Adversarial Top-K Ranking

Changho Suh, *Member, IEEE*, Vincent Y. F. Tan, *Senior Member, IEEE*, and Renbo Zhao, *Member, IEEE*

***Abstract*— We study the top-K ranking problem where the goal is to recover the set of top-K ranked items out of a large collection of items based on partially revealed preferences. We consider an *adversarial crowdsourced* setting where there are two population sets, and pairwise comparison samples drawn from one of the populations follow the standard Bradley-Terry-Luce model (i.e., the chance of item i beating item j is proportional to the relative score of item i to item j), while in the other population, the corresponding chance is inversely proportional to the relative score. When the relative size of the two populations is known, we characterize the minimax limit on the sample size required (up to a constant) for reliably identifying the top-K items, and demonstrate how it scales with the relative size. Moreover, by leveraging a tensor decomposition method for disambiguating mixture distributions, we extend our result to the more realistic scenario, in which the relative population size is unknown, thus establishing an upper bound on the fundamental limit of the sample size for recovering the top-K set.**

***Index Terms*— Adversarial population, Bradley-Terry-Luce model, crowdsourcing, minimax optimality, sample complexity, top-K ranking, tensor decompositions.**

I. Introduction

RANKING is one of the fundamental problems that has proved crucial in a wide variety of contexts—social choice [1], [2], web search and information retrieval [3], recommendation systems [4], ranking individuals by group comparisons [5] and crowdsourcing [6], to name a few. Due to its wide applicability, a large volume of work on ranking has been done. The two main paradigms in the literature include spectral ranking algorithms [3], [7], [8] and maximum likelihood estimation (MLE) [9]. While these ranking schemes yield reasonably good estimates which are faithful globally w.r.t. the latent preferences (i.e., low ℓ_2 loss), it is not necessarily guaranteed that this results in optimal ranking accuracy. Accurate ranking has more to do with how well the *ordering* of the estimates matches that of the true preferences (a discrete/combinatorial optimization problem), and less to do with how well we can estimate the true preferences (a continuous optimization problem).

In applications, a ranking algorithm that outputs a total ordering of all the items is not only overkill, but it also unnecessarily increases complexity. Often, we pay attention to only a *few* significant items. Thus, recent work such as that by Chen and Suh [10] studied the top-K identification task. Here, one aims to recover a correct set of top-ranked items only. This work characterized the minimax limit on the sample size required (i.e., the sample complexity) for reliable top-K ranking, assuming the Bradley-Terry-Luce (BTL) model [11], [12].

While this result is concerned with practical issues, there are still limitations when modeling other realistic scenarios. The BTL model considered in [10] assumes that the quality of pairwise comparison information which forms the basis of the model is the same across annotators. In reality (e.g., crowdsourced settings), however, the quality of the information can vary significantly across different annotators. For instance, there may be a non-negligible fraction of spammers who provide answers in an *adversarial* manner. In the context of *adversarial web search* [13], web contents can be maliciously manipulated by spammers for commercial, social, or political benefits in a robust manner. Alternatively, there may exist false information such as false voting in social networks and fake ratings in recommendation systems [14].

A. Our Model and Justifications for the Model

As an initial effort to address this challenge, we investigate a so-called *adversarial BTL* model, which postulates the existence of two sets of populations—the *faithful* and *adversarial* populations, each of which has proportion η and $1-\eta$ respectively. Specifically we consider a BTL-based pairwise comparison model in which there exist latent variables indicating ground-truth preference scores of items. In this model, it is assumed that comparison samples drawn from the faithful population follow the standard BTL model (the probability of item i beating item j is proportional to item i's relative score to item j), and those of the adversarial population act in an "opposite" manner, i.e., the probability of i beating j is inversely proportional to the relative score. See Fig. 1.

This model may, at a first glance, seem somewhat contrived and artificial. However, we believe that it is justified as an

Manuscript received February 15, 2016; revised October 11, 2016; accepted January 23, 2017. Date of publication January 26, 2017; date of current version March 15, 2017. C. Suh was supported by the National Research Foundation of Korea within MSIP through the Korean Government under Grant 2015R1C1A1A02036561. V. Y. F. Tan and R. Zhao were supported in part by the National University of Singapore (NUS) through the NUS Young Investigator Award under Award R-263-000-B37-133 and in part by MoE AcRF Tier 1 Grant under Grant R-263-000-C12-112. This paper was presented at the 2016 Information Theory and Applications Workshop.

C. Suh is with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, South Korea (e-mail: chsuh@kaist.ac.kr).

V. Y. F. Tan is with the Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore 117583, and the Department of Mathematics, NUS, Singapore 119076 (e-mail: vtan@nus.edu.sg)

R. Zhao is with the Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore 117583, the Department of Mathematics, NUS, Singapore 119076, and the Department of Industrial and Systems Engineering, NUS, Singapore 117576 (e-mail: elezren@nus.edu.sg).

Communicated by C. Caramanis, Associate Editor for Machine Learning.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2017.2659660

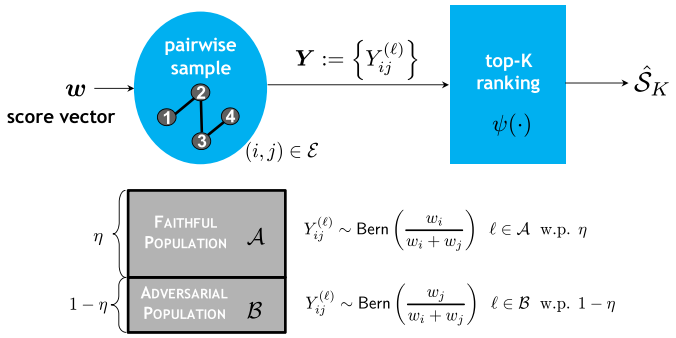

Fig. 1. Adversarial top-K ranking given samples $\boldsymbol{Y} = \{Y_{ij}^{(\ell)}\}$ where $(i, j) \in \mathcal{E}$ and $\mathcal{E}$ is the edge set of an Erdős-Rényi random graph.

initial step to study the fundamental limits of ranking in adversarial settings due to the following reasons.

- Nowadays, it is of paramount importance for retailers (like hotels and restaurants) to have outstanding recommendations of their establishments on the Web. For example, the websites Tripadvisor and Zagat (or Yelp) provide recommendation and ratings for hotels and restaurants respectively. Let us work with hotels. Suppose there are two hotels A and B owned by Alice and Bob respectively and they are competing for businesses in the same city, say $\mathfrak{C}$. All of the tourists who come to $\mathfrak{C}$ are nice and honest and provide decent ratings (more precisely, pairwise rankings) of both hotels and these rankings follow a standard BTL model parametrized by a latent vector $\boldsymbol{w}$. However, because Alice wants to beat Bob in attracting tourists to her hotel A, Alice hires spammers to flood Tripadvisor with "opposite" and indeed unfavorable ratings. These are generated from the opposite BTL model parametrized by $\boldsymbol{w}' = \text{fliplr}(\boldsymbol{w})$ (i.e., the preference vector is flipped). Now, because Alice wants to be circumspect in her egregious actions (to minimize the likelihood of being caught), she will not ask her spammers to say that B is terribly bad, corresponding to B having completely negative ratings all the time. Alice will ask spammers to write less-than-positive reviews (i.e., "flipped" reviews) about B on Tripadvisor so as to induce some subtle yet adversarial behavior into the system (parametrized by $\boldsymbol{w}'$), resulting in B having worse ratings than it should have. Now the machine learning task at hand is to be cognizant of the presence of spammers, uncover the true rankings of hotels in $\mathfrak{C}$, and possibly also to learn the proportion of spammers. In a more realistic scenario, there is a *network* of hotels represented by a sparse graph in which pairs of hotels that are linked in the graph try to undermine each others' businesses as for the pairwise scenario involving A and B.
- The other motivation is mathematical tractability: By having the adversarial population have the same (yet reordered) parameters compared to the faithful population, the number of parameters in the model is reduced (halved), hence the learning task becomes significantly more tractable (cf. Lemma 6). However, the complexity of the task of obtaining provable bounds, especially if one seeks globally optimal estimates for the unknown parameters (e.g., η), is still readily apparent from our analyses and results.

B. Main Contributions

We seek to characterize the fundamental limits on the sample size required for top-K ranking, and to develop computationally efficient ranking algorithms. There are two main contributions in this paper.

Building upon *RankCentrality* [7] and *SpectralMLE* [10], we develop a ranking algorithm to characterize the minimax limit required for top-K ranking, up to constant factors, for the η-known scenario. We also show the minimax optimality of our ranking scheme by proving a converse or impossibility result that applies to *any* ranking algorithm using information-theoretic methods. As a result, we find that the sample complexity is inversely proportional to $(2\eta - 1)^2$, which suggests that less distinct the population sizes, the larger the sample complexity. We also demonstrate that our result recovers that of the $\eta = 1$ case in [10], so the work contained herein is a strict generalization of that in [10].

The second contribution is to establish an upper bound on the sample complexity for the more practically-relevant scenario where η is unknown. A novel procedure based on tensor decomposition approaches in Jain and Oh [15] and Anandkumar *et al.* [16] is proposed to first obtain an estimate of the parameter η that is in a neighborhood of η, i.e., we seek to obtain an ε-globally optimal solution. This is usually not guaranteed by traditional iterative methods such as Expectation Maximization [17]. Subsequently, the estimate is then used in the ranking algorithm that assumes knowledge of η. We demonstrate that this algorithm leads to an order-wise worse sample complexity relative to the η-known case. Our theoretical analyses suggest that the degradation is unavoidable if we employ this natural two-step procedure.

Let us, at this point, comment informally on some of the mathematical complexities in the derivations of our results, specifically for the case in which η is unknown. After a careful analysis of a non-asymptotic sample complexity bound on disambiguating mixtures in [15, Th. 3], we deduce that for any $\varepsilon, \delta > 0$, with probability exceeding $1 - \delta$, there exists an algorithm (based on tensor decompositions) that produces an estimate of η, called $\hat{\eta}$, satisfying

$$|\hat{\eta} - \eta| \leq \varepsilon \tag{1}$$

if the sample size L satisfies

$$L = \Omega\left(\frac{1}{\varepsilon^2} \log \frac{n}{\delta}\right). \tag{2}$$

For a precise statement of claim, please see Lemma 5. Here we observe a subtle tradeoff. If we drive the estimation error $|\hat{\eta} - \eta|$ down (i.e., ε in (1) is small), then the ℓ_∞ error of the preference vector $\|\boldsymbol{w} - \hat{\boldsymbol{w}}\|_\infty$ (in the actual ranking algorithm) would also go down. However, according to (2), the sample size would then increase. On the other hand, if $|\hat{\eta} - \eta|$ is not small, then the sample size required decreases but the ranking accuracy, measured according to $\|\boldsymbol{w} - \hat{\boldsymbol{w}}\|_\infty$,

would be adversely affected. Thus, our analysis entails finding a "sweet spot" for the accuracy of learning η (for (1)) by balancing the two competing objectives (of sample size and ranking accuracy) to find the best achievable sample complexity using the current strategy. This, we feel, is the main non-trivial aspect of our analyses. Such a technique, while being extremely useful in our context, may also be applicable to other multi-stage estimation and learning tasks in machine learning as one needs to carefully "balance" the estimation quality of the first stage with the eventual learning outcome after all stages are completed.

