G = (V; E)$. Let $V = \{1, 2, \ldots, n\}$ be the vertex set of $n$ items and $E = \{(u, i, j) : i, j \in V, u \in U\}$ be the set of edges, where $U$ is the set of all users who compared items. User $u$ provides his/her preference between choice $i$ and $j$, such that $y^u_{ij} > 0$ means $u$ prefers $i$ to $j$ and $y^u_{ij} \leq 0$ otherwise. Hence we may assume $y : E \to R$ with skew-symmetry (orientation) $y^u_{ij} = -y^u_{ji}$. The magnitude of $y^u_{ij}$ can represent the degree of preference and it varies in applications. The simplest setting is the binary choice, where $y^u_{ij} = 1$ if $u$ prefers $i$ to $j$ and $y^u_{ij} = -1$ otherwise.

Traditionally, a statistical ranking is commonly understood as a strict total order, i.e., an irreflexive, asymmetric, and transitive relation $>$, specifying for all pairs whether $i$ precedes $j$, denoted $i > j$, or $j$ precedes $i$. The key property of transitivity can be seen as a principle of consistency: If $i$ is preferred to $j$ and $j$ is preferred to $k$, then $i$ must be preferred to $k$. However, in real-world applications, some pairs are intrinsically ambiguous, in this case, the rater cannot reliably decide whether the former should precede the latter or the other way around, he may abstain from this decision and instead declare these alternatives as being incomparable. Therefore, it might be misleading to merely look at a global total ranking (i.e., in which every pair of distinct elements is comparable) while ignoring the intrinsic ambiguity among items. In this paper, we focus on deriving a partial ranking based on a margin-based MLE method.

### 3.2 Partial Order Ranking

A partial order $\succ$ is a generalization of the relation $>$ mentioned above that preserves the consistency principle but is not necessarily total. Define $i \succ j$ as $(i > j) \wedge (\urcorner(j > i))$ . If, for two alternatives $i$ and $j$, neither $i \succ j$ nor $j \succ i$, then these alternatives are considered as incomparable, we then denote $i \perp j$ or equivalently $j \perp i$. In other words, if $i$ and $j$ are too similar such that we think neither $i$ precedes $j$ nor $j$ precedes $i$, we then claim that $i$ and $j$ are incomparable.

In [4], it proposes to learn a Partial Order Relation (POR) by a "$\alpha$-cut" of preference relation. Suppose $P(i, j)$ is a measure of support for the order (preference) relation $i \succ j$ with property $P(j, i) = 1 - P(i, j)$. Then a POR is defined as

$$\mathcal{R}_\alpha = \{i \succ j : P(i, j) \geq \alpha\}$$

by setting $\alpha$ big enough.

However, this construction of POR requires the preference relation between every pair of items (i.e., $n(n - 1)/2$ pairs in total for $n$ items). And each $P(i, j)$ is usually estimated by empirical probability between $i$ and $j$. Therefore, a good estimation of $P(i, j)$ needs a large number of comparisons.

### 3.3 Probability Model for Binary Data

In order to extend the methods to the case of small number of samples, we introduce the probability model for binary data. Suppose that the true scaling scores for $n$ items are $\boldsymbol{s} = [s_1, \cdots, s_n]$ and we collect $N$ pairwise comparison samples $\{(i_k, j_k, y_k)\}_{k=1}^N$ in total. Here $(i_k, j_k)$ is a pair of items, and $y_k$ is the corresponding comparison label. Suppose that, for the $k$th observation, $y_k$ is generated by:

$$y_k = \text{sign}(s_{i_k} - s_{j_k} + \epsilon_k),$$

where $\epsilon_k$ are $i.i.d$ and have a c.d.f $\Phi(t)$. Different $\Phi$ leads to different models. For example:

- Uniform model: $\Phi(t) = \frac{t+1}{2}$.
- Bradley-Terry model: $\Phi(t) = \frac{e^t}{1+e^t}$.
- Thurstone-Mosteller model: $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$.

Note that $P(y_k = 1) = 1 - \Phi(s_{j_k} - s_{i_k})$, a $\alpha$-cut of preference relation is thus equivalent to a $-\Phi^{-1}(1-\alpha)$-cut of the score difference function $f(k) = s_{i_k} - s_{j_k}$. Therefore, a POR can be obtained if the score can be estimated, which allows the comparison samples to be incomplete. It can be proved, such a cut indeed implies a POR.

**Proposition** 1. *For any* $\mathbf{s}$ *and* $\lambda > 0$*, the relation*

$$\mathcal{R}_\lambda = \{i \succ j : s_i - s_j > \lambda\} \tag{1}$$

*is a Partial Order.*

### 3.4 Extended Probability Model

As stated that, any value of $\lambda$ can produce a POR, then how can we choose a proper one is a key step. In real-world applications, as some pairs are intrinsically ambiguous, raters usually provide a third option, "I can't decide" or "They are comparable". Such kind of data can help us to determine the "optimal cut" here. To fit this kind data with three options, we extend the probability model as follows:

$$y_k = \begin{cases} 1, & s_{i_k} - s_{j_k} + \epsilon_k > \lambda; \\ -1, & s_{i_k} - s_{j_k} + \epsilon_k < -\lambda; \\ 0, & \text{else.} \end{cases} \tag{2}$$

where $y_k = 0$ indicates the annotator thinks that $i$ and $j$ are too close to judge. Then under this model, the POR in (1) has an explicit meaning: an oracle annotator, whose $\epsilon_{ij} = 0$, will give exactly this POR!

