HodgeRank-based Analysis of Crowdsourced Data

Chen Liu
Department of Mathematics
cliudh@connect.ust.hk

Hangyu Lin
Department of Mathematics
hlinbh@connect.ust.hk

Abstract

As crowdsourcing gains popularity as a data collection method, ensuring the quality of such data becomes increasingly critical. This report addresses this issue specifically in the context of ranking data obtained through crowdsourcing. The report introduces HodgeRank as a methodology for analyzing pairwise comparison data and obtaining global rankings, as well as conducting inconsistency analysis. Additionally, the report discusses extensions of HodgeRank, such as the generalized linear model-based HodgeRank, and investigates positional bias among annotators. The report then presents experiments conducted on the WorldCollege and HumanAge datasets, utilizing various algorithms and examining annotator information through positional bias and Mixed Effect models. To ensure regularization and stable results, LBI and Knockoffs are also tested. Overall, the report offers valuable insights into the analysis of crowdsourced data and highlights the potential for future research in this area.

Contribution Remark:
Hangyu LIN: HodgeRank-GLM, Mixed Effect Model,
Chen LIU: LBI Regularization, Knockoffs

1 Introduction

Crowdsourcing has emerged as a popular method for collecting data and information in a wide range of fields, from business and marketing to social science and engineering. The use of crowdsourcing has grown rapidly in recent years, as advances in technology have made it easier to connect with large groups of people online. How to analyze such kind of data to gain information on both the items and the crowd becomes a popular and meaningful research topic. HodgeRank[2] was proposed to study pairwise comparison data which can provide global ranking and other inconsistency analysis. Specifically, a pairwise comparison dataset will be decomposed into gradient flow, harmonic flow, and curl flow which represent the globally acyclic information(gradient flow), locally acyclic and globally cyclic information(harmonic flow), and locally cyclic information(curl flow). All the above analysis is restricted to the information or relationship between the items. Xu[5] extend the original HodgeRank algorithm by using the generalized linear model to model the pairwise comparison matrix. In addition to studying the information of items, some researchers[7] try to find out whether there is some positional bias within some annotators like some annotators just click the left button. In [7], they proposed to incorporate Linearized Bregman Iteration(LBI)[3] and Knockoffs into HodgeRank to make the algorithm more stable and higher performance. Mixed Effect model[6] is developed to further study the user's preference effect which is a generalized model of positional bias model.

In this report, we implement algorithms and conduct experiments on WorldCollege and HumanAge datasets. First, we utilize the HodgeRank and generalized linear model-based HodgeRank to derive the global scores and ranking for colleges and humans in these two datasets. And the inconsistency analysis is achieved by HodgeRank algorithm. Furthermore, we investigate the information of annotators by using the position bias model and Mixed Effect model. Besides, LBI and Knockoffs are tested to find the regularization paths and stable results.

2 Methodology

In this section, we will introduce the methods or models we use to analyze the crowdsourced data. Given the comparison-based crowdsourced data, we will first use HodgeRank[2] to derive the global scores, harmonic flow, and curl flow for each item in the dataset. We will leverage different kinds of variants of HodgeRank based on some generalized linear models[5]. Rather than consider the item effect in the data, we will try to find the user/annotator effect in the data by using the Mixed Effect model[6]. In another way, some regularization methods[6, 8] or FDR control[7] can be used to make the model more compact and to find out the parsimonious path of multi-level models.

2.1 HodgeRank

First of all, we will introduce the HodgeRank method. HodgeRank is based on the Hodge decomposition, which is a mathematical concept from algebraic topology that decomposes a complex object into simpler parts. In the context of ranking, the key idea of the HodgeRank is the pairwise comparison data or edge flow can be decomposed into three parts,

$$Y = im(grad) \oplus ker(\Delta_0) \oplus im(curl^*). \tag{1}$$

Specifically, here we use Y_{ij}^u to represent that whether the u-user/annotator think i-item is better than j-item,

$$Y_{ij}^u = \begin{cases} 1 & \text{if } u\text{-user/annotator think } i \text{ is better than } j \\ -1 & \text{otherwise} \end{cases} \tag{2}$$

In a similar way, we define the weight matrix W_{ij}^u by,

$$W_{ij}^u = \begin{cases} 1 & \text{if } u\text{-user/annotator make a comparison between } ij \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