C. Related Work

The most relevant related works are those by Chen and Suh [10], Negahban *et al.* [7], and Chen *et al.* [6]. Chen and Suh [10] focused on top-K identification under the standard BTL model, and derived an ℓ_∞ error bound on preference scores which is intimately related to top-K ranking accuracy. Negahban *et al.* [7] considered the same comparison model and derived an ℓ_2 error bound. A key distinction in our work is that we consider a different measurement model in which there are two population sets, although the ℓ_∞ and ℓ_2 norm error analyses in [7] and [10] play crucial roles in determining the sample complexity.

The statistical model introduced by Chen *et al.* [6] attempts to represent crowdsourced settings and forms the basis of our adversarial comparison model. We note that no theoretical analysis of the sample complexity is available in [6] or other related works on crowdsourced rankings [18]–[20]. For example, Kim *et al.* [20] employed variational EM-based algorithms to estimate the latent scores; *global* optimality guarantees for such algorithms are difficult to establish. Jain and Oh [15] developed a tensor decomposition method [16] for learning the parameters of a mixture model [21]–[23] that includes our model as a special case. We specialize their model and relevant results to our setting for determining the accuracy of the estimated η. This allows us to establish an upper bound on the sample complexity when η is unknown.

Recently, Shah and Wainwright [24] showed that a simple counting method [25] achieves order-wise optimal sample complexity for top-K ranking under a general comparison model which includes, as special cases, a variety of parametric ranking models including the one under consideration in this paper (the BTL model). However, the authors made assumptions on the statistics of the pairwise comparisons which are different from that in our model. Hence, their result is not directly applicable to our setting.

D. Paper Organization

This paper is organized as follows. In Section II, we detail our system model and define the performance criterion. In Section III, we state our main results and provide interpretations of and intuitions behind these results. In Sections IV (achievability) and V (minimax converse) we prove the result concerning the sample complexity of learning the top-K ranking when the proportion of adversaries η is known. In Section VI we provide a high-level description of the algorithm for the case in which η is unknown and we sketch the achievability proof. In Section VII, we perform experiments on synthetic data to corroborate our result for the known η scenario. We conclude our discussion and suggest avenues for future research in Section VIII. Proofs of more technical results and other auxiliary discussions are deferred to the appendices.

E. Notations

We provide a brief summary of the notations used throughout the paper. Let $[n]$ represent $\{1, 2, \cdots, n\}$. We denote by $\|\boldsymbol{w}\|$, $\|\boldsymbol{w}\|_1$, $\|\boldsymbol{w}\|_\infty$ the ℓ_2 norm, ℓ_1 norm, and ℓ_∞ norm of $\boldsymbol{w}$, respectively. Additionally, for any two sequences $f(n)$ and $g(n)$, $f(n) \gtrsim g(n)$ or $f(n) = \Omega(g(n))$ mean that there exists a (universal) constant c such that $f(n) \geq cg(n)$; $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ mean that there exists a constant c such that $f(n) \leq cg(n)$; and $f(n) \asymp g(n)$ or $f(n) = \Theta(g(n))$ mean that there exist constants c_1 and c_2 such that $c_1 g(n) \leq f(n) \leq c_2 g(n)$. The notation poly$(n)$ denotes a sequence in $O(n^c)$ for some $c > 0$.

II. Problem Setup

We now describe the model which we will analyze subsequently. We assume that the observations used to learn the rankings are in the form of a limited number of pairwise comparisons over n items. In an attempt to reflect the adversarial crowdsourced setting of our interest in which there are two population sets—the *faithful* and *adversarial* sets—we adopt a comparison model introduced by Chen *et al.* [6]. This is a generalization of the BTL model [11], [12]. We now delve into the details of the components of the model.

1) (*Preference scores*): As in the standard BTL model, this model postulates the existence of a ground-truth preference score vector $\boldsymbol{w} = (w_1, w_2, \ldots, w_n) \in \mathbb{R}_+^n$. Each w_i represents the underlying preference score of item i. Without loss of generality, we assume that the scores are in non-increasing order:

$$w_1 \geq w_2 \geq \ldots \geq w_n > 0. \tag{3}$$

It is assumed that the dynamic range of the score vector is fixed irrespective of n:

$$w_i \in [w_{\min}, w_{\max}], \qquad \forall\, i \in [n], \tag{4}$$

for some positive constants $w_{\min}$ and $w_{\max}$. In fact, the case in which the ratio $\frac{w_{\max}}{w_{\min}}$ grows with n can be readily translated into the above setting by first separating out those items with vanishing scores (e.g., via a simple voting method like Borda count [25], [26]).

2) (*Comparison graph*): Let $\mathcal{G} := ([n], \mathcal{E})$ be the comparison graph such that items i and j are compared by an annotator if the node pair (i, j) belongs to the edge set $\mathcal{E}$. We will assume throughout that the edge set $\mathcal{E}$ is drawn in accordance to the Erdős-Rényi (ER) model $\mathcal{G} \sim \mathcal{G}_{n,p}$. That is node pair (i, j) appears independently of any other node pair with an observation probability $p \in (0, 1)$.

3) (*Pairwise comparisons*): For each edge $(i, j) \in \mathcal{E}$, we observe L comparisons between i and j. Each outcome, indexed by $\ell \in [L]$ and denoted by $Y_{ij}^{(\ell)}$, is drawn from a mixture of Bernoulli distributions weighted by an unknown parameter $\eta \in (1/2, 1]$. The ℓ-th observation of edge (i, j) has distribution $\mathsf{Bern}(\frac{w_i}{w_i+w_j})$ with probability η and distribution $\mathsf{Bern}(\frac{w_j}{w_i+w_j})$ with probability $1-\eta$. Hence,

$$Y_{ij}^{(\ell)} \sim \mathsf{Bern}\left(\eta \frac{w_i}{w_i + w_j} + (1-\eta)\frac{w_j}{w_i + w_j}\right). \quad (5)$$

See Fig. 1. When $\eta = 1/2$, all the observations are fair coin tosses. In this case, no information can be gleaned about the rankings. Thus we exclude this degenerate setting from our study. The case of $\eta \in [0, 1/2)$ is equivalent to the "mirrored" case of $1-\eta \in (1/2, 1]$ where we flip 0's to 1's and 1's to 0's. So without loss of generality, we assume that $\eta \in (1/2, 1]$. We allow η to depend on n.

We remark that in Chen *et al.* [6], each annotator indexed by ℓ has its own quality parameter η_ℓ. In our model all the η_ℓ's are equal for the sake of tractability and to model the scenario in which there are two sub-populations—faithful and adversarial.

Conditioned on a realization of the random graph $\mathcal{G} = ([n], \mathcal{E})$, for each $(i, j) \in \mathcal{E}$, the $Y_{ij}^{(\ell)}$'s are independent and identically distributed across all ℓ's, each according to the distribution in (5). Let $\underline{Y}^{(\ell)}$ be the column vector with entries $Y_{ij}^{(\ell)}$where these entries are indexed according to the lexicographical order of $(i, j) \in \mathcal{E}$. Define $\underline{\boldsymbol{Y}} := \{\underline{Y}^{(\ell)} : \ell \in [L]\}$. We will also often employ the sufficient statistics

$$Y_{ij} := \frac{1}{L}\sum_{\ell=1}^{L} Y_{ij}^{(\ell)}, \quad \forall\, (i, j) \in \mathcal{E}. \quad (6)$$

Note that Y_{ij} is a function of L, the *per-edge number of samples* and L is measure of the quality of the measurements. We let $\boldsymbol{Y}_i := \{Y_{ij}\}_{j:(i,j)\in\mathcal{E}}$ and $\boldsymbol{Y} := \{Y_{ij}\}_{(i,j)\in\mathcal{E}}$ be various collections of the sufficient statistics.

4) (*Performance metric*): We are interested in recovering the top-K ranked items in the collection of n items from the data $\boldsymbol{Y}$. We denote the true set of top-K ranked items by $\mathcal{S}_K$ which, by our ordering assumption, is the set $[K]$. We would like to design a ranking scheme $\psi : \{0, 1\}^{|\mathcal{E}|\times L} \to \binom{[n]}{K}$ that maps from the available measurements to a set of K indices. Given a ranking scheme ψ, the performance metric we consider is the *probability of error*

$$\mathrm{P_e}(\psi) := \Pr\left[\psi(\boldsymbol{Y}) \neq \mathcal{S}_K\right]. \quad (7)$$

We consider the fundamental *admissible region* $\mathcal{R}_{\boldsymbol{w}}$ of (p, L) pairs in which top-K ranking is feasible for a given $\boldsymbol{w}$, i.e., $\mathrm{P_e}(\psi)$ can be arbitrarily small for large enough n. In particular, we are interested in the *sample complexity*

$$S_\delta := \inf_{p\in[0,1], L\in\mathbb{Z}^+} \sup_{\boldsymbol{a}\in\Omega_\delta} \left\{\binom{n}{2} pL : (p, L) \in \mathcal{R}_{\boldsymbol{a}}\right\}, \quad (8)$$

where $\boldsymbol{\Omega}_\delta := \{\boldsymbol{a} \in \mathbb{R}^n : (a_K - a_{K+1})/a_{\max} \geq \delta\}$. Here we consider a *minimax* scenario in which, given a score estimator, nature can behave in an adversarial manner, and so she chooses the worst preference score vector that maximizes the probability of error under the constraint that the normalized score separation between the K-th and $(K+1)$-th items is at least δ. Note that $\binom{n}{2}p$ is the expected number of edges of the ER graph so $\binom{n}{2}pL$ is the expected number of pairwise samples drawn from the model of our interest.

III. Main Results

As suggested in [10], a crucial parameter for successful top-K ranking is the separation between the two items near the decision boundary,

$$\Delta_K := \frac{w_K - w_{K+1}}{w_{\max}}. \quad (9)$$

The sample complexity depends on $\boldsymbol{w}$ and K only through Δ_K—more precisely, it decreases as Δ_K increases. Our contribution is to identify relationships between η and the sample complexity when η is known and unknown. We will see that the sample complexity increases as Δ_K decreases. This is intuitively true as Δ_K captures how distinguishable the top-K set is from the rest of the items.

We assume that the graph $\mathcal{G}$ is drawn from the ER model $\mathcal{G}_{n,p}$ with edge appearance probability p. We require p to satisfy

$$p > \frac{\log n}{n}. \quad (10)$$

From random graph theory, this implies that the graph is connected with high probability. If the graph were not connected, rankings cannot be inferred [9].

We start by considering the η-known scenario in which key ingredients for ranking algorithms and analysis can be easily digested, as well as which forms the basis for the η-unknown setting.

Theorem 1 (Known η): Suppose that η is known and $\mathcal{G} \sim \mathcal{G}_{n,p}$. Also assume that $L = O(\mathrm{poly}(n))$ and $Lnp \geq \frac{c_0}{(2\eta-1)^2}\log n$. Then with probability $\geq 1 - c_1 n^{-c_2}$, the set of top-K set can be identified exactly provided that

$$L \geq c_3 \frac{\log n}{(2\eta-1)^2 np\,\Delta_K^2}. \quad (11)$$

Conversely, for a fixed $\epsilon \in (0, \frac{1}{2})$, if

$$L \leq c_4 \frac{(1-\epsilon)\log n}{(2\eta-1)^2 np\,\Delta_K^2} \quad (12)$$

holds, then for any top-K ranking scheme ψ, there exists a preference vector $\boldsymbol{w}$ with separation Δ_K such that $P_e(\psi) \geq \epsilon$. Here, and in the following, $c_i > 0, i \in \{0, 1, \ldots, 4\}$ are finite universal constants.