### 3.5 Maximum Likelihood Estimator

With the label distribution modeled, in this section we elaborate a Maximum Likelihood method to estimate the model parameters. First, we construct the design matrix $\boldsymbol{X}$ as $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \cdots \boldsymbol{x}_N^\top]^\top$, where $\boldsymbol{x}_k = \boldsymbol{e}_{j_k} - \boldsymbol{e}_{i_k}$. Furthermore, we denote $\theta = [\lambda, \boldsymbol{s}]$ . With the notations above, we could calculate the possibility that $y_k = 1, 0, -1$ as follows:

$$\begin{aligned} P\{y_k = 1\} &= P\{\epsilon_k > \lambda - s_{i_k} + s_{j_k}\} = 1 - \Phi\Big([1, \boldsymbol{x}_k^\top]^\top \boldsymbol{\theta}\Big), \\ P\Big\{y_k = 0\Big\} &= P\{-\lambda - s_{i_k} + s_{j_k} < \epsilon_k \le \lambda - s_{i_k} + s_{j_k}\}, \\ &= \Phi\Big([1, \boldsymbol{x}_k^\top]^\top \boldsymbol{\theta}\Big) - \Phi\Big([-1, \boldsymbol{x}_k^\top]^\top \boldsymbol{\theta}\Big), \\ P\{y_k = -1\} &= P\{\epsilon_k \le -\lambda - s_{i_k} + s_{j_k}\} = \Phi\Big([-1, \boldsymbol{x}_k^\top]^\top \boldsymbol{\theta}\Big). \end{aligned}$$

Therefore:

$$P\{y_k\} = \prod_{label \in \{-1,0,1\}} \Big[ P\{y_k = label\} \Big]^{1\{y_k = label\}}.$$

Given all above, it is easy to write out the negative log-likelihood via denoting $\zeta_k^+$ as $[1, \boldsymbol{x}_k^\top]^\top \boldsymbol{\theta}$ and $\zeta_k^-$ as $[-1, \boldsymbol{x}_k^\top]^\top \boldsymbol{\theta}$:

$$\begin{aligned} \ell(y|\boldsymbol{s}, \lambda) = -\sum_k \Big( & 1\{y_k = 1\} log\Big[1 - \Phi(\zeta_k^+)\Big] \\ & + 1\{y_k = 0\} log\Big[\Phi(\zeta_k^+) - \Phi(\zeta_k^-)\Big] + 1\{y_k = -1\} log\Big[\Phi(\zeta_k^-)\Big]\Big). \end{aligned} \tag{3}$$

To solve our proposed model, one just needs to minimize $\ell(y|\boldsymbol{s}, \lambda)$ with respect to $(\lambda, \boldsymbol{s})$. Furthermore, if we assume that $\sum_i s_i = 0$, then we could replace $s_n$ with $-\sum_{i=1}^{n-1} s_i$. Correspondingly, we can rewrite the loss function $\ell(y|\boldsymbol{s}, \lambda)$ as a function of $(\lambda, s_1, ..., s_{n-1})$: $\ell(y|\boldsymbol{s}/s_n, \lambda)$ , where $\boldsymbol{s}/s_n \triangleq (s_1, ..., s_{n-1})$. Then, under Assumption 1 we could prove that Theorem 3.2 holds.

**Table 1: The definition of FDR and Power.**

| | Comparable | Incomparable |
|---|---|---|
| Detected as Comparable | $\mathcal{N}_{0,0}$ | $\mathcal{N}_{0,1}$ |
| Detected as Incomparable | $\mathcal{N}_{1,0}$ | $\mathcal{N}_{1,1}$ |

**Assumption** 1. *Define* $\phi(x)$ *as* $\Phi'(x)$*, we assume that at least one of the following assumptions holds for* $\Phi(x), \phi(x)$ *and* $\phi'(x)$*:*

a) $\phi'(x) \equiv 0$ *and* $\forall x, \phi(x) \neq 0$*;*

b) $\phi(x)$ *is an even function,* $\phi(x)$ *and* $\Phi(x) - \frac{\phi^2(x)}{\phi(x)}$ *is strictly decreasing on* $(0, +\infty)$*. Furthermore,* $\lim_{x \to +\infty} \frac{\phi^2(x)}{\phi'(x)} = 0, \phi'(x) \neq 0$ *for* $x \neq 0$*,* $\forall x, \phi(x) \neq 0$ *and* $\phi'(x) < 0$ *if* $x > 0$ *.*

**Theorem** 1. $\ell(y|\boldsymbol{s}/s_n, \lambda)$ *is strictly convex with respect to* $(\lambda, s_1, \cdots, s_{n-1})$*.*

It is easy to check all the three models satisfy Assumption 1, thus all these models are strictly convex.

Putting all these together, we conclude that the MLE of these models are just solutions of strictly convex problems which can be solved efficiently.

### 3.6 FDR and Power Control

In this part, we show the theoretical analysis of model performance on separating the incomparable pairs from the comparable ones. To measure this ability, we introduce two criteria: False Discovery Rate (FDR) and Power, as is shown in Table 1. The definition of FDR and Power in our setting are:

$$\begin{aligned} FDR &= \frac{\mathcal{N}_{1,0}}{\mathcal{N}_{1,0} + \mathcal{N}_{1,1}}, \\ Power &= \frac{\mathcal{N}_{1,1}}{\mathcal{N}_{0,1} + \mathcal{N}_{1,1}}. \end{aligned}$$

In the following, we will propose a conservative threshold bound to guarantee FDR to be 0 and an aggressive threshold bound to guarantee Power to be 1.

Let $\big(\{s_i^\star\}_{i=1}^n, \lambda^\star\big)$ be the corresponding true parameters, $\big(\{\hat{s}_i\}_{i=1}^n, \hat{\lambda}\big)$ be the corresponding estimated parameters returned by our proposed method. Denote $\delta$ as the maximum of the variance of the estimated model parameters i.e. $\delta = max(\sigma_{\hat{\lambda}}^2, \sigma_{\hat{s}_1}^2, \cdots, \sigma_{\hat{s}_n}^2)$. Furthermore, we denote $\hat{\delta}$ as the estimation of $\delta$ on the observed dataset and $\Delta = \frac{\sqrt{4log(n+1)\hat{\delta}}}{\sqrt{N}}$. With the notations above, we construct the set of all incomparable pairs as $\mathcal{M}$, a conservative set as $\widehat{\mathcal{M}}$ and the aggressive set as $\widetilde{\mathcal{M}}$ :

$$\mathcal{M} = \{(i,j) : |s_i^\star - s_j^\star| \le \lambda^*\}, \tag{4}$$

$$\widehat{\mathcal{M}} = \{(i,j) : |\hat{s}_i - \hat{s}_j| \le \hat{\lambda} - 3\Delta\}, \tag{5}$$

$$\widetilde{\mathcal{M}} = \{(i,j) : |\hat{s}_i - \hat{s}_j| \le \hat{\lambda} + 3\Delta\}, \tag{6}$$

where $N$ is the number of samples. Now we first propose a theorem which shows that with high probability, $\widehat{\mathcal{M}} \subseteq \mathcal{M} \subseteq \widetilde{\mathcal{M}}$, followed by a practical interpretation via the remark that comes right after the theorem.