From the definition, we can know that Y^u is a skew-symmetric matrix and W^u is a symmetric matrix. The HodgeRank to find the global score s is to solve a weighted least square optimization problem,

$$\min \sum_{u,ij} W_{ij}^u (s_i - s_j - Y_{ij}^u)^2 \tag{4}$$

By summing over the u we can reformulate the problem into,

$$\min \sum_{ij} W_{ij} (s_i - s_j - Y_{ij})^2 \tag{5}$$

where $Y_{ij} = \sum_u W_{ij}^u Y_{ij}^u / \sum_u W_{ij}^u$, $W_{ij} = \sum_u W_{ij}^u$

After solving the problem and getting the s, we can derive the gradient flow as $grad(s)(ij) = s_i - s_j$. Then solving another projection problem, we can get the curl flow by $proj_{im(curl^*)}\hat{Y}$ and harmonic flow by $Y - grad(s) - proj_{im(curl^*)}\hat{Y}$

2.2 Generalized Linear Model

From the formulation above, if we define a_{ij} as the participants have a preference on i over j and $n_{ji} = a_{ij} + a_{ji}$, then we can get the preference probability by $\hat{\pi_{ij}} = a_{ij}/n_{ij}$. In this case, we have $Y_{ij} = 2 * \hat{\pi_{ij}} - 1$. If we formulate Y_{ij} as a function of $\hat{\pi_{ij}}$,

$$Y_{ij} = F^{-1}(\hat{\pi_{ij}}) \tag{6}$$

we can deploy some generalized linear model to get F

1. *Uniform* model:

$$Y_{ij} = 2 * \hat{\pi_{ij}} - 1 \tag{7}$$

2. *Bradley-Terry* model:

$$Y_{ij} = \log \frac{\hat{\pi_{ij}}}{1 - \hat{\pi_{ij}}} \tag{8}$$

3. *Thurstone-Mosteller* model:
$$Y_{ij} = F^{-1}(\hat{\pi_{ij}}) \tag{9}$$
where F is essentially the gauss error function
4. *Angular* model:
$$Y_{ij} = \arcsin(2 * \hat{\pi_{ij}} - 1) \tag{10}$$

By varying the target Y we can get different HodgeRank results with this generalized linear models.

2.3 Position Bias Model

For the ranking data, it is labeled by different annotators and the quality is not guaranteed. Some annotators may tend to quickly finish the labeling so they may just choose one side or choose one side by higher probability. To model such bias, we follow the work [7] to use the following linear model.

Linear Model for Position Bias Using the notation of previous section, we use Y_{ij}^u to denote the annotation by annotator u for the (ij) pair. We suppose that the annotation can be decomposed into the combination of the grad of global ranking $grad(s, i, j)$, which is defined as $s_i - s_j$ and the annotator's bias z_{ij}^u. And $z_{ij}^u = \gamma^u + \epsilon_{ij}^u$, γ^u is the positional bias and ϵ_{ij}^u is the random noise. Here the model can be formulated as follows,

$$Y = \delta_0 s + A\gamma + \epsilon \tag{11}$$

Here δ_0 is the matrix for grad and A is the matrix for positional bias. To find good estimation for s and γ, we can use the square loss $\frac{1}{2}\|Y - \delta_0 s - A\gamma\|_2^2$. But for the annotators, a natural assumption is that the majority of the annotators are reliable, so the ones with positional bias is the minority which means the γ should be sparse, so the problem us formulate as follows,

$$\min_{s,\gamma} \frac{1}{2}\|Y - \delta_0 s - A\gamma\|_2^2 + \lambda\|\gamma\|_1 \tag{12}$$

Linearized Bregman Iteration Regularization To solve the problem in equation. 21, a natural way is to use Lasso [4]. But the estimation of Lasso is biased, here we follow the work [7] to use Linearized Bregman Iteration(LBI) [3] to solve the problem. Here we have the following dynamics,

$$\begin{aligned} \frac{dp}{dt} &= A^T(Y - \delta_0 s - A\gamma) && (13) \\ 0 =& \delta_0^T(Y - \delta_0 s - A\gamma) && (14) \\ p \in& \partial\|\gamma\|_1 && (15) \end{aligned}$$

For such a dynamics, important feature tend to become nonzero more quickly. In practice, we use the discretization of this dynamics, we update the parameters as follows,

$$\begin{aligned} w^{t+1} &= w^t + \alpha A^T(Y - \delta_0 s - A\gamma) && (16) \\ \gamma^{t+1} &= \kappa shrink(w^{t+1}) && (17) \\ s^{k+1} &= s^k + \kappa\alpha\delta_0^T(Y - \delta_0 s - A\gamma) && (18) \end{aligned}$$

Here w is the augment variable and $\alpha$is the step size. κ is the hyperparameter larger κ can make discretization close to the dynamics. And $shrink(x) = sign(x)max(|x| - 1, 0)$ is the proximal mapping function. Here we explore the positional bias of annotators by viewing the regularization path of γ.