Proof: See Section IV for the algorithm and a sketch of the achievability proof (sufficiency). The proof of the converse (impossibility part) can be found in Section V. ■

This theorem asserts that the sample complexity scales as

$$S_{\Delta_K} \asymp \frac{n \log n}{(2\eta - 1)^2 \, \Delta_K^2}. \tag{13}$$

This result recovers that for the faithful scenario where $\eta = 1$ in [10]. When $\eta - \frac{1}{2}$ is uniformly bounded above 0, we achieve the same order-wise sample complexity. This suggests that the ranking performance is not substantially worsened if the sizes of the two populations are sufficiently distinct. For the challenging scenario in which $\eta \approx \frac{1}{2}$, the sample complexity depends on how $\eta - \frac{1}{2}$ scales with n. Indeed, this dependence is quadratic. Also notice that since the sample complexity in (13) is stated in terms of S_{Δ_K}, for a fixed number of pairwise samples $\binom{n}{2} pL$, it does not matter how to choose p and L as long as p satisfies (10) and the product $\binom{n}{2} pL$ satisfies (13). Theorem 1 will be validated by experimental results in Section VII. Several other remarks are in order.

1) *(No computational barrier)*: Our proposed algorithm is based primarily upon two popular ranking algorithms: spectral methods [7] and MLE [9], both of which enjoy nearly-linear time complexity in our ranking problem context. Hence, the information-theoretic limit promised by (13) can be achieved by a computationally efficient algorithm.
2) *(Implication of the minimax lower bound)*: The minimax lower bound continues to hold when η is unknown, since we can only do better for the η-known scenario, and hence the lower bound is also a lower bound in the η-unknown scenario.
3) *(Another adversarial scenario)*: Our results readily generalize to another adversarial scenario in which samples drawn from the adversarial population (of proportion η) are *completely noisy*, i.e., they follow the distribution $\mathsf{Bern}(\frac{1}{2})$. With a slight modification of our proof techniques, one can easily verify that the sample complexity is on the order of

$$\widetilde{S}_{\Delta_K} \asymp \frac{n \log n}{\eta^2 \Delta_K^2} \tag{14}$$

if η is known. Hence the closer the faithful proportion η is to 0, the worse the sample complexity, which is intuitively true. The result in (14) will be evident after we describe the algorithm in Section IV. For a sketch of the argument to obtain (14), please refer to Section IV-C.

Theorem 2 (Unknown η): Suppose that η is unknown and $\mathcal{G} \sim \mathcal{G}_{n,p}$. Also assume that $L = O(\mathrm{poly}(n))$ and $Lnp \geq \frac{c_0}{(2\eta-1)^4} \log^2 n$. Then with probability $\geq 1 - c_1 n^{-c_2}$, the top-K set can be identified exactly provided that

$$L \geq c_3 \frac{\log^2 n}{(2\eta - 1)^4 np \Delta_K^4}. \tag{15}$$

Proof: See Section VI for the key ideas in the proof. ■

This theorem implies that the sample complexity satisfies

$$S_{\Delta_K} \lesssim \frac{n \log^2 n}{(2\eta - 1)^4 \, \Delta_K^4}. \tag{16}$$

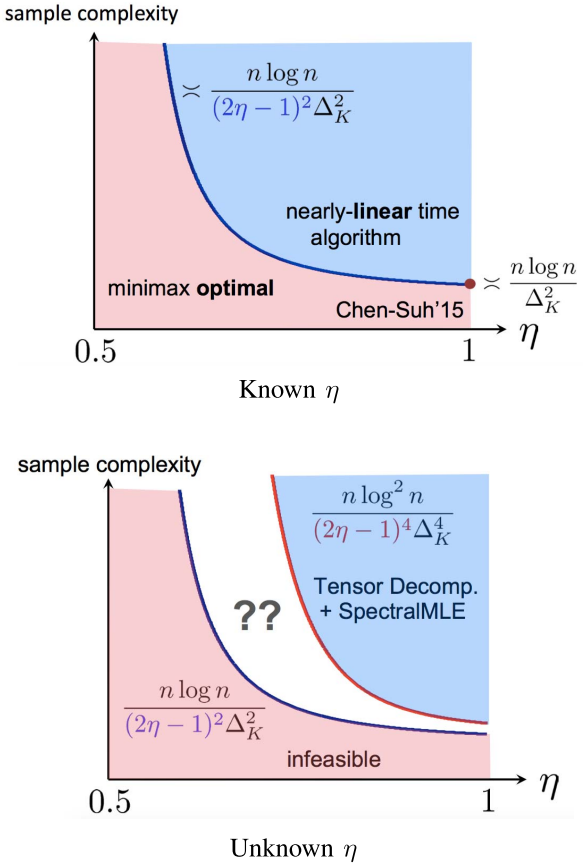

Fig. 2. Illustrations of our main theorems (Theorems 1 and 2)

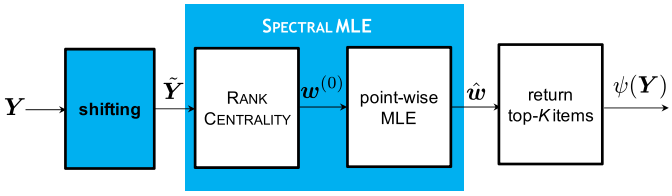

Fig. 3. Ranking algorithm for the η-known scenario: (1) shifting the empirical mean of pairwise measurements to get $\tilde{Y}_{ij} = \frac{Y_{ij} - (1-\eta)}{2\eta - 1}$, which converges to $\frac{w_i}{w_i + w_j}$ as $L \to \infty$; (2) performing SpectralMLE [10] seeded by $\tilde{\boldsymbol{Y}}$ to obtain a score estimate $\hat{\boldsymbol{w}}$; (3) return a ranking based on the estimate $\hat{\boldsymbol{w}}$. Our analysis reveals that the ℓ_∞ norm bound w.r.t. $\hat{\boldsymbol{w}}$ satisfies $\|\hat{\boldsymbol{w}} - \boldsymbol{w}\|_\infty \lesssim \frac{1}{2\eta-1}\sqrt{\frac{\log n}{npL}}$, which in turn ensures $P_e \to 0$ under $\Delta_K \gtrsim \frac{1}{2\eta-1}\sqrt{\frac{\log n}{npL}}$.

This bound is worse than (13)—the inverse dependence on $(2\eta - 1)^2 \Delta_K^2$ is now an inverse dependence on $(2\eta - 1)^4 \Delta_K^4$. This is because our algorithm involves estimating η, incurring some loss. Whether this loss is fundamentally unavoidable (i.e., whether the algorithm is order-wise optimal or not) is open. See detailed discussions in Section VIII. Moreover, since the estimation of η is based on tensor decompositions with polynomial-time complexity, our algorithm for the η-unknown case is also, in principle, computationally efficient. Note that minimax lower bound in (13) also serves as a lower bound in the η-unknown scenario.

The conclusions of Theorems 1 and 2 are illustrated graphically in Fig. 2.

IV. Algorithm and Achievability Proof of Theorem 1

A. Algorithm Description

Inspired by the consistency between the preference scores $\boldsymbol{w}$ and ranking under the BTL model, our scheme also adopts a two-step approach where $\boldsymbol{w}$ is first estimated and then the top-K set is returned.

Recently a top-K ranking algorithm *SpectralMLE* [10] has been developed for the faithful scenario and it is shown to have order-wise optimal sample complexity. The algorithm yields a small ℓ_∞ loss of the score vector $\boldsymbol{w}$ which ensures a small point-wise estimate error. Establishing a key relationship between the ℓ_∞ norm error and top-K ranking accuracy, Chen and Suh [10] then identify an order-wise tight bound on the ℓ_∞ norm error required for top-K ranking, thereby characterizing the sample complexity. Our ranking algorithm builds on SpectralMLE, which proceeds in two stages: (1) an appropriate initialization that concentrates around the ground truth in an ℓ_2 sense, which can be obtained via spectral methods [3], [7], [8]; (2) a sequence of T iterative updates sharpening the estimates in a point-wise manner using MLE.

We observe that *RankCentrality* [7] can be employed as a spectral method in the first stage. In fact, RankCentrality exploits the fact that the empirical mean Y_{ij} converges to the relative score $\frac{w_i}{w_i+w_j}$ as $L \to \infty$. This motivates the use of the empirical mean for constructing the transition probability from j to i of a Markov chain. Note that the detailed balance equation $\pi_i \frac{w_j}{w_i+w_j} = \pi_j \frac{w_i}{w_i+w_j}$ that holds as $L \to \infty$ will enforce that the stationary distribution of the Markov chain is identical to $\boldsymbol{w}$ up to some constant scaling. Hence, the stationary distribution is expected to serve as a reasonably good global score estimate. However, in our problem setting where η is not necessarily 1, the empirical mean does not converge to the relative score, instead it behaves as

$$Y_{ij} \overset{L\to\infty}{\longrightarrow} \eta\frac{w_i}{w_i+w_j} + (1-\eta)\frac{w_j}{w_i+w_j}. \tag{17}$$

(The above convergence is "in probability".) Note, however, that the limit is linear in the desired relative score and η, implying that knowledge of η with an appropriate linear transformation then leads to the relative score. A natural idea then arises. We construct a shifted version of the empirical mean:

$$\tilde{Y}_{ij} := \frac{Y_{ij}-(1-\eta)}{2\eta-1} \overset{L\to\infty}{\longrightarrow} \frac{w_i}{w_i+w_j}, \tag{18}$$

and take this as an input to *RankCentrality*. This then forms a Markov chain that yields a stationary distribution that is proportional to $\boldsymbol{w}$ as $L \to \infty$ and hence a good estimate of the ground-truth score vector when L is large. This serves as a good initial estimate to the second stage of *SpectralMLE* as it guarantees a small point-wise error.

A formal and more detailed description of the procedure is summarized in Algorithm 1. Also see the block diagram of the procedure in Fig. 3. For completeness, we also include the procedure of *RankCentrality* in Algorithm 2. Here we emphasize two distinctions w.r.t. the second stage of *SpectralMLE*.

Algorithm 1 Adversarial top-K ranking for the η-known scenario

Input: The average comparison outcome Y_{ij} for all $(i,j) \in \mathcal{E}$; the score range $[w_{\min}, w_{\max}]$.

Partition $\mathcal{E}$ randomly into two sets $\mathcal{E}^{\text{init}}$ and $\mathcal{E}^{\text{iter}}$ each containing $\frac{1}{2}|\mathcal{E}|$ edges. Denote by $\boldsymbol{Y}_i^{\text{init}}$ (resp. $\boldsymbol{Y}_i^{\text{iter}}$) the components of $\boldsymbol{Y}_i$ obtained over $\mathcal{E}^{\text{init}}$ (resp. $\mathcal{E}^{\text{iter}}$).

Compute the shifted version of the average comparison output: $\tilde{Y}_{ij} = \frac{Y_{ij}-(1-\eta)}{2\eta-1}$. Denote by $\tilde{\boldsymbol{Y}}_i^{\text{init}}$ the components of $\tilde{\boldsymbol{Y}}_i$ obtained over $\mathcal{E}^{\text{init}}$

Initialize $\boldsymbol{w}^{(0)}$ to be the estimate computed by *Rank Centrality* on $\tilde{\boldsymbol{Y}}_i^{\text{init}}$ $(1 \le i \le n)$.