**Theorem** 2. *Let $\theta = (\lambda, \boldsymbol{s})$. Then with probability at least $1 - 2(n+1)^{\frac{\delta - 2\hat{\delta}}{\delta}}$, we will have that $\widehat{\mathcal{M}} \subseteq \mathcal{M} \subseteq \widetilde{\mathcal{M}}$.*

**Remark** 1. *If $\widehat{\mathcal{M}} \subseteq \mathcal{M}$ occurs, we set the threshold $\lambda$ as $\underline{\lambda} = \hat{\lambda} - 3\Delta$. Then all the detected incomparable pairs are truly incomparable, thus FDR = 0 is guaranteed. Likewise, if $\mathcal{M} \subseteq \widetilde{\mathcal{M}}$ i.e. $\widetilde{\mathcal{M}}^c \subseteq \mathcal{M}^c$ occurs, we have $|\hat{s}_i - \hat{s}_j| > \hat{\lambda} + 3\Delta$ indicating $|s_i^\star - s_j^\star| > \lambda^\star$. Consequently, if we set the threshold as $\overline{\lambda} = \hat{\lambda} + 3\Delta$, then all the comparable pairs will be detected as comparable and thus Power = 1 is guaranteed.*

To evaluate $\overline{\lambda}$ and $\underline{\lambda}$, one must first evaluate $\hat{\delta}$. Next, we propose a method to estimate $\hat{\delta}$ with the well-known asymptotic normality of MLE [28]. First, according to Section 3.5, we know that $\ell(y|\boldsymbol{s}/s_n, \lambda)$ is strictly convex for all mentioned distributions. Denote $\tilde{I}((\hat{\lambda}, \hat{\mathbf{s}}/\hat{s}_n))$ as the estimated Fisher Information matrix, we have:

$$\tilde{I}((\hat{\lambda}, \hat{\mathbf{s}}/\hat{s}_n)) = -\nabla^2_{\lambda, \boldsymbol{s}/s_n} \left[ \ell(y|(\hat{\lambda}, \hat{\mathbf{s}}/\hat{s}_n))/N \right] \succ 0,$$

and

$$E\left[ \frac{-\nabla^2_{\lambda, \boldsymbol{s}/s_n} \ell(y|\boldsymbol{s}^\star / s_n^\star, \lambda^\star)}{N} \right] = I((\lambda^\star, \boldsymbol{s}^\star / s_n^\star)) \succ 0,$$

where $I((\lambda^\star, \boldsymbol{s}^\star / s_n^\star))$ is the true Fisher Information matrix. Hence, these two matrices are invertible while the inversion has positive diagonal elements. Accordingly we have:

$$I^{-1}((\lambda^\star, \boldsymbol{s}^\star / s_n^\star))_{1,1} = \sigma^2_{\hat{\lambda}};$$
$$I^{-1}((\lambda^\star, \boldsymbol{s}^\star / s_n^\star))_{i,i} = \sigma^2_{\hat{s}_i}, \quad \forall i = 1, ..., n-1.$$

Then, we could estimate the variances as:

$$\hat{\sigma}^2_{\hat{\lambda}} \triangleq \tilde{I}^{-1}((\hat{\lambda}, \hat{\mathbf{s}}/\hat{s}_n))_{1,1};$$
$$\hat{\sigma}^2_{\hat{s}_i} \triangleq \tilde{I}^{-1}((\hat{\lambda}, \hat{\mathbf{s}}/\hat{s}_n))_{i,i}, \quad \forall i = 1, ..., n-1;$$
$$\hat{\sigma}^2_{\hat{s}_n} \triangleq (0, 1, 1, .., 1)\tilde{I}^{-1}((\hat{\lambda}, \hat{\mathbf{s}}/\hat{s}_n))(0, 1, 1, ..., 1)^\top.$$

Similarly, we can estimate $\hat{\delta}$ as:

$$\hat{\delta} = max\{\hat{\sigma}^2_{\hat{\lambda}}, \hat{\sigma}^2_{\hat{s}_1}, ..., \hat{\sigma}^2_{\hat{s}_n}\}.$$

## 4 EXPERIMENTS

In this section, four examples are exhibited with both simulated and real-world data to illustrate the validity of the analysis above and applications of the methodology proposed. The first example is with simulated data while the latter three exploit real-world data collected by crowdsourcing.

### 4.1 Simulated Study

**Settings** We validate the proposed algorithm on simulated data with $n = |V| = 20$ labeled by users. Specifically, we first randomly create a global ranking score $\boldsymbol{s}^\star \sim 10 \times N(0,1)$ as the ground-truth for $n$ candidates. Then pairwise comparisons are generated by Bradley-Terry model, i.e. $y_{i,j} = 1$ with probability $\left\{\frac{exp(s_i^\star - s_j^\star - \lambda^\star)}{1+exp(s_i^\star - s_j^\star - \lambda^\star)}\right\}$, $y_{i,j} = 0$ with probability $\left\{\frac{exp(s_i^\star - s_j^\star + \lambda^\star)}{1+exp(s_i^\star - s_j^\star + \lambda^\star)} - \frac{exp(s_i^\star - s_j^\star - \lambda^\star)}{1+exp(s_i^\star - s_j^\star - \lambda^\star)}\right\}$, and $y_{i,j} = -1$ with probability $\left\{\frac{1}{1+exp(s_i^\star - s_j^\star + \lambda^\star)}\right\}$. Here we set $\lambda = 0.5 : 0.5 : 2$. Finally, we obtain a dataset with 10000 samples. The experiments are repeated 20 times and ensemble statistics for the estimator are recorded.