FDR Control Via Knockoffs Here one problem is when to stop the iteration, as shown in the work [3], early stopping is a powerful tool for feature selection which is verified theoretically and practically. However, early stopping needs to use cross validation. Here, if we can use an adaptive way to control the false discovery with no need to use cross validation. Follow the work [7], we use knockoffs to control the sparsity. Here, we use the difference between the time for element of γ become nonzero and the time for element of constructed $\tilde{\gamma}$ become nonzero to conduct FDR control.

2.4 Mixed Effect Model

In addition to the Position Bias Model, we can further separate the preference of each user/annotator and the common preference of some items.

$$Y_{ij}^u = (\phi_i^T \eta + \phi_i^T \xi^u) - (\phi_j^T \eta + \phi_j^T \xi^u) + \gamma^u + \varepsilon_{ij}^u \tag{19}$$

or in matrix form

$$Y = d\Phi\eta + X\beta + \varepsilon \tag{20}$$

the γ will be included into β, then following the same argument, the problem will be formulated as follows,

$$\min_{s,\eta,\beta} \frac{1}{2}\|Y - d\Phi\eta - X\beta\|_2^2 + \lambda(\|\gamma\|_1 + \sum_u \|\xi^u\|_1) \tag{21}$$

3 Main Results

3.1 Real World Data

First of all, we introduce the datasets we used in further study.

WolrdCollege dataset, which is composed of 261 colleges, is collected on the Allourideas crowdsourcing platform. 400 distinct annotators from various countries are given a pair of universities and asked to choose which university is more attractive to attend. Finally, a total of 9,408 pairwise comparisons are collected, and 8,544 pairwise comparisons are valid among the total data. There is no ground-truth ranking for these colleges.

Human Age dataset comprises 30 images from the FG-NET 1 dataset, which have been annotated by 94 volunteer users on ChinaCrowds platform. Each annotator was presented with two images and asked to choose which one appears older. A total of 14,011 pairwise comparisons were collected in this way. The ground-truth age ranking for the images is already known. Therefore, the dataset contains pairwise annotations of human age, and it has been obtained through a crowdsourcing approach.

3.2 Global Ranking Results

For crowdsourcing pairwise comparison datasets, the first question should be what is the global ranking of the items used in the comparison, e.g., each person in the Human Age, college in WolrdCollege. Here, we use several metrics to evaluate the performance of the ranking algorithm like HodgeRank, and generalized linear model-based HodgeRank. As shown in Tab.1, we measure the similarity of the predicted ranking and the ground-truth ranking using the Kendall τ which measures the mismatch rate between two rankings. It should be noticed that there is no actual ground-truth ranking for WorldCollege, so we leverage the score-based ranking as a reference to measure our algorithms. On the Human Age dataset, we can find the Bradley-Terry model and Thurstone-Mosteller model significantly improve the performance of the global ranking while the Angular-based model does not work well. However, compared to the score-based ranking, we can find almost all models have similar performance and even worse results when applying generalized linear models. It may be caused by that the score-based ranking is not ground truth ranking and it is not so reliable to measure the performance of algorithms.