Successive Refinement: for $t = 0 : T$ **do**

1) Compute the coordinate-wise MLE
$$w_i^{\mathsf{mle}} \leftarrow \arg\max_\tau \mathcal{L}\left(\tau, \boldsymbol{w}_{\setminus i}^{(t)}; \boldsymbol{Y}_i^{\mathsf{iter}}\right)$$
where $\mathcal{L}$ is the likelihood function defined in (19).

2) For each $1 \le i \le n$, set
$$w_i^{(t+1)} \leftarrow \begin{cases} w_i^{\mathsf{mle}}, & |w_i^{\mathsf{mle}} - w_i^{(t)}| > \xi_t; \\ w_i^{(t)}, & \text{else}, \end{cases}$$
where ξ_t is the replacement threshold defined in (20).

Output the indices of the K largest components of $\boldsymbol{w}^{(T)}$.

Algorithm 2 Rank Centrality [7]

Input: The shifted average comparison outcome $\tilde{Y}_{ij}$ for all $(i,j) \in \mathcal{E}^{\text{iter}}$.

Compute the transition matrix $\hat{P} = [\hat{p}_{ij}]_{1\le i,j\le n}$ such that for $(i,j) \in \mathcal{E}^{\text{iter}}$
$$\hat{p}_{ij} = \begin{cases} \frac{\tilde{Y}_{ji}}{d_{\max}}, & \text{if } i \ne j; \\ 1 - \frac{1}{d_{\max}} \sum_{k:(i,k)\in\mathcal{E}^{\text{iter}}} \tilde{Y}_{ki}, & \text{if } i = j. \end{cases}$$
where $d_{\max}$ is the maximum out-degrees of vertices in $\mathcal{E}^{\text{iter}}$.

Output the stationary distribution of $\hat{P}$.

First, the computation of the pointwise MLE w.r.t. say, item i, requires knowledge of η:

$$\begin{aligned}\mathcal{L}\big(\tau, \boldsymbol{w}_{\setminus i}^{(t)}; \boldsymbol{Y}_i\big) &= \prod_{j:(i,j)\in\mathcal{E}} \Bigg[\left(\eta\frac{\tau}{\tau+w_j^{(t)}} + (1-\eta)\frac{w_j^{(t)}}{\tau+w_j^{(t)}}\right)^{Y_{ij}} \\ &\quad \times \left(\eta\frac{w_j^{(t)}}{\tau+w_j^{(t)}} + (1-\eta)\frac{\tau}{\tau+w_j^{(t)}}\right)^{1-Y_{ij}}\Bigg]. \end{aligned} \tag{19}$$

Here, $\mathcal{L}(\tau, \boldsymbol{w}_{\setminus i}^{(t)}; \boldsymbol{Y}_i)$ is the profile likelihood of the preference score vector $[w_1^{(t)}, \cdots, w_{i-1}^{(t)}, \tau, w_{i+1}^{(t)}, \cdots, w_n^{(t)}]$ where $\boldsymbol{w}^{(t)}$ indicates the preference score estimate in the t-th iteration, $\boldsymbol{w}_{\setminus i}^{(t)}$ denotes the score estimate excluding the i-th component, and $\boldsymbol{Y}_i$ is the data available at node i. The second difference is the use of a different threshold ξ_t which incorporates the

effect of η:

$$\xi_t := \frac{c}{2\eta - 1}\left\{\sqrt{\frac{\log n}{npL}} + \frac{1}{2^t}\left(\sqrt{\frac{\log n}{pL}} - \sqrt{\frac{\log n}{npL}}\right)\right\}, \quad (20)$$

where $c > 0$ is a constant. This threshold is used to decide whether $w_i^{(t+1)}$ should be set to be the pointwise MLE w_i^{mle} in (25) (if $|w_i^{\mathsf{mle}} - w_i^{(t)}| > \xi_t$) or remains as $w_i^{(t)}$ (otherwise). The design of ξ_t is based on (1) the ℓ_∞ loss incurred in the first stage; and (2) a desirable ℓ_∞ loss that we intend to achieve at the end of the second stage. Since these two values are different, ξ_t needs to be adapted accordingly. Notice that the computation of ξ_t requires knowledge of η. The two modifications in (19) and (20) result in a more complicated analysis vis-à-vis Chen and Suh [10].

B. Achievability Proof of Theorem 1

Let $\hat{\boldsymbol{w}}$ be the final estimate $\boldsymbol{w}^{(T)}$ in the second stage. We carefully analyze the ℓ_∞ loss of the $\boldsymbol{w}$ vector, showing that under the conditions in Theorem 1

$$\|\hat{\boldsymbol{w}} - \boldsymbol{w}\|_\infty \le \frac{c_1}{2\eta - 1}\sqrt{\frac{\log n}{npL}}. \quad (21)$$

holds with probability exceeding $1 - c_2 n^{-c_3}$. This bound together with the following observation completes the proof. Observe that if $w_K - w_{K+1} \ge \frac{c_4}{2\eta-1}\sqrt{\frac{\log n}{npL}}$, then for a top-$K$ item $1 \le i \le K$ and a non-top-K item $j \ge K+1$,

$$\hat{w}_i - \hat{w}_j \ge w_i - w_j - |w_i - \hat{w}_i| - |w_j - \hat{w}_j| \quad (22)$$

$$\ge w_K - w_{K+1} - 2\|\hat{\boldsymbol{w}} - \boldsymbol{w}\|_\infty > 0. \quad (23)$$

This implies that our ranking algorithm outputs the top-K ranked items as desired. Hence, as long as $w_K - w_{K+1} \gtrsim \frac{1}{2\eta-1}\sqrt{\frac{\log n}{npL}}$ holds (coinciding with the claimed bound in Theorem 1), we can guarantee perfect top-K ranking, which completes the proof of Theorem 1.

The remaining part is the proof of (21). The proof builds upon the analysis made in [10], which demonstrates the relationship between $\frac{\|\boldsymbol{w}^{(0)} - \boldsymbol{w}\|}{\|\boldsymbol{w}\|}$ and $\|\boldsymbol{w}^{(T)} - \boldsymbol{w}\|_\infty$. We establish a new relationship for the arbitrary η case, formally stated in the following lemma. We will then use this to prove (21).

Lemma 1: Fix $\delta, \xi > 0$. Consider $\hat{\boldsymbol{w}}^{\mathrm{ub}}$ such that it is independent of $\mathcal{G}$ and satisfies

$$\frac{\|\hat{\boldsymbol{w}}^{\mathrm{ub}} - \boldsymbol{w}\|}{\|\boldsymbol{w}\|} \le \delta \quad \textit{and} \quad \|\hat{\boldsymbol{w}}^{\mathrm{ub}} - \boldsymbol{w}\|_\infty \le \xi. \quad (24)$$

Consider an estimate of the score vector $\hat{\boldsymbol{w}}$ such that $|\hat{w}_i - w_i| \le |\hat{w}_i^{\mathrm{ub}} - w_i|$ for all $i \in [n]$. Let

$$w_i^{\mathsf{mle}} := \arg\max_\tau \mathcal{L}(\tau, \hat{\boldsymbol{w}}_{\setminus i}; \boldsymbol{Y}_i). \quad (25)$$

Then, the pointwise error

$$|w_i^{\mathsf{mle}} - w_i| \le c_0 \max\left\{\delta + \frac{\log n}{np}\cdot\xi, \frac{c_1}{2\eta - 1}\sqrt{\frac{\log n}{npL}}\right\} \quad (26)$$

holds with probability at least $1 - c_2 n^{-c_3}$.

Proof: The relationship in the faithful scenario $\eta = 1$, which was proved in [10], means that the point-wise MLE w_i^{mle} is close to the ground truth w_i in a component-wise manner, once an initial estimate $\hat{\boldsymbol{w}}$ is accurate enough. Unlike the faithful scenario, in our setting, we have (in general) noisier measurements $\boldsymbol{Y}_i$ due to the effect of η. Nonetheless this lemma reveals that the relationship for the case of $\eta = 1$ is almost the same as that for an arbitrary η case only with a slight modification. This implies that a small point-wise loss is still guaranteed as long as we start from a reasonably good estimate. Here the only difference in the relationship is that the multiplication term of $\frac{1}{2\eta-1}$ additionally applies in the upper bound of (26). See Appendix A for the proof. ∎

Obviously the accuracy of the point-wise MLE reflected in the ℓ_∞ error depends crucially on an initial error $\|\boldsymbol{w}^{(0)} - \boldsymbol{w}\|$. In fact, Lemma 1 leads to the claimed bound (21) once the initial estimation error is properly chosen as follows:

$$\frac{\|\boldsymbol{w}^{(0)} - \boldsymbol{w}\|}{\|\boldsymbol{w}\|} \lesssim \frac{1}{2\eta - 1}\sqrt{\frac{\log n}{npL}}. \quad (27)$$

Here we demonstrate that the desired initial estimation error can indeed be achieved in our problem setting, formally stated in Lemma 2 (see below). On the other hand, adapting the analysis in [10], one can verify that with the replacement threshold ξ_t defined in (20), the ℓ_2 loss is monotonically decreasing in an order-wise sense, i.e.,

$$\frac{\|\boldsymbol{w}^{(t)} - \boldsymbol{w}\|}{\|\boldsymbol{w}\|} \lesssim \frac{\|\boldsymbol{w}^{(0)} - \boldsymbol{w}\|}{\|\boldsymbol{w}\|}. \quad (28)$$

We are now ready to prove (21) when $L = O(\mathrm{poly}(n))$ and

$$\frac{\|\boldsymbol{w}^{(t)} - \boldsymbol{w}\|}{\|\boldsymbol{w}\|} \asymp \delta \asymp \frac{1}{2\eta - 1}\sqrt{\frac{\log n}{npL}}. \quad (29)$$

Lemma 1 asserts that in this regime, the point-wise MLE $\boldsymbol{w}^{\mathsf{mle}}$ is expected to satisfy

$$\|\boldsymbol{w}^{\mathsf{mle}} - \boldsymbol{w}\|_\infty \lesssim \frac{\|\boldsymbol{w}^{(t)} - \boldsymbol{w}\|}{\|\boldsymbol{w}\|} + \frac{\log n}{np}\|\boldsymbol{w}^{(t)} - \boldsymbol{w}\|_\infty. \quad (30)$$

Using the analysis in [10], one can show that the choice of ξ_t in (20) enables us to detect outliers (where an estimation error is large) and drag down the corresponding point-wise error, thereby ensuring that $\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}\|_\infty \asymp \|\boldsymbol{w}^{\mathsf{mle}} - \boldsymbol{w}\|_\infty$. This together with the fact that

$$\frac{\|\boldsymbol{w}^{(t)} - \boldsymbol{w}\|}{\|\boldsymbol{w}\|} \lesssim \frac{\|\boldsymbol{w}^{(0)} - \boldsymbol{w}\|}{\|\boldsymbol{w}\|} \lesssim \frac{1}{2\eta - 1}\sqrt{\frac{\log n}{npL}} \quad (31)$$

(see (29) above and Lemma 2) gives

$$\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}\|_\infty \lesssim \frac{1}{2\eta - 1}\sqrt{\frac{\log n}{npL}} + \frac{\log n}{np}\|\boldsymbol{w}^{(t)} - \boldsymbol{w}\|_\infty. \quad (32)$$

A straightforward computation with this recursion yields (21) if $\frac{\log n}{np}$ is sufficiently small (e.g., $p > \frac{2\log n}{n}$) and T, the number of iterations in the second stage of *SpectralMLE*, is sufficiently large (e.g., $T = O(\log n)$).