**Table 2: Experimental results of 3 models on simulated data ($\lambda$ = 1).**

(a) Macro-F1

| | min | mean | max | std |
|---|---|---|---|---|
| **Uniform** | 0.7842 | **0.8454** | 0.9632 | 0.0437 |
| **Bradley-Terry** | 0.8309 | **0.9794** | 1.0000 | 0.0265 |
| **Thurstone-Mosteller** | 0.8747 | **0.9679** | 1.0000 | 0.0312 |

(b) Micro-F1

| | min | mean | max | std |
|---|---|---|---|---|
| **Uniform** | 0.7872 | **0.8611** | 0.9677 | 0.0389 |
| **Bradley-Terry** | 0.8214 | **0.9803** | 1.0000 | 0.0263 |
| **Thurstone-Mosteller** | 0.8908 | **0.9749** | 1.0000 | 0.0260 |

**Table 3: Experimental results of 3 models on simulated data as $\lambda$ varies ($\lambda$ = 0.5, 1, 1.5, 2).**

(a) Macro-F1

| $\lambda$ | 0.5 | 1 | 1.5 | 2 |
|---|---|---|---|---|
| **Uniform** | 0.8017 | 0.8454 | 0.8369 | 0.8068 |
| **Bradley-Terry** | 0.9753 | 0.9794 | 0.9761 | 0.9818 |
| **Thurstone-Mosteller** | 0.9628 | 0.9679 | 0.9714 | 0.9727 |

(b) Macro-F1

| $\lambda$ | 0.5 | 1 | 1.5 | 2 |
|---|---|---|---|---|
| **Uniform** | 0.8520 | 0.8611 | 0.8273 | 0.7987 |
| **Bradley-Terry** | 0.9794 | 0.9803 | 0.9761 | 0.9814 |
| **Thurstone-Mosteller** | 0.9822 | 0.9749 | 0.9710 | 0.9704 |

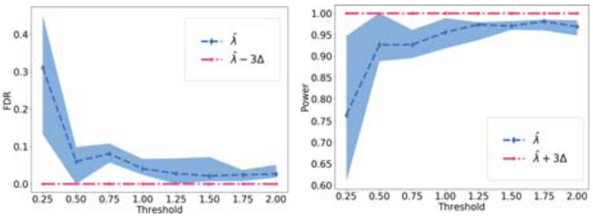
**Evaluation metrics** We measure the experimental results via two evaluation criteria, i.e., Macro-F1, and Micro-F1, which take both precision and recall into account. The larger the value of Micro-F1 and Macro-F1, the better the performance. More details about the evaluation metric please refer to [39].

**Results** Table 2(a) and 2(b) show the Macro-F1 and Micro-F1 of three models with $\lambda = 1$. Since the observed dataset is generated from the Bradley-Terry model, it obtains the best performance in terms of both metrics. Moreover, we also show the experimental results as $\lambda$ varies in Table 3(a) and 3(b), and it is easy to find that Bradley-Terry model again exhibits the best performance in most cases. In the following real-world datasets, we will also show the experimental results of Bradley-Terry Model.

**Validation of the FDR and Power guarantee.** To demonstrate the correctness of Theorem 2, we plot the FDR and Power results in Figure 2 for $\lambda = 0.25 : 0.25 : 2$ when $\hat{\lambda}$, $\hat{\lambda} - 3\Delta$ and $\hat{\lambda} + 3\Delta$ are employed as the estimated threshold, respectively. From the results we can easily find that, when $\hat{\lambda} - 3\Delta$ is employed as the estimated threshold, the FDR could always reach 0; while $\hat{\lambda} + 3\Delta$ could ensure the Power to be 1. This observation effectively demonstrates the correctness of the constructed conservative/aggressive set for FDR/Power.

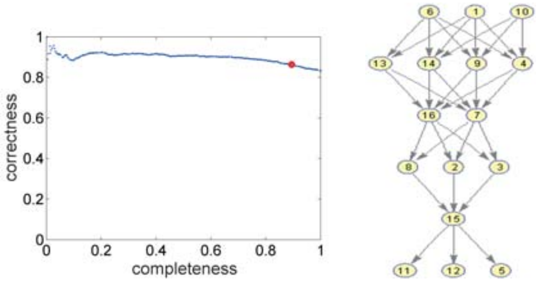
### 4.2 Image Quality Assessment

**Dataset Description** The second dataset is for subjective image quality assessment (IQA), which contains 15 reference images and 15 distorted versions of each reference, for a total of 240 images which come from two publicly available datasets LIVE, [29] and IVC [18]. Totally, 342 observers, each of whom performs a varied number of comparisons via Internet, provide 52,043 feedbacks (i.e.,



(a) FDR (b) Power

**Figure 2: An illustration of FDR and Power control.**



(a) Optimal $\lambda$ (b) Partial ranking

**Figure 3: Experimental results of Bradley-Terry model on IQA dataset.**

1, 0, -1) for crowdsourced subjective image quality assessment. For simplicity, we randomly take reference 1 as an illustrative example while other reference images exhibit similar results.

**Competitors** Now we introduce the competitors employed in our experiments. As mentioned in the Section 2, the $\alpha$-cut algorithm shares the most similar problem setting with our proposed algorithm and thus is adopted as our main competitor. Seeing that the $\alpha$-cut algorithm employs bagging ensembles of weak learners, we further compare our proposed algorithms with $\alpha$-cut algorithm when different types of such weak learners are adopted.

**Experiment Setting** Different from simulated data, as there are no ground-truth in real-world data, one can not compute Macro-F and Micro-F as in simulated data to evaluate the method we proposed. To see whether our proposed method could provide precise partial ranking, we generate 20 repetitions of training/testing splits with 80% of the samples are selected as the training set and the rest as the testing set. Regarding the parameter-tuning of the weak learners in $\alpha$-cut, we tune the coefficient for Ridge/LASSO regularization from the range $\{2^{-7}, 2^{-6}, \cdots, 2^{-3}\}$ and the best parameter is selected via a 5-fold cross-validation on the training set.