Kendall τ	Uniform	BT	TM	Angular
WorldCollege	0.9204	0.9172	0.9183	0.9198
Human Age	-0.0896	0.1586	0.1908	-0.0758

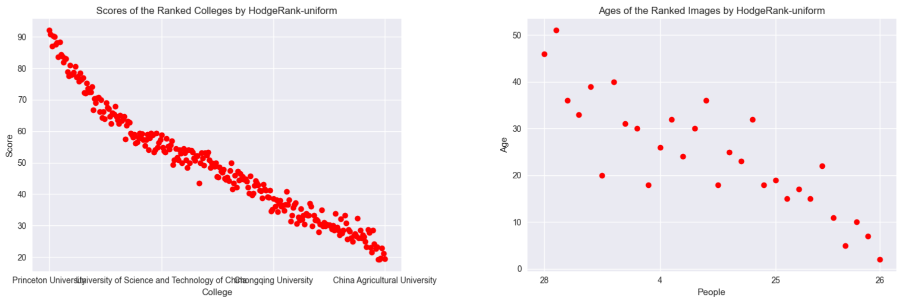
Table 1: The Kendall τ value of HodgeRank-GLM ranking w.r.t score-based ranking on WorldCollege and HodgeRank-GLM ranking w.r.t groundtruth age ranking on HumanAge.

To evaluate the ranking performance of the algorithms better, we calculate another test error metric, where the ratio of the mismatch on a test comparison set will be used to measure the performance. In practice, we randomly split the whole pairwise comparison set into 70%/30% train/test sets and apply algorithms on the train set to get a global score/ranking then compute the test error on the test

Test Error	Uniform	BT	TM	Angular
WorldCollege	0.2985(±0.0070)	0.2986(±0.0094)	0.2991(±0.0057)	0.3007(±0.0085)
Human Age	0.1826(±0.0047)	0.2139(±0.0090)	0.1965±(0.0065)	0.1835(±0.0055)

Table 2: The test error values of HodgeRank-GLM on WorldCollege and HumanAge. Randomly split the whole pairwise comparison votes into 70%/30% train/test datasets 20 times and compute the average and standard deviation of test errors.

set. For fairness, we repeat this process 20 times and compute the average and standard deviation. From Tab.2, we can find generalized linear models improve the performance on Human Age and WorldCollege consistently.



(a) The Scores of the Ranked Colleges by HodgeRank.

(b) The Ages of the Ranked Images by HodgeRank.

Figure 1: Scatter Plot of the Ranking Results.

Furthermore, we visualize the ranking result by scatter plots, where the x-axis represents the ranked items (college and human id), and the y-axis represents the scores and ages. From a global view, we can find the ranking result of the WorldCollege is better than HumanAge where the scatter plot has a linear shape. In fact, making a decision on which person is more younger or older from images is a harder task than give a decision on which college is better since the appearance of age is affected by many other effects not only age. And it is clear that the ranking result based on the comparison not only depends on the items but also the annotators and we will to some analysis for annotator bias in the following sections.

3.3 Inconsistency Analysis

Rather than the global ranking results, from the Hodge theory, we know there are still harmonic flow(global cyclic, local acyclic) and curl flow(local cyclic) which indicate the inconsistency in the graph. Following the analysis in [2], we compute the *cyclicity ratio*$C_p = ||R*||_w^2/||\overline{Y}||_w^2$ of HodgeRank results on WorldCollege and Human Age. The cyclicity ratio is 0.37 for WorldCollege and 0.19 for HumanAge. In another way, we find out the triangles in both datasets with the highest curl score. On the WorldCollege dataset, the id of the college just indicates the score-based ranking (which to some extent represents the true ranking of the college). In Tab.3, the triangles all have a significant inconsistent edge, like (80-153) in (153,135,80), (80-162) in (162, 124, 80). Similar results can find in Tab.4 on the HumanAge dataset that there is a significant inconsistent edge in a triangle (0,9,25) where the age of them is (10, 7, 2). In addition, we can find the 25th image which is an image of a 2-year-old child occurs frequently in the inconsistent triangles. It means that this image may be hard to tell age for many people.

3.4 Regularization & FDR Control

In this section, we apply the LBI method to identify abnormal annotators and analyze the selected annotators. We follow the notation and formulation presented in Section 2.3 to identify abnormal annotators and conduct normalization before the iteration. To verify whether the selection of the LBI [3] regularization path is meaningful, we record the evolution of the augment variable Γ during

Triangle Φ	curl Φ	Pred Scores
(153, 135, 80)	53.3880	(0.0311, -0.0470,0.1657)
(162, 124, 80)	44.6433	(-0.1533, 0.0965, 0.1657)
(135, 249, 243)	39.9422	(-0.0470, -0.4995, -0.4476)
(80, 239, 135)	39.7352	(0.1657, -0.3993, -0.0470)
(67, 153, 142)	37.3203	(0.1607, 0.0311, -0.0359)

Table 3: Top-10 curl Φ values triangles in the WolrdCollege Dataset.