Lemma 2: Let $L = O(\text{poly}(n))$ *and* $Lnp \geq \frac{c_0}{(2\eta-1)^2}\log n$. *Let* $\boldsymbol{w}^{(0)}$ *be an initial estimate: an output of RankCentrality [7] when seeded by* $\tilde{\boldsymbol{Y}} := \{\tilde{Y}_{ij}\}_{(i,j)\in\mathcal{E}}$. *Then,*

$$\frac{\|\boldsymbol{w} - \boldsymbol{w}^{(0)}\|}{\|\boldsymbol{w}\|} \leq \frac{c_1}{2\eta - 1}\sqrt{\frac{\log n}{npL}} \tag{33}$$

holds with probability exceeding $1 - c_2 n^{-c_3}$.

Proof: Here we provide only a sketch of the proof, leaving details to Appendix B. The proof builds upon the analysis structured by [7, Lemma 2], which bounds the deviation of the Markov chain w.r.t. the transition matrix $\hat{P}$ after t steps:

$$\frac{\|\hat{p}_t - \boldsymbol{w}\|}{\|\boldsymbol{w}\|} \leq \rho^t \frac{\|\hat{p}_0 - \boldsymbol{w}\|}{\|\boldsymbol{w}\|}\sqrt{\frac{w_{\max}}{w_{\min}}} + \frac{1}{1-\rho}\|\Delta\|\sqrt{\frac{w_{\max}}{w_{\min}}} \tag{34}$$

where $\hat{p}_t$ denotes the distribution w.r.t. $\hat{P}$ at time t seeded by an arbitrary initial distribution $\hat{p}_0$, the matrix $\Delta := \hat{P} - P$, indicates the fluctuation of the transition probability matrix[1] around its mean $P := \mathbb{E}[\hat{P}]$, and $\rho := \lambda_{\max} + \|\Delta\|\sqrt{\frac{w_{\max}}{w_{\min}}}$. Here $\lambda_{\max} = \max\{\lambda_2, -\lambda_n\}$ and λ_i indicates the i-th eigenvalue of P.

Unlike the faithful scenario $\eta = 1$, in the arbitrary η case, the bound on $\|\Delta\|$ depends on η:

$$\|\Delta\| \lesssim \frac{1}{2\eta - 1}\sqrt{\frac{\log n}{npL}}, \tag{35}$$

which will be proved in Lemma B by using various concentration bounds (e.g., Hoeffding and Tropp [27]). Adapting the analysis in [7], one can easily verify that $\rho < 1$ under one of the conditions in Theorem 1 that $Lnp \gtrsim \frac{\log n}{(2\eta-1)^2}$. Applying the bound on $\|\Delta\|$ and $\rho < 1$ to (34) gives the claimed bound, which completes the proof. ∎

C. The Completely Noisy Case

As a final remark, in the third remark following Theorem 1, we mentioned that if the adversarial population generates *completely noisy* $\mathsf{Bern}(\frac{1}{2})$ observations, then the sample complexity is as in (14). This is easily seen as follows: Analogously to (17), the samples Y_{ij} are now generated from a Bernoulli distribution with bias $\eta\frac{w_i}{w_i+w_j} + (1-\eta)\frac{1}{2}$. Thus, we ought to shift and scale the observed samples to form $\tilde{Y}_{ij} := \frac{1}{\eta}(Y_{ij} - \frac{1}{2}) + \frac{1}{2}$ and this sequence of random variables will converge to the desired proportion $\frac{w_i}{w_i+w_j}$ as $L \to \infty$. Due to the factor η in the denominator (instead of $2\eta - 1$ as in (18)), we see that the sample complexity is as in (14) (with η^2 in the denominator).

V. Converse Proof of Theorem 1

As in Chen and Suh's work [10], by Fano's inequality, we see that it suffices for us to upper bound the mutual information between a set of appropriately chosen rankings $\mathcal{M}$ of cardinality $M := \min\{K, n-K\}+1$. More specifically, let $\sigma : [n] \to [n]$ represent a permutation over $[n]$. We also denote by $\sigma(i)$ and $\sigma([K])$ the corresponding index of the i-th ranked item and the index set of all top-K items, respectively. We subsequently impose a uniform prior over $\mathcal{M}$ as follows: If $K < n/2$ then

$$\Pr[\sigma([K]) = \mathcal{S}] = \frac{1}{M} \quad \text{for } \mathcal{S} = \{2, \ldots, K\} \cup \{i\}, \quad i = 1, K+1, \ldots, n \tag{36}$$

and if $K \geq n/2$, then

$$\Pr[\sigma([K]) = \mathcal{S}] = \frac{1}{M} \quad \text{for } \mathcal{S} = \{1, \ldots, K+1\} \setminus \{i\}, \quad i = 1, \ldots, K+1. \tag{37}$$

In words, each alternative hypothesis is generated by swapping *only two* indices of the hypothesis (ranking) obeying $\sigma([K]) = [K]$. Clearly, the original minimax error probability is lower bounded by the corresponding error probability of this reduced ensemble.

Let the set of observations for the edge $(i, j) \in \mathcal{E}$ be denoted as $\vec{Y}_{ij} := \{Y_{ij}^{(\ell)} : \ell \in [L]\}$. We also find it convenient to introduce an "erased" version of the observations $\boldsymbol{Z} = \{\vec{Z}_{ij} : (i, j) \in [n]^2\}$ which is related to the true observations $\boldsymbol{Y} := \{\vec{Y}_{ij} : (i, j) \in \mathcal{E}\}$ as follows,

$$\vec{Z}_{ij} = \begin{cases} \vec{Y}_{ij} & (i,j) \in \mathcal{E} \\ \mathrm{e} & (i,j) \notin \mathcal{E} \end{cases}. \tag{38}$$

Here e is an *erasure* symbol. Let σ, a chance variable, be a uniformly distributed ranking in $\mathcal{M}$ (the ensemble of rankings created in (36)–(37)). Let $P_{\vec{Y}_{ij}|\sigma_j}$ be the distribution of the observations given that the ranking is $\sigma_j \in \mathcal{M}$ where $j \in [M]$ and a similar notation is used for when $\vec{Y}_{ij}$ is replaced by $\vec{Z}_{ij}$. Now, by the convexity of the relative entropy and the fact that the rankings are uniform, the mutual information can be bounded as

$$I(\sigma; \boldsymbol{Z}) \leq \frac{1}{M^2}\sum_{\sigma_1,\sigma_2\in\mathcal{M}} D\left(P_{\boldsymbol{Z}|\sigma_1} \middle\| P_{\boldsymbol{Z}|\sigma_2}\right) \tag{39}$$

$$= \frac{1}{M^2}\sum_{\sigma_1,\sigma_2\in\mathcal{M}}\sum_{i\neq j} D\left(P_{\vec{Z}_{ij}|\sigma_1} \middle\| P_{\vec{Z}_{ij}|\sigma_2}\right) \tag{40}$$

$$= \frac{p}{M^2}\sum_{\sigma_1,\sigma_2\in\mathcal{M}}\sum_{i\neq j} D\left(P_{\vec{Y}_{ij}|\sigma_1} \middle\| P_{\vec{Y}_{ij}|\sigma_2}\right) \tag{41}$$

$$= \frac{p}{M^2}\sum_{\sigma_1,\sigma_2\in\mathcal{M}}\sum_{i\neq j}\sum_{\ell=1}^{L} D\left(P_{Y_{ij}^{(\ell)}|\sigma_1} \middle\| P_{Y_{ij}^{(\ell)}|\sigma_2}\right). \tag{42}$$

Assume that under ranking σ_1, the score vector is $\boldsymbol{w} := (w_1, \ldots, w_n)$ and under ranking σ_2, the score vector is $\boldsymbol{w}' := (w_{\pi(1)}, \ldots, w_{\pi(n)})$ for some fixed permutation $\pi : [n] \to [n]$. By using the statistical model in Section II, we know that

$$D\left(P_{Y_{ij}^{(\ell)}|\sigma_1} \middle\| P_{Y_{ij}^{(\ell)}|\sigma_2}\right) = D\left(\eta\frac{w_i}{w_i+w_j} + (1-\eta)\frac{w_j}{w_i+w_j} \middle\| \eta\frac{w_{\pi(i)}}{w_{\pi(i)}+w_{\pi(j)}} + (1-\eta)\frac{w_{\pi(j)}}{w_{\pi(i)}+w_{\pi(j)}}\right) \tag{43}$$

[1]The notation $\Delta = \hat{P} - P$, a matrix, should not be confused with the scalar normalized score separation Δ_K, defined in (9).

where $D(\alpha\|\beta) := \alpha\log\frac{\alpha}{\beta} + (1-\alpha)\log\frac{1-\alpha}{1-\beta}$ is the binary relative entropy. For brevity, write

$$a := \frac{w_i}{w_i + w_j}, \quad \text{and} \quad b := \frac{w_{\pi(i)}}{w_{\pi(i)} + w_{\pi(j)}}. \tag{44}$$

Furthermore, we note that the chi-squared divergence is an upper bound for the relative entropy between two distributions $P = \{P_i\}_{i\in\mathcal{X}}$ and $Q = \{Q_i\}_{i\in\mathcal{X}}$ on the same (countable) alphabet $\mathcal{X}$ (see e.g. [28, Lemma 6.3]), i.e.,

$$D(P\|Q) \le \chi^2(P\|Q) := \sum_{i\in\mathcal{X}} \frac{(P_i - Q_i)^2}{Q_i}. \tag{45}$$

We also use the notation $\chi^2(\alpha\|\beta)$ to denote the binary chi-squared divergence similarly to the binary relative entropy. Now, we may bound (43) using the following computation

$$\begin{aligned} &D\big(\eta a + (1-\eta)(1-a)\big\|\eta b + (1-\eta)(1-b)\big) \\ &\quad\le \chi^2\big(\eta a + (1-\eta)(1-a)\big\|\eta b + (1-\eta)(1-b)\big) \qquad (46)\\ &\quad= \frac{(2\eta-1)^2(a-b)^2}{\big((2\eta-1)b + (1-\eta)\big)\big(\eta - (2\eta-1)b\big)} \qquad (47)\end{aligned}$$

Now

$$|a-b| \le \frac{w_K}{w_K + w_{K+1}} - \frac{w_{K+1}}{w_K + w_{K+1}} \le \frac{w_{\max}}{2w_{\min}}\Delta_K. \tag{48}$$

Hence, if we consider the case where $\eta = (1/2)^+$ (which is the regime of interest), uniting (47) and (48) we obtain

$$\begin{aligned} &D\big(\eta a + (1-\eta)(1-a)\big\|\eta b + (1-\eta)(1-b)\big) \\ &\quad\lesssim (2\eta-1)^2\Delta_K^2. \end{aligned} \tag{49}$$

By construction of the hypotheses in (36)–(37), conditional on any two distinct rankings $\sigma_1, \sigma_2 \in \mathcal{M}$, the distributions of $\vec{Y}_{ij}$ (namely $P_{\vec{Y}_{ij}|\sigma_1}$ and $P_{\vec{Y}_{ij}|\sigma_2}$) are different over at most $2n$ locations so

$$\sum_{i\ne j}\sum_{l=1}^{L} D\left(P_{Y_{ij}^{(\ell)}|\sigma_1}\middle\|P_{Y_{ij}^{(\ell)}|\sigma_2}\right) \lesssim nL(2\eta-1)^2\Delta_K^2. \tag{50}$$

Thus, plugging this into the bound on the mutual information in (42), we obtain

$$I(\sigma;\boldsymbol{Z}) \lesssim pnL(2\eta-1)^2\Delta_K^2. \tag{51}$$

Plugging this into Fano's inequality, and using the fact that $M \le n/2$ (from $M = \min\{K, n-K\}+1$), we obtain

$$\begin{aligned} \mathrm{P_e}(\psi) &\ge 1 - \frac{I(\sigma;\boldsymbol{Z})}{\log M} - \frac{1}{\log M} \qquad (52)\\ &\ge 1 - \frac{I(\sigma;\boldsymbol{Z})}{\log(n/2)} - \frac{1}{\log(n/2)}. \qquad (53)\end{aligned}$$

Thus, if $S_{\Delta_K} = \binom{n}{2}pL \le \frac{c_2(1-\epsilon)\log n}{(2\eta-1)^2\Delta_K^2}$ for some small enough but positive c_2, we see that

$$\mathrm{P_e}(\psi) \ge \epsilon > 0. \tag{54}$$

Since this is independent of the decoder ψ, the converse part is proved.