**Evaluation metrics** To test whether the edges we added in the graph are reasonable or not, we employ two metrics called correctness and completeness, respectively. Given the true partial order relation $\succ_*$, the estimated partial order relation $\succ$, the concordant set:

$$C = \{(i,j) : (i \succ j \wedge i \succ_* j) \vee (j \succ i \wedge j \succ_* i)\}$$

and discordant set

$$\mathcal{D} := \{(i,j) : (i \succ j \wedge j \succ_* i) \vee (j \succ i \wedge i \succ_* j)\}$$

**Table 4: Experimental results on IQA dataset.**

| types | algorithms | correctness | | completeness | | geomean | |
|---|---|---|---|---|---|---|---|
| | | median | std | median | std | median | std |
| $\alpha$-cut [4] | LRLASSO | 0.9137 | 0.0173 | 0.8309 | 0.0325 | 0.8760 | 0.0200 |
| | LRRidge | 0.9227 | 0.0150 | 0.8044 | 0.0301 | 0.8582 | 0.0148 |
| | SVMLASSO | 0.9158 | 0.0137 | 0.8310 | 0.0297 | 0.8721 | 0.0166 |
| | SVMRidge | 0.9184 | 0.0099 | 0.8083 | 0.0484 | 0.8594 | 0.0246 |
| | LSLASSO | 0.9154 | 0.0117 | 0.8095 | 0.0285 | 0.8623 | 0.0146 |
| | LSRidge | 0.9139 | 0.0126 | 0.8218 | 0.0336 | 0.8668 | 0.0182 |
| | SVRLASSO | 0.9236 | 0.0119 | 0.7405 | 0.0291 | 0.8311 | 0.0167 |
| | SVRRidge | 0.9191 | 0.0145 | 0.7594 | 0.0386 | 0.8378 | 0.0187 |
| ours | **Uniform** | **0.9137** | 0.0107 | **0.8623** | 0.0142 | **0.8867** | 0.0081 |
| | **Bradley-Terry** | **0.9113** | 0.0124 | **0.9254** | 0.0141 | **0.9064** | 0.0082 |
| | **Thurstone-Mosteller** | **0.9146** | 0.0122 | **0.9077** | 0.0122 | **0.9084** | 0.0075 |

we could define a metric for completeness as :

$$completeness = \frac{|C| + |\mathcal{D}|}{|\{(i,j) : i \succ_* j \vee j \succ_* i\}|}.$$

It is easy to find that the completeness metric measures the ability to detect a comparable pair. Likewise, correctness is defined as follows:

$$correctness = \frac{|C|}{|C| + |\mathcal{D}|}.$$

According to the definition, we see that a higher correctness implies a more accurate prediction for the pairs which are detected as comparable. Actually, there is always a trade-off between these two criteria: correctness on the one side and completeness on the other side. An ideal learner is correct in the sense of making few mistakes, but also complete in the sense of abstaining rarely. In other words, the two criteria are conflicting: increasing completeness typically might as well come along with reducing correctness and vice versa. Here we plot the trade-off between completeness and correctness as $\lambda$ varies. After all, every $\lambda$ can induce a partial ranking. The partial ranking obtained by $\lambda$-cut of MLE is highlighted as red circle, as is shown in Figure 3(a).

**Performance Comparison** Table 4 shows the corresponding performance of our proposed algorithms and the $\alpha$-cut algorithms. In this table, the second column shows the weak learner and regularization term employed in $\alpha$-cut and three models proposed in our algorithm. Specifically, LR represents for logistics regression [6], SVM stands for the Support Vector Machine [8] method, LS stands for the method of least squares [6] while SVR stands for the Support Vector Regression [13] method. For regularization, we employ the Ridge [15] and LASSO [33] regularization term. In order to comprehensively aggregate the performance, an overall metric should be defined based on both criteria. This leads to our inclusion of the last column which records the corresponding statistics for the geometric mean of the two mentioned criteria. According to this table, we find that our proposed algorithms significantly outperform other competitors in terms of completeness, and reach comparable results in terms of correctness. Moreover, the advantage in terms of the third metric also suggests the comprehensive superiority of our proposed algorithms.

**Partial Order Visualization** Here Figure 3(b) depicts a diagram for the partial order induced by Bradley-Terry Model. Disconnected nodes in the diagram indicate the incomparability of their corresponding subjective quality. Take the fourth level (ID=8,2,3) as an example. These three images come from LIVE datasets [29], and the corresponding names in LIVE dataset are ID=8 (img91-jp2k.bmp),

Table 5: Experimental results on human age dataset.

| type | algorithms | correctness | | completemess | | geomean | |
|---|---|---|---|---|---|---|---|
| | | median | std | median | std | median | std |
| $\alpha$-cut[4] | LRLASSO | 0.8640 | 0.0095 | 0.8352 | 0.0974 | 0.8511 | 0.0562 |
| | LRRidge | 0.8693 | 0.0070 | 0.8467 | 0.0186 | 0.8584 | 0.0090 |
| | SVMLASSO | 0.8674 | 0.0084 | 0.8565 | 0.0315 | 0.8619 | 0.0144 |
| | SVMRidge | 0.8660 | 0.0076 | 0.8447 | 0.1049 | 0.8542 | 0.0597 |
| | LSLASSO | 0.8688 | 0.0072 | 0.8583 | 0.0265 | 0.8617 | 0.0128 |
| | LSRidge | 0.8681 | 0.0072 | 0.8513 | 0.0193 | 0.8556 | 0.0096 |
| | SVRLASSO | 0.8732 | 0.0087 | 0.7687 | 0.0380 | 0.8177 | 0.0188 |
| | SVRRidge | 0.8732 | 0.0082 | 0.7750 | 0.0229 | 0.8237 | 0.0118 |
| ours | **Uniform** | **0.8655** | 0.0056 | **0.8523** | 0.0098 | **0.8591** | 0.0056 |
| | **Bradley-Terry** | **0.8671** | 0.0061 | **0.8990** | 0.0070 | **0.8826** | 0.0042 |
| | **Thurstone-Mosteller** | **0.8682** | 0.0062 | **0.8949** | 0.0067 | **0.8816** | 0.0044 |