Triangle Φ	curl Φ	True Ages	Pred Scores
(0, 9, 25)	1.5776	(10, 7, 2)	(-0.6658, -0.7024, -0.7674)
(15, 9, 25)	1.4924	(17, 7, 2)	(-0.2783, -0.7024, -0.7674)
(14, 9, 25)	1.4523	(15, 7, 2)	(-0.2780, -0.7024, -0.7674)
(9, 25, 1)	1.4000	(7, 2, 15)	(-0.7024, -0.7674, -0.3354)
(27, 2, 20)	1.3980	(46, 51, 20)	(0.7823, 0.6620, 0.4333)

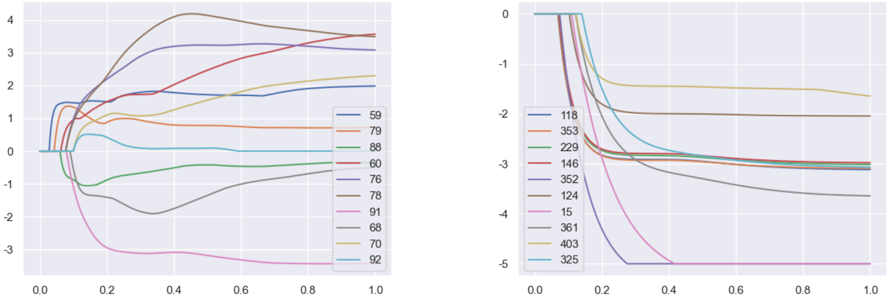
Table 4: Top-10 Φ values triangles in the Human Age Datasets.

the algorithm and pick the 30 annotators that appear earlier during the iteration. We present the detailed selection bias of these annotators in Table 5 and Table 6 for the Age and College datasets, respectively. We also visualize the regularization path for the Age and College datasets in Figure.2(a) and Figure.2(b), respectively. We note that we do not use the optimization procedure presented in Osher et al. [3], which involves a sequence of non-negative least squares. Instead, we optimize the loss function iteratively, and the resulting curve may not have the piece-wise constant shape as illustrated in their work. By choosing a moderate value of κ and reducing the step size α, we can approximate the inverse scale space dynamics, but this requires more iterations. Therefore, we choose a moderate value of κ for fast verification. By examining Table 5, we observe that some annotators with strong positional preferences are selected. For example, annotator 39 selects all 40 samples as left-click, and annotator 50 chooses left for all 63 choices. Other annotators also exhibit significant positional preference. For the College dataset shown in Table 6, we observe both strong left and right preferences. For instance, annotator 189 selects the left option for all 35 pairs. However, some preferences are due to a lack of samples. For example, annotators 147 and 353 only score one pair, so their preferences are not statistically significant.

One potential issue with the exploration of inverse scale space is that some normal annotators may be falsely identified as abnormal during the optimization process. To address this issue, we adopt a technique proposed in the work of Xu et al. [7] and use Knockoffs [1] to control the False Discovery Rate (FDR) in expectation. To evaluate the effectiveness of our method in controlling the FDR, we visualize scatter plots for the Age and College datasets in Figures 3(a) and 3(b), respectively. In these plots, the x-axis and y-axis represent the left-click count and right-click count, respectively. The blue points represent annotators that were not selected by our algorithm, while the red points represent those that were selected. By examining Figure 3(a), we observe that some annotators that strongly prefer the left side are selected as abnormal annotators. Additionally, most selections show a positional preference. For the College dataset shown in Figure 3(b), we enlarge the markers for abnormal annotators since many of them have fewer samples. We also observe that annotators selected by our algorithm tend to exhibit a significant preference for either the left or right side.

3.5 Preference Analysis

Based on the Mixed Effect model, we can analyze the positional bias and annotator preference. In practice, the positional bias achieved by Mixed Effect model is almost the same as the positional bias model. So we only focus on the preference analysis here. For both HumanAge and WorldCollege datasets, the absolute value of the effect of each user is relatively small than the common value to each item like $0.19/1.0$. So the personal preference in these two datasets maybe not so significant.



(a) Path of 10 ealry annotators in Age dataset. (b) Path of 10 ealry annotators in College dataset.