As a final remark, let us mention that it is easy to use techniques in [29] to show that if L satisfies the upper bound in (13), not only does the probability of error fail to tend to zero as in (54), but it is arbitrarily close to 1 for sufficiently large n, a so-called *strong converse* statement.

VI. Algorithm and Proof of Theorem 2

A. High-Level Description of Algorithm

The proof of Theorem 2 follows by combining the results of Jain and Oh [15] with the analysis for the case when η is known in Theorem 1. Jain and Oh were interested in disambiguating a mixture distribution from samples. This corresponds to our model in (5). They showed using tensor decomposition methods that it is possible to find a globally optimal solution for the mixture weight η using a computationally efficient algorithm. They also provided an ℓ_2 bound on the error of the distributions but as mentioned, we are more interested in controlling the ℓ_∞ error so we estimate $\boldsymbol{w}$ separately. The use of the ℓ_2 bound in [15] leads to a worse sample complexity for top-K ranking.

Thus, in the first step, we will use the method in [15] to estimate η given the sufficient statistics of the data samples (pairwise comparisons) $\boldsymbol{Y}$. The estimate is denoted as $\hat{\eta}$. It turns out that one can specialize the result in [15] with suitably parametrized "distribution vectors"

$$\pi_0 := \left[\cdots \quad \frac{w_i}{w_i+w_j} \quad \frac{w_j}{w_i+w_j} \quad \frac{w_{i'}}{w_{i'}+w_{j'}} \quad \frac{w_{j'}}{w_{i'}+w_{j'}} \quad \cdots\right]^T \tag{55}$$

and $\pi_1 := \mathbf{1}_{2|\mathcal{E}|} - \pi_0 \in \mathbb{R}^{2|\mathcal{E}|}$ and where in (55), (i, j) runs through all edges in $\mathcal{E}$. Hence, we are in fact applying [15] to a more restrictive setting where the two probability distributions represented by π_0 and π_1 are "coupled" but this does not preclude the application of the results in [15]. In fact, this assumption makes the calculation of relevant parameters (in Lemma 6) easier. The relevant second and third moments are

$$M_2 := \eta\pi_0\otimes\pi_0 + (1-\eta)\pi_1\otimes\pi_1, \tag{56}$$

$$M_3 := \eta\pi_0\otimes\pi_0\otimes\pi_0 + (1-\eta)\pi_1\otimes\pi_1\otimes\pi_1, \tag{57}$$

where $\pi_j \otimes \pi_j \in \mathbb{R}^{(2|\mathcal{E}|)\times(2|\mathcal{E}|)}$ is the outer product and $\pi_j\otimes\pi_j\otimes\pi_j \in \mathbb{R}^{(2|\mathcal{E}|)\times(2|\mathcal{E}|)\times(2|\mathcal{E}|)}$ is the 3-fold tensor outer product. If one has the *exact* M_2 and M_3, we can obtain the mixture weight η *exactly*. The intuition as to why tensor methods are applicable to problems involving latent variables has been well-documented (e.g. [16]). Essentially, the second- and third-moments contained in M_2 and M_3 provide sufficient statistics for identifying and hence estimating *all* the parameters of an appropriately-defined model with latent variables (whereas second-order information contained in M_2 is, in general, not sufficient for reconstructing the parameters). Thus, the problem boils down to analyzing the precision of η when we only have access to *empirical* versions of M_2 and M_3 formed from pairwise comparisons in $\mathcal{G}$. As shown in Lemma 5 to follow, there is a tradeoff between the sample size per edge L and the quality of the estimate of η. Hence, this causes a degradation to the overall sample complexity reflected in Theorem 2.

In the second step, we plug the estimate $\hat{\eta}$ into the algorithm for the η-known case by shifting the observations $\boldsymbol{Y}$ similarly to (18) but with $\hat{\eta}$ instead of η. See Fig. 4. However, here there are a couple of important distinctions relative to the case where η is known exactly. First, the likelihood function $\mathcal{L}(\cdot)$ in (19) needs to be modified since it is a function of η in which now

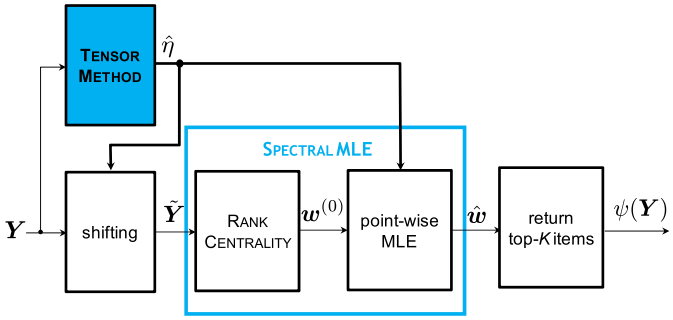

Fig. 4. Ranking algorithm for the unknown η scenario. The key distinction relative to the known η case is that we estimate η based on the tensor decomposition method [15], [16] and the estimate $\hat{\eta}$ is employed for shifting $\boldsymbol{Y}$ and performing the point-wise MLE. This method allows us to get $\|\hat{\boldsymbol{w}} - \boldsymbol{w}\|_\infty \lesssim \frac{1}{2\eta-1}\sqrt[4]{\frac{\log^2 n}{npL}}$, which ensures that $P_e \to 0$ under $\Delta_K \gtrsim \frac{1}{2\eta-1}\sqrt[4]{\frac{\log^2 n}{npL}}$.

we only have its estimate $\hat{\eta}$. Second, since the guarantee on the ℓ_∞ loss of the preference score vector $\boldsymbol{w}$ is different (and in fact worse), we need to design the threshold ξ_t differently from (20). We call the modified threshold $\hat{\xi}_t$, to be defined precisely in (66).

B. Detailed Algorithm for Estimating the Mixing Coefficient η

The algorithm for estimating η is shown in Algorithm 3, with a subroutine (alternating minimization for matrix completion) shown in Algorithm 4. Some notations that appear in the algorithms are defined as follows. First, we define two sets of indices $\Omega_2 := \{(i,j) \in [2|\mathcal{E}|] \times [2|\mathcal{E}|] : \lceil i/2 \rceil \neq \lceil j/2 \rceil\}$ and $\Omega_3 := \{(i,j,k) \in [2|\mathcal{E}|] \times [2|\mathcal{E}|] \times [2|\mathcal{E}|] : \lceil i/2 \rceil \neq \lceil j/2 \rceil \neq \lceil k/2 \rceil \neq \lceil i/2 \rceil\}$. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, let $\otimes^3 \boldsymbol{x} := \boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}$ be the 3-fold (tensor) outer product. Given a m-th order tensor $A \in \mathbb{R}^{d \times d \times \ldots \times d} \cong \mathbb{R}^{d^m}$ and an index set $\Omega \subseteq [d]^m$, we define the projection operator $\mathcal{P}_\Omega : \mathbb{R}^{d^m} \to \mathbb{R}^{d^m}$ as

$$[\mathcal{P}_\Omega(A)]_{i_1,\ldots,i_m} := \begin{cases} A_{i_1,\ldots,i_m}, & (i_1,\ldots,i_m) \in \Omega \\ 0, & \text{otherwise,} \end{cases} \tag{58}$$

for all $(i_1,\ldots,i_m) \in [d]^m$. If A is symmetric and $m = 2$, denote its (unique) eigen-decomposition as $A = U_A \Sigma_A U_A^T$, where the diagonal entries of Σ_A are arranged in the deceasing order. Then we define $P_A := (U_A \Sigma_A^{1/2})^T$ and $Q_A := U_A \Sigma_A^{-1/2}$. If A is symmetric and $m = 3$, we define an $r \times r \times r$ operation with respect to $R \in \mathbb{R}^{d \times r}$ as

$$(A[R]_3)_{j_1,j_2,j_3} := \sum_{i_1,i_2,i_3 \in [d]} A_{i_1,i_2,i_3} \prod_{k=1}^{3} R_{i_k,j_k}, \tag{59}$$

for all $j_1, j_2, j_3 \in [r]$. For any $A, B \in \mathbb{R}^{d^m}$, define their inner product $\langle A, B \rangle := \sum_{i_1,\ldots,i_m \in [d]} A_{i_1,\ldots,i_m} B_{i_1,\ldots,i_m}$ and $\|A\|_F := \sqrt{\langle A, A \rangle}$. Finally, define two set of indices $\mathcal{I}_1 := \{1,\ldots,\lfloor L/2 \rfloor\}$ and $\mathcal{I}_2 := \{\lfloor L/2 \rfloor + 1,\ldots,L\}$.

Note that in Algorithm 3, a crucial step is to find the third-order statistic $\widehat{G}$ by solving the least squares problem in (60). Our theoretical guarantee for exact identification of

Algorithm 3 Estimating mixing coefficient η [15]

Input: The collection of observed pairwise comparisons $\underline{\boldsymbol{Y}}$

Split $\boldsymbol{Y}$ evenly into two subsets of samples $\underline{\boldsymbol{Y}}^{(1)} := \{\underline{Y}^{(\ell)} : \ell \in \mathcal{I}_1\}$ and $\underline{\boldsymbol{Y}}^{(2)} := \{\underline{Y}^{(\ell)} : \ell \in \mathcal{I}_2\}$

Compute $\widehat{M}_2$, the estimated second-order moment matrix M_2 in (56) based on $\underline{\boldsymbol{Y}}^{(1)}$ using Algorithm 4

Compute $\widehat{G}$, the estimated third-order statistic of $G := M_3[Q_{M_2}]_3$ by solving the least squares problem[2]

$$\widehat{G} \in \operatorname*{argmin}_{Z \in \mathbb{R}^{2\times 2\times 2}} \left\| \mathcal{P}_{\Omega_3} \left(Z\left[P_{\widehat{M}_2}\right]_3 - \sum_{t \in \mathcal{I}_2} \frac{\otimes^3 \underline{Y}^{(t)}}{|\mathcal{I}_2|} \right) \left[Q_{\widehat{M}_2}\right]_3 \right\|_F^2. \tag{60}$$

Compute the first eigenvalue λ_1 of $\widehat{G}$ using the *robust power method* in Anandkumar *et al.* [16]

Return the estimated mixing coefficient $\hat{\eta} = \lambda_1^{-2}$

Algorithm 4 Alternating Minimization for Matrix Completion [15]

Input: The collection of observed pairwise comparisons $\underline{\boldsymbol{Y}}^{(1)}$, maximum number of iterations T

Compute a summary statistic $S_2 := \frac{1}{|\mathcal{I}_1|} \sum_{t \in \mathcal{I}_1} \underline{Y}^{(t)} \otimes \underline{Y}^{(t)}$.