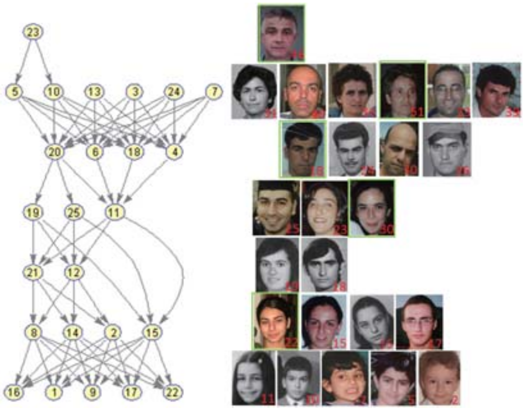


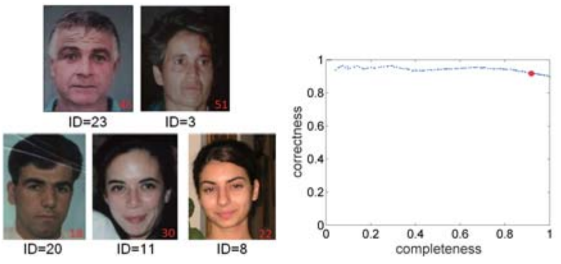
Figure 4: Partial ranking of human age dataset.

ID=2 (img95-fastfading.bmp), ID=3 (img91-fastfading.bmp). Via our proposed partial ranking algorithm, the quality of three images are treated as confusing thus located on the same level. To see whether they are really confusing or not, we go back to check the mean opinion score (MOS) of three images provided by LIVE dataset. We are pleasantly surprised to find that their MOS are so close: 50.96, 50.29, 48.62, respectively. From this viewpoint, the partial ranking we obtained is reasonable. However, MOS is not always accurate enough, which suffers from: i) Unable to concretely define the concept of scale; ii) Dissimilar interpretations of the scale among users; iii) Difficult to verify whether a participant gives false ratings either intentionally or carelessly. In this case, the results derived from our method could stand up to undertake the mission of being the ground-truth for image quality assessment.

## 4.3 Human Age

**Dataset Description** In this dataset, 25 images from human age dataset FG-NET [2] are annotated by a group of volunteer users on ChinaCrowds platform. The groundtruth age ranking is known to us. The annotator is presented with two images and given a choice of which one is older (or difficult to judge). Totally, we obtain 9589 feedbacks from 91 annotators.

**Performance Comparison** For age dataset, we adopt the same experiment setting, competitors and hyperparameter tuning strategy as the IQA dataset. Table 5 shows the comparable results on this dataset. Similar with the results on the IQA dataset, we can find that our proposed algorithms reach comparable performance in the sense of correctness. While, for the last two models (i.e. Bradley-Terry and Thurstone-Mosteller), our proposed algorithm significantly outperforms the competitors in terms of completeness. This leverages a better geometric mean of our algorithm with the last two models.



(a) Conflict images (b) Optimal $\lambda$

Figure 5: Conflict images and optimal $\lambda$ of human age dataset.

**Partial Ranking Visualization** Moreover, Figure 4 (Left) shows the partial ranking we obtained with 7 hierarchical levels on this dataset. It is easy to see that ID=23, the oldest, stands on the first level, while ID=5,10,13,3,24,7 are on the second level, and so on. On the leaf nodes, individuals with ID=16,1,9,17,22 are the youngest group of this dataset. To demonstrate whether the partial ranking we derived is reasonable or not, the original images are shown on the right panel, with ground truth ages painted red on the right corner of each image. From top to down, we can see that ID=23 (46 years old) is indeed older than most of the individuals on level 2 except ID=3 (51 years old). In other words, the partial ranking by mistake thinks 46 older than 51! If we look into the details of these two individuals, as is shown in Figure 5(a), the man with ID=23 gets more wrinkles, especially around his forehead and eyes, compared with the woman with ID=3. Besides, the man has white hair on his temples while the woman not. Another three conflicts happen on level 3 of ID=20 (18 years old), level 4 of ID=11 (30 years old), and level 6 of ID=8 (22 years old), respectively. We guess the reason behind lies in the three individuals have more or less the gap with their actual ages. For example, ID=20 looks older than he really is, while other two women (ID=11 and ID=8) look significantly younger than they really are. Especially the woman ID=8 with 22 years old looks even younger than other two girls (ID=14 and ID=2) who are actually 7 years younger than her. From this viewpoint, the partial ranking derived from our proposed method is reasonable. Moreover, the optimal $\lambda$ on this dataset is highlighted as red circle, as is shown in Figure 5(b).

## 4.4 WorldCollege Ranking

**Data Description** We now apply the proposed method to the worldCollege ranking dataset, which is composed of 261 colleges. Using the Allourideas crowdsourcing platform, a total of 340 distinct annotators from various countries (e.g., USA, Canada, Spain, France, Japan, China, etc.) are shown randomly with pairs of these colleges, and asked to decide which of the two universities is more attractive to attend. If the voter thinks the two colleges are incomparable, he can choose the third option by clicking “I can’t decide".