Figure 2: Path Visualization for two dataset

Annotator	Left Click	Right Click	Ratio
59	5	2	0.714286
79	4	6	0.400000
88	5	3	0.625000
60	21	13	0.617647
76	11	1	0.916667
78	35	20	0.636364
91	21	16	0.567568
68	19	23	0.452381
70	32	21	0.603774
92	6	13	0.315789
90	79	5	0.940476
4	71	79	0.473333
66	43	34	0.558442
2	24	23	0.510638
85	8	10	0.444444
30	15	10	0.600000
50	63	0	1.000000
44	30	20	0.600000
35	22	10	0.687500
39	40	0	1.000000

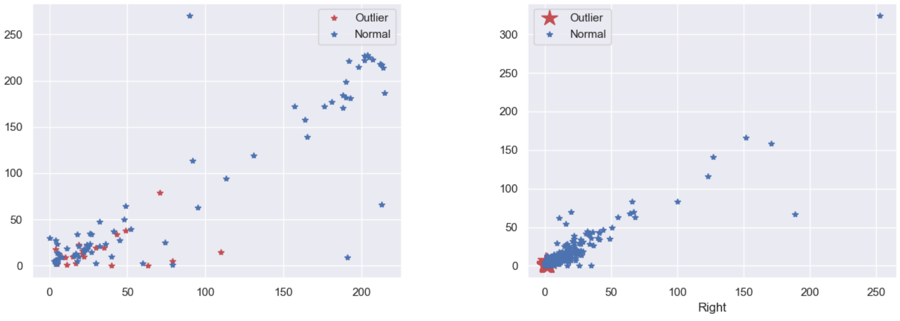
Table 5: Top 20 annotator on the path for Age data

4 Conclusion

In this study, we focus on the problem of ranking and employ a specific methodology to identify a global ranking while examining the curl component, positional bias, and preference of annotators. To validate our approach, we utilize two ranking datasets, namely Age and College. Our findings reveal that the quality of annotations heavily influences the outcome of global ranking. Specifically, for the Age dataset with relatively low-quality annotations, the global ranking is not satisfactory. We also conduct an inconsistency analysis on the Age dataset, which enables us to identify a local inconsistent triangle. Additionally, we successfully employ the LBI method to detect the positional bias of annotators. However, our analysis does not reveal any significant evidence of preference.

Annotator	Left Click	Right Click	Ratio
119	4	1	0.800000
354	9	6	0.600000
230	11	9	0.550000
147	1	0	1.000000
353	0	1	0.000000
125	2	1	0.666667
16	19	10	0.655172
362	1	0	1.000000
404	1	1	0.500000
326	3	5	0.375000
101	66	83	0.442953
329	0	1	0.000000
54	2	3	0.400000
306	1	3	0.250000
60	4	0	1.000000
99	22	24	0.478261
405	2	5	0.285714
385	5	5	0.500000
189	35	0	1.000000
87	26	18	0.59090

Table 6: Top 20 annotator on the path for College data



(a) Scatter for annotators in Age dataset.

(b) Scatter for annotators in College dataset

Figure 3: Scatter for two datasets

References

[1] Rina Foygel Barber and Emmanuel J Candes. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[2] Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.

[3] Stanley Osher, Feng Ruan, Jiechao Xiong, Yuan Yao, and Wotao Yin. Sparse recovery via differential inclusions. *Applied and Computational Harmonic Analysis*, 41(2):436–469, 2016.

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

[5] Qianqian Xu, Qingming Huang, Tingting Jiang, Bowei Yan, Weisi Lin, and Yuan Yao. Hodgerank on random graphs for subjective video quality assessment. *IEEE Transactions on Multimedia*, 14(3):844–857, 2012.

[6] Qianqian Xu, Jiechao Xiong, Xiaochun Cao, Qingming Huang, and Yuan Yao. From social to individuals: A parsimonious path of multi-level models for crowdsourced preference aggregation. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):844–856, 2018.

[7] Qianqian Xu, Jiechao Xiong, Xiaochun Cao, and Yuan Yao. False discovery rate control and statistical quality assessment of annotators in crowdsourced ranking. In *International conference on machine learning*, pages 1282–1291. PMLR, 2016.

[8] Qianqian Xu, Ming Yan, Chendi Huang, Jiechao Xiong, Qingming Huang, and Yuan Yao. Exploring outliers in crowdsourced ranking for qoe. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1540–1548, 2017.