Initialize $U_0 \in \mathbb{R}^{2|\mathcal{E}| \times 2}$ using the top two eigenvectors of $\mathcal{P}_{\Omega_2}(S_2)$.

For $\tau = 0,\ldots,T-1$

Compute the (projected) Cholesky decomposition

$$\widehat{U}_{\tau+1} \in \operatorname*{argmin}_{U \in \mathbb{R}^{2|\mathcal{E}| \times 2}} \left\| \mathcal{P}_{\Omega_2}\left(S_2 - U U_\tau^T\right) \right\|_F^2. \tag{61}$$

Compute the QR decomposition

$$[U_{\tau+1}, R_{\tau+1}] = \mathrm{QR}(\widehat{U}_{\tau+1}). \tag{62}$$

end for

Return the estimated second-order moment matrix $\widehat{M}_2 := \widehat{U}_T U_{T-1}^T$

the top-K set (Theorem 2) requires both the sample size L and the number of observed edges $|\mathcal{E}|$ to be sufficiently large. Therefore, (60) is a large-scale optimization problem. Recently, Huang *et al.* [30] proposed to use the stochastic tensor gradient descent (STGD) method to perform tensor decomposition in an online manner specifically for problems where L is large. In Appendix G, we discuss how to make use of this method to solve (60). We also discuss the difficulties of applying such a method when $|\mathcal{E}|$ (i.e., the number of "features") is also large.

[2] $\widehat{G}$ can be taken to be any minimizer of (60), and similar for $\widehat{U}_{\tau+1}$ in (61).

C. Proof of Theorem 2

As in Section IV-B, the crux is to analyze the ℓ_∞ loss of the $\boldsymbol{w}$ vector. We show that

$$\|\hat{\boldsymbol{w}} - \boldsymbol{w}\|_\infty \leq \frac{c_0}{2\eta - 1}\sqrt[4]{\frac{\log^2 n}{npL}} \tag{63}$$

holds with probability $\geq 1 - c_1 n^{-c_2}$. To guarantee that top-K ranking is accurate, we then follow the same argument as in (22)–(23). We lower bound $\|\hat{\boldsymbol{w}} - \boldsymbol{w}\|_\infty$ in (63) by Δ_K and solve for L. Thus, it suffices to show (63) under the conditions of Theorem 2.

The proof of (63) follows from several lemmas, two of which we present in this section. These are the analogues of Lemmas 1 and 2 for the η-known case. Once we have these two lemmas, the strategy to proving (63) is almost the same as that in the η-known setting in Section IV-B so we omit the details.

The first lemma concerns the relationship between the normalized ℓ_2 error and the ℓ_∞ error when we do not have access to the true mixture weight η, but only an estimate of it given via Algorithm 3.

Lemma 3: Consider $\hat{\boldsymbol{w}}^{\mathsf{ub}}$ such that it is independent of $\mathcal{G}$ and satisfies (24)*. Consider $\hat{\boldsymbol{w}}$ such that $|\hat{w}_i - w_i| \leq |\hat{w}_i^{\mathsf{ub}} - w_i|$ for all $i \in [n]$. Now define*

$$w_i^{\mathsf{mle}} := \arg\max_{\tau} \hat{\mathcal{L}}(\tau, \hat{\boldsymbol{w}}_{\setminus i}; \boldsymbol{Y}_i), \tag{64}$$

where $\hat{\mathcal{L}}(\cdot)$ is the surrogate likelihood (cf. (19)*) constructed with $\hat{\eta}$ in place of η. Then, for all i, the same pointwise MLE bound in* (26) *holds with probability $\geq 1 - c_0 n^{-c_1}$.*

Proof: The proof parallels that of Lemma 1 but is more technical. We analyze the fidelity of the estimate $\hat{\eta}$ relative to η as a function of L (Lemma 5). This requires the specialization of Jain and Oh [15] to our setting. By proving several continuity statements, we show that the estimated normalized log-likelihood (NLL) $\frac{1}{L}\log\hat{\mathcal{L}}(\cdot)$ is *uniformly* close to the true NLL $\frac{1}{L}\log\mathcal{L}(\cdot)$ w.h.p. This leads us to prove (26), which is the *same* as the η-known case. The details are deferred to Appendix C. ∎

Similarly to the case where η is known, we need to subsequently control the initial error $\|\boldsymbol{w}^{(0)} - \boldsymbol{w}\|$. For the η-known case, this is done in Lemma 2 so the following lemma is an analogue of Lemma 2.

Lemma 4: Assume the conditions of Theorem 2 hold. Let $\boldsymbol{w}^{(0)}$ be an initial estimate, i.e., an output of RankCentrality when seeded by $\check{\boldsymbol{Y}}$ which consists of the shifted observations with $\hat{\eta}$ in place of η (cf. (18)*). Then,*

$$\frac{\|\boldsymbol{w} - \boldsymbol{w}^{(0)}\|}{\|\boldsymbol{w}\|} \leq \frac{c_0}{2\eta - 1}\sqrt[4]{\frac{\log^2 n}{npL}} \tag{65}$$

holds with probability $\geq 1 - c_1 n^{-c_2}$.

Proof: See Section VI-D for a sketch of the proof and Appendix D for a detailed calculation of an upper bound on the spectral norm of the fluctuation matrix, which is a key ingredient of the proof of Lemma 4. ∎

We remark that (65) is worse than its η-known counterpart in (33). In particular, there is now a fourth root inverse dependence on L (compared to a square root inverse dependence), which means we potentially need many more observations to drive the normalized ℓ_2 error $\frac{\|\boldsymbol{w} - \boldsymbol{w}^{(0)}\|}{\|\boldsymbol{w}\|}$ down to the same level. This loss is present because there is a penalty incurred in estimating η via the tensor decomposition approach, especially when η is close to $1/2$. In the analysis, we need to control the Lipschitz constants of functions such as $t \mapsto \frac{1}{2t-1}$ and $t \mapsto \frac{1-t}{2t-1}$ (see e.g. (18)). Such functions behave badly near $1/2$. In particular, the gradient diverges as $t \downarrow 1/2$. We have endeavored to optimize (65) so that it is as tight as possible, at least using the proposed methods.

Using Lemmas 3 and 4 and invoking a similar argument as in the η-known scenario, we can now to prove (63). One key distinction here lies in the choice of the threshold:

$$\hat{\xi}_t := \frac{c}{2\hat{\eta} - 1}\left\{\sqrt[4]{\frac{\log^2 n}{npL}} + \frac{1}{2^t}\left(\sqrt[4]{\frac{n\log^2 n}{pL}} - \sqrt[4]{\frac{\log^2 n}{npL}}\right)\right\}. \tag{66}$$

The rationale behind this choice, which is different from (20), is that it drives the initial ℓ_∞ loss (associated to the initial ℓ_2 loss in Lemma 4) to approach the desired ℓ_∞ loss in (63). Taking this choice, which we optimized, and adapting the analysis in [10] with Lemma 3, one can verify that the ℓ_∞ loss is monotonically decreasing in an order-wise sense: $\frac{\|\boldsymbol{w}^{(t)} - \boldsymbol{w}\|}{\|\boldsymbol{w}\|} \lesssim \frac{\|\boldsymbol{w}^{(0)} - \boldsymbol{w}\|}{\|\boldsymbol{w}\|}$ similarly to (28). By applying Lemma 3 to the regime where $L = O(\mathrm{poly}(n))$ and

$$\frac{\|\boldsymbol{w}^{(t)} - \boldsymbol{w}\|}{\|\boldsymbol{w}\|} \asymp \delta \asymp \frac{1}{2\eta - 1}\sqrt[4]{\frac{\log^2 n}{npL}}, \tag{67}$$

we get

$$\|\boldsymbol{w}^{\mathsf{mle}} - \boldsymbol{w}\|_\infty \lesssim \frac{\|\boldsymbol{w}^{(t)} - \boldsymbol{w}\|}{\|\boldsymbol{w}\|} + \frac{\log n}{np}\|\boldsymbol{w}^{(t)} - \boldsymbol{w}\|_\infty. \tag{68}$$

As in the η-known setting, one can show that the replacement threshold $\hat{\xi}_t$ leads to $\|\boldsymbol{w}^{\mathsf{mle}} - \boldsymbol{w}\|_\infty \asymp \|\boldsymbol{w}^{(t)} - \boldsymbol{w}\|_\infty$. This together with Lemma 4 gives

$$\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}\|_\infty \lesssim \frac{1}{2\eta - 1}\sqrt[4]{\frac{\log^2 n}{npL}} + \frac{\log n}{np}\|\boldsymbol{w}^{(t)} - \boldsymbol{w}\|_\infty. \tag{69}$$

A straightforward computation with this recursion yields the claimed bound as long as $\frac{\log n}{np}$ is sufficiently small (e.g., $p > \frac{2\log n}{n}$) and T is sufficiently large (e.g., $T = O(\log n)$). This completes the proof of (63).

D. Proof Sketch of Lemma 4

The proof of Lemma 4 relies on the fidelity of the estimate $\hat{\eta}$ as a function of L when we use the tensor decomposition approach by Jain and Oh [15] on the problem at hand.

Lemma 5 (Fidelity of η Estimate): If the number of observations per observed node pair L satisfies

$$L \gtrsim \frac{1}{\varepsilon^2}\log\frac{n}{\delta},, \tag{70}$$

then the estimate $\hat{\eta}$ is ε-close to the true value η with probability exceeding $1 - \delta$.

Proof: The complete proof using Theorem 3 and Lemma 6 is provided in Section VI-E. ∎

We take $\delta = n^{-c_0}$ (for some constant $c_0 > 0$) in the sequel so (70) reduces to $L \gtrsim \frac{1}{\varepsilon^2} \log n$. A major contribution in the present paper is to find a "sweet spot" for ε; if it is chosen too small, $\|\hat{\boldsymbol{w}} - \boldsymbol{w}\|_\infty$ is reduced (improving the estimation error) but L increases (worsening the overall sample complexity). Conversely, if ε is chosen to be too large, the requirement on L in (70) is relaxed, but $\|\hat{\boldsymbol{w}} - \boldsymbol{w}\|_\infty$ increases and hence, the overall sample complexity grows (worsens) eventually. The estimate in (70) is reminiscent of a Chernoff-Hoeffding bound estimate of the sample size per edge L required to ensure that the average of i.i.d. random variables is ε-close to its mean with probability $\geq 1 - \delta$. However, the justification is more involved and requires specializing Theorem 3 (to follow) to our setting.