[2] http://www.fgnet.rsunit.com/

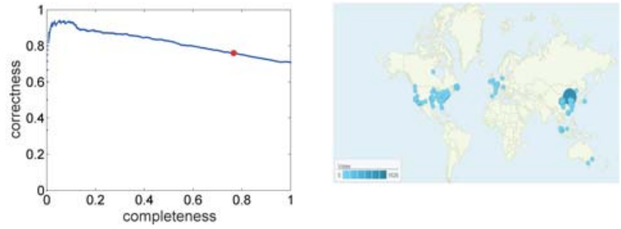
Table 6: Experimental results on worldCollege dataset.

| types | algorithms | correctness | | completeness | | geomean | |
|---|---|---|---|---|---|---|---|
| | | median | std | median | std | median | std |
| $\alpha$-cut [4] | LRLASSO | 0.5100 | 0.0121 | 1.0000 | 0.0412 | 0.7135 | 0.0120 |
| | LRRidge | 0.7439 | 0.0139 | 0.7475 | 0.0639 | 0.7391 | 0.0285 |
| | SVMLASSO | 0.5090 | 0.0091 | 1.0000 | 0.0012 | 0.7135 | 0.0063 |
| | SVMRidge | 0.7488 | 0.0150 | 0.7531 | 0.0655 | 0.7448 | 0.0297 |
| | LSLASSO | 0.5090 | 0.0091 | 1.0000 | 0.0000 | 0.7135 | 0.0064 |
| | LSRidge | 0.7490 | 0.0117 | 0.6518 | 0.0699 | 0.7020 | 0.0325 |
| | SVRLASSO | 0.5090 | 0.0091 | 1.0000 | 0.0000 | 0.7135 | 0.0064 |
| | SVRRidge | 0.7463 | 0.0151 | 0.7394 | 0.0671 | 0.7373 | 0.0300 |
| ours | **Uniform** | **0.7557** | 0.0101 | **0.7478** | 0.0104 | **0.7501** | 0.0074 |
| | **Bradley-Terry** | **0.7629** | 0.0108 | **0.7566** | 0.0087 | **0.7583** | 0.0069 |
| | **Thurstone-Mosteller** | **0.7619** | 0.0110 | **0.7586** | 0.0082 | **0.7576** | 0.0066 |

Finally, we obtain a total of 11012 feedbacks, among which 9409 samples are pairwise comparisons with clear opinions and the remaining 1603 are records with voter clicking "I can't decide".

**Performance Comparisons** Table 6 shows the comparable results on the college dataset. It is easy to see that our proposed algorithms again attain better correctness than all the $\alpha$-cut variants. Moreover, we find that all the LASSO-based $\alpha$-cut variants exhibit almost perfect completeness. Nonetheless, this superiority on completeness comes at a fatal price: the corresponding correctness results are close to 0.5, a value for a random ranker. Having perfect completeness alone thus does not make LASSO variants the top rankers. Consequently, in view of the aggregated metric, we see that all the LASSO variants show unreasonable performance on the third column while our proposed algorithms attain better comprehensive performance than $\alpha$-cut variants. Furthermore, compared to the two real-world datasets above, the performance on this dataset is a little bit worse than the IQA and human age datasets. We then go back to the crowdsourcing platform and find out that the reason behind lies in the "I can't decide" button. Though most voters click this button when he thinks two colleges are incomparable and difficult to choose, there are also some voters click this button because he does not know both of these two colleges or one of them. From this viewpoint, colleges with distinguishable difference even have the possibility to be treated as incomparable, just because the voters are not familiar with them. Due to the existence of these contaminated samples, though the performance of our proposed method declines by 10% approximately on this dataset, we still think it a reasonable phenomenon. Besides, the optimal $\lambda$ on this dataset is illustrated in Figure 6(a).

**Partial Order Visualization** Considering the partial ranking on 261 colleges is difficult to show, we only illustrate the partial ranking on top-20 colleges in Figure 7. It is easy to see that Yale, Princeton, and Harvard are the top 3 at the first level, while MIT, UC. Berkeley, Stanford, Cornell, UCLA are the second level. These results we derived are basically matched with the college ranking in reality. But a mystery has emerged from the experimental results. That is, Peking University (PKU) magically jumped into the third echelon together with Cambridge, Oxford, CMU, etc. To investigate the reason behind this phenomenon, we go back to see the world map of all the annotators. As is shown in Figure 6(b), most of the annotators come from China, thus significantly raises the ranking of PKU located in the capital of China—Beijing.



(a) Optimal $\lambda$ (b) World map

Figure 6: Optimal $\lambda$ and world map of all the annotators in worldCollege dataset.

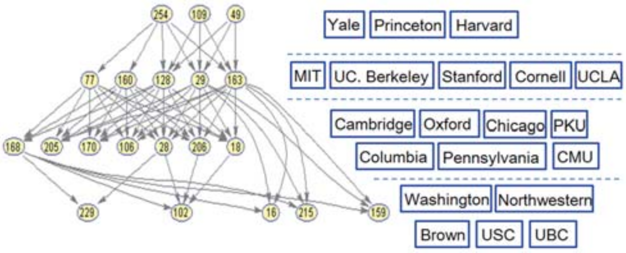


Figure 7: Partial ranking of worldCollege dataset.

## 5 CONCLUSIONS

In this paper, we propose a partial ranking algorithm based on margin-based MLE to learn partial but more accurate (i.e., making less incorrect statements) orders in crowdsourced ranking. In this scheme, three kinds of models are systematically discussed, including the uniform model, the Bradley-Terry model, and the Thurstone-Mosteller model. Moreover, we conduct theoretical analysis on FDR and Power control to demonstrate the effectiveness of the proposed method. Experimental studies conducted on simulated examples and three real-world datasets show that our proposed method could exhibit better performance compared with the traditional methods. Our results suggest that the proposed methodology is an effective tool to provide partial ranking for modern crowdsourced preference data.

## 6 ACKNOWLEDGMENTS

The research of Qianqian Xu was supported in part by National Key R&D Program of China (Grant No. 2016YFB0800403), National Natural Science Foundation of China (No.61672514, 61390514, 61572042), Beijing Natural Science Foundation (4182079), Youth Innovation Promotion Association CAS, and CCF-Tencent Open Research Fund. The research of Zhiyong Yang and Qingming Huang was supported in part by National Natural Science Foundation of China: 61332016, 61620106009, U1636214 and 61650202, in part by National Basic Research Program of China (973 Program): 2015CB351800, in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013. The research of Yuan Yao was supported in part by Hong Kong Research Grant Council (HKRGC) grant 16303817, National Basic Research Program of China (No. 2015CB85600, 2012CB825501), National Natural Science Foundation of China (No. 61370004, 11421-110001), as well as awards from Tencent AI Lab, Si Family Foundation, Baidu Big Data Institute, and Microsoft Research-Asia.

## REFERENCES

[1] K. Arrow. 1963. *Social Choice and Individual Values, 2nd Ed.* Yale University Press, New Haven, CT.

[2] Sergey Brin and Larry Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *International Conference on World Wide Web.* 107–117.

[3] Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei. 2009. A crowdsourceable QoE evaluation framework for multimedia content. In *ACM International Conference on Multimedia.* 491–500.

[4] Weiwei Cheng, Michaël Rademaker, Bernard De Baets, and Eyke Hüllermeier. 2010. Predicting partial orders: ranking with abstention. *Machine Learning and Knowledge Discovery in Databases* (2010), 215–230.

[5] C. Chow. 1970. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory* 16, 1 (1970), 41–46.

[6] M Bishop Christopher. 2006. *Pattern Recognition and Machine Learning.* Springer-Verlag New York.

[7] C. Cortes, M. Mohri, and A. Rastogi. 2007. Magnitude-preserving ranking algorithms. In *International Conference on Machine learning.* 169–176.

[8] N. Cristianini and J. Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge Unversity Press.

[9] D. Cynthia, K. Ravi, N. Moni, and S. Dandapani. 2001. Rank aggregation methods for the web. In *International Conference on World Wide Web.* 613–622.

[10] H. David. 1988. *The Method of Paired Comparisons.* Oxford University Press, New York, NY.

[11] J. de Borda. 1781. *Mémoire sur les Elections au Scrutin.* Histoire de l'Académie Royale des Sciences.

[12] Marquis de Condorcet. 1785. *Éssai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix (Essay on the Application of Analysis to the Probability of Majority Decisions).* Imprimerie Royale, Paris.

[13] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alexander J. Smola, and Vladimir Vapnik. 1996. Support Vector Regression Machines. In *Advances in Neural Information Processing Systems.* 155–161.

[14] Aristides Gionis, Heikki Mannila, Kai Puolamäki, and Antti Ukkonen. 2006. Algorithms for discovering bucket orders from data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 561–566.

[15] Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 1 (1970), 55–67.

[16] X. Jiang, L-H. Lim, Y. Yao, and Y. Ye. 2011. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming* 127, 6 (2011), 203–244.

[17] Jon Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *J. ACM* 46, 5 (1999), 604–632.

[18] Patrick Le Callet and Florent Autrusseau. 2005. Subjective quality assessment IRCCyN/IVC database. (2005). http://www.irccyn.ec-nantes.fr/ivcdb/.

[19] Guy Lebanon and Yi Mao. 2008. Non-parametric modeling of partially ranked data. *Journal of Machine Learning Research* 9 (2008), 2401–2429.

[20] W. Ma, J. M. Morel, S. Osher, and A. Chien. 2011. An $L_1$-based variational model for Retinex theory and its application to medical images. In *IEEE Conference on Computer Vision and Pattern Recognition.* 153–160.

[21] S. Negahban, S. Oh, and D. Shah. 2012. Iterative ranking from pair-wise comparisons. In *Annual Conference on Neural Information Processing Systems.* 2483–2491.

[22] G. Noether. 1960. Remarks about a paired comparison model. *Psychometrika* 25 (1960), 357–367.

[23] Braxton Osting, Jérôme Darbon, and Stanley Osher. 2013. STATISTICAL RANKING USING THE $L_1$-NORM ON GRAPHS. *Inverse Problems and Imaging* 7, 3 (2013), 907–926.

[24] Devi Parikh and Kristen Grauman. 2011. Relative attributes. In *IEEE International Conference on Computer Vision.* 503–510.

[25] Arun Rajkumar and Shivani Agarwal. 2014. A Statistical Convergence Perspective of Algorithms for Rank Aggregation from Pairwise Data. In *International Conference on Machine Learning.* 118–126.

[26] T. Saaty. 1977. A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology* 15, 3 (1977), 234–281.

[27] Thomas L Saaty and Mujgan S Ozdemir. 2003. Why the magic number seven plus or minus two. *Mathematical and computer modelling* 38, 3-4 (2003), 233–244.

[28] Mark J Schervish. 2012. *Theory of Statistics.* Springer Science & Business Media.

[29] H.R. Sheikh, Z.Wang, L. Cormack, and A.C. Bovik. 2008. LIVE Image & Video Quality Assessment Database. (2008).

[30] Yannis Sismanis. 2010. How I won the "Chess Ratings – Elo vs. the Rest of the World" Competition. *arxiv.org/abs/1012.4571v1* (2010).

[31] R. Stefani. 1977. Football and Basketball Predictions Using Least Squares. *IEEE Transactions on Systems, Man, and Cybernetics* 7 (1977), 117–121.

[32] L.L. Thurstone. 1927. A law of comparative judgement. *Psychological Review* 34 (1927), 278–286.

[33] R. Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58, 1 (1996), 267–288.

[34] Qianqian Xu, Qingming Huang, and Yuan Yao. 2012. Online crowdsourcing subjective image quality assessment. In *ACM International Conference on Multimedia.* 359–368.

[35] Qianqian Xu, Tingting Jiang, Yuan Yao, Qingming Huang, Bowei Yan, and Weisi Lin. 2011. Random partial paired comparison for subjective video quality assessment via Hodgerank. In *ACM International Conference on Multimedia.* 393–402.

[36] Qianqian Xu, Jiechao Xiong, Xiaochun Cao, and Yuan Yao. 2016. Parsimonious Mixed-Effects HodgeRank for Crowdsourced Preference Aggregation. In *ACM International Conference on Multimedia.* 841–850.

[37] S. Yu. 2009. Angular Embedding: From Jarring Intensity Differences to Perceived Luminance. In *IEEE Conference on Computer Vision and Pattern Recognition.* 2302–2309.

[38] S. Yu. 2012. Angular Embedding: A Robust Quadratic Criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 1 (2012), 158–173.

[39] Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26, 8 (2014), 1819–1837.