Now, we denote the difference matrix $\Delta := \hat{P} - P$ in which $\hat{\eta}$ is used in place of η as $\hat{\Delta}$. Now using Lemma 5, several continuity arguments, and some concentration inequalities, we are able to establish that

$$\|\hat{\Delta}\| \lesssim \frac{1}{2\eta - 1} \sqrt[4]{\frac{\log^2 n}{npL}} \tag{71}$$

with probability $\geq 1 - c_1 n^{-c_2}$. The inequality (71) is proved in Appendix D. Now similarly to the proof of Lemma 1, $\rho < 1$ under the conditions of Theorem 2. Applying the bound on the spectral norm of $\|\hat{\Delta}\|$ in (71) to (34) (which continues to hold in the η-unknown setting) completes the proof of Lemma 4.

E. Proof of Lemma 5

To prove Lemma 5, we specialize the non-asymptotic bound on the recovery of parameters in a mixture model in [15] to our setting; cf. (55). Before stating this, we introduce a few notations. Let the singular value decomposition of M_2, defined in (56), be written as $M_2 = U \Sigma V^T$ where $\Sigma = \mathrm{diag}(\sigma_1(M_2), \sigma_2(M_2))$ and $U \in \mathbb{R}^{(2|\mathcal{E}|)\times 2}$ the matrix consisting of the left-singular vectors, is decomposed as

$$U = \left[((U^{(1)})^T \ (U^{(2)})^T \ \ldots \ (U^{(|\mathcal{E}|)})^T \right]^T. \tag{72}$$

Each submatrix $U^{(k)} \in \mathbb{R}^{2\times 2}$ where k denotes a node pair. We say that M_2 is *$\tilde{\mu}$-block-incoherent* if the operator norms for all $|\mathcal{E}|$ blocks of U, namely $U^{(k)}$, are upper bounded as

$$\|U^{(k)}\|_2 \leq \tilde{\mu} \sqrt{\frac{2}{|\mathcal{E}|}}, \qquad \forall\, k \in \mathcal{E}. \tag{73}$$

For M_2, the smallest block-incoherent constant $\tilde{\mu}$ is known as the *block-incoherence of M_2*. We denote this as $\mu(M_2) := \inf\{\tilde{\mu} : M_2 \text{ is } \tilde{\mu}\text{-block-incoherent}\}$.

Theorem 3 (Jain and Oh [15]): Fix any $\varepsilon, \delta > 0$. There exists a polynomial-time algorithm in $|\mathcal{E}|$, $\frac{1}{\varepsilon}$ and $\log \frac{1}{\delta}$ [15, Algorithm 1] such that if

$$|\mathcal{E}| \gtrsim \frac{\sigma_1(M_2)^{4.5} \mu(M_2)}{\sigma_2(M_2)^{4.5}} \tag{74}$$

and for a large enough (per-edge) sample size L satisfying

$$L \gtrsim \frac{\mu(M_2)\sigma_1(M_2)^6 |\mathcal{E}|^3}{\min\{\eta, 1-\eta\}\sigma_2(M_2)^9} \cdot \frac{\log(n/\delta)}{\varepsilon^2}, \tag{75}$$

the estimate of the mixture weight $\hat{\eta}$ is ε-close to the true mixture weight η with probability exceeding $1 - \delta$.

It remains to estimate the scalings of $\sigma_1(M_2), \sigma_2(M_2)$ and $\mu(M_2)$. These require calculations based on π_0, π_1 and M_2 and are summarized in the following crucial lemma.

Lemma 6: For a fixed (deterministic) sequence of graphs with $|\mathcal{E}|$ edges,

$$\sigma_i(M_2) = \Theta(|\mathcal{E}|), \quad i = 1, 2, \tag{76}$$

$$\mu(M_2) = \Theta(1). \tag{77}$$

Proof: The proof of this lemma can be found in Appendix E. It hinges on the fact that $\|\pi_0\|^2 = \|\pi_1\|^2$, as the faithful and adversarial populations have "permuted" preference score vectors. This lemma is where the assumption that the preference scores for the two populations are coupled is essential. ∎

Now the proof of Lemma 5 is immediate upon substituting (76) into (74)–(75). We then notice that $|\mathcal{E}| = \Theta(n^2 p) = \omega(1)$ with high probability so (74) is readily satisfied. Also $\frac{\mu(M_2)\sigma_1(M_2)^6|\mathcal{E}|^3}{\min\{\eta,1-\eta\}\sigma_2(M_2)^9} = \Theta(1)$ so we recover (70) as desired.

VII. Experimental Results

For the case where η is known, a number of experiments on synthetic data were conducted to validate Theorem 1. We first state parameter settings common to all experiments. The total number of items is $n = 1000$ and the number of ranked items $K = 10$. In the pointwise MLE step in Algorithm 1, we set the number of iterations $T = \lceil \log n \rceil$ and $c = 1$ in the formula for the threshold ξ_t in (20). The observation probability of each edge of the Erdős-Rényi graph is $p = \frac{6 \log n}{n}$. The latent scores are uniformly generated from the dynamic range [0.5, 1]. Each (empirical) success rate is averaged over 1000 Monte Carlo trials.

We first examine the relations between success rates and η for various values of the normalized separation of the scores $\Delta_K \in \{0.1, 0.2, \ldots, 0.5\}$. Here we consider two different scenarios, one being such that η is close to $1/2$ and the other being such that η is close to 1. We set the number of samples per edge, $L = 1000$ for the first case and $L = 10$ for the second. This is because when η is small, more data samples are needed to achieve non-negligible success rates. The results for these two scenarios are shown in Figs. 5(a) and 5(b) respectively. For both cases, when L is fixed, we observe as η increases, the success rates increase accordingly. However, the effect of η on success rates is more prominent when η is close to $1/2$. This is in accordance to (13) in Theorem 1 since $1/(2\eta - 1)^2$ has sharp decrease (as η increases) near $1/2$ and a gentler decrease near 1. Also, success rates increase when Δ_K increases. This again corroborates (13) which says that the sample complexity is proportional to $1/\Delta_K^2$.

Next we examine the relations between success rates and normalized sample size

$$S_{\text{norm}} := \frac{S_{\Delta_K}}{(n \log n)/[(2\eta - 1)^2 \Delta_K^2]}, \tag{78}$$

for $\eta \in \{0.6, 0.7, \ldots, 1\}$. We fix $\Delta_K = 0.4$ in this case. The results are shown in Fig. 6. We observe the relations between

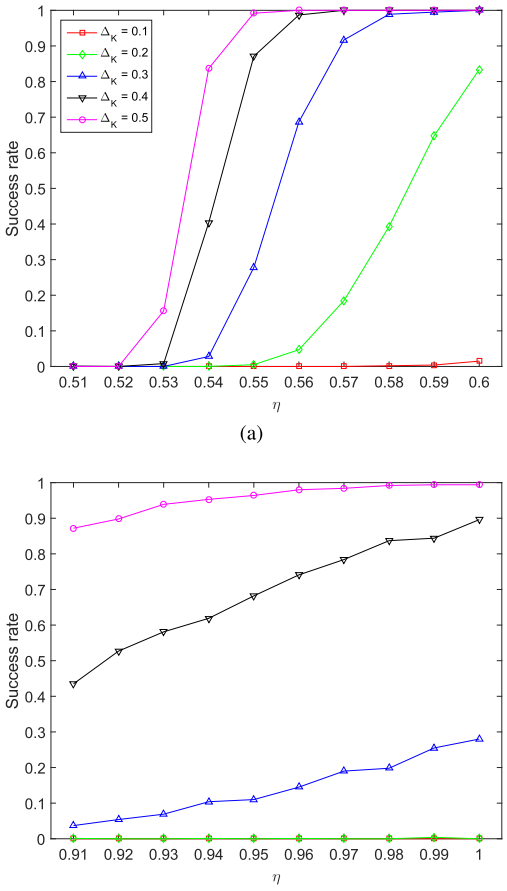

Fig. 5. Success rates across η for (a) η close to $1/2$ and (b) η close to 1.

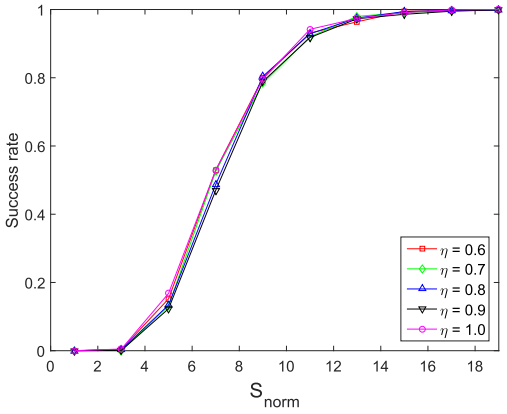

Fig. 6. Success rates across normalized sample size S_{norm}.

success rates and S_{norm} are almost the same for all η's so the implied constant factor in $\asymp$ notation in (13) depends very weakly on η (if at all).

Finally we numerically examine the relation between the sample complexity and η. We fix $\Delta_K = 0.4$ and focus on the regime where η is close to $1/2$. For each η, we use the bisection method to approximately find the minimum sample size per edge $\hat{L}$ that achieves a high success rate $q_{\text{th}} = 0.99$. Specifically, the bisection procedure terminates

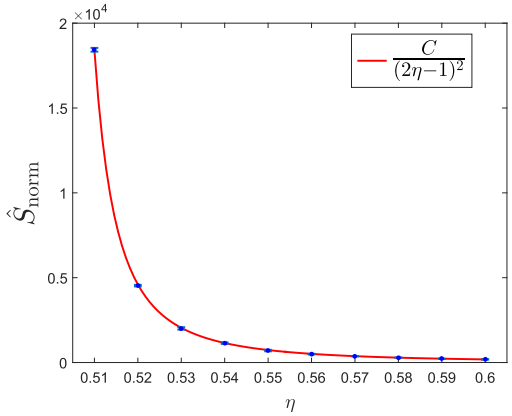

Fig. 7. Normalized empirical sample size $\hat{S}_{\text{norm}}$ for η close to $1/2$.

when the empirical success rate $\hat{q}$ corresponding to $\hat{L}$ satisfies $|\hat{q} - q_{\text{th}}| < \epsilon$, where ϵ is set to 5×10^{-3}. We repeat such a procedure 10 times to get an average result $\hat{L}_{\text{ave}}$. We also compute the resulting standard deviation and observe that it is small across the 10 independent runs. Define the expected minimum total sample size

$$\hat{S} := \binom{n}{2} p \hat{L}_{\text{ave}}. \tag{79}$$

To illustrate the explicit dependence of $\hat{S}$ on η, we further normalize $\hat{S}$ to

$$\hat{S}_{\text{norm}} := \frac{\hat{S}}{(n \log n)/\Delta_K^2}, \tag{80}$$

thus isolating the dependence of minimum total sample size on η only. We then fit a curve $C/(2\eta-1)^2$ to $\hat{S}_{\text{norm}}$, where C is chosen to best fit the points by optimizing a least-squares-like objective function. The empirical results (mean and one standard deviation) together with the fitted curve are shown in Fig. 7. We observe $\hat{S}_{\text{norm}}$ depends on η via $1/(2\eta-1)^2$ almost perfectly up to a constant. This corroborates our theoretical result in (13), i.e., the reciprocal dependence of the sample complexity on $(2\eta-1)^2$.

For the case where η is not known, the computational and storage costs turn out to be prohibitive even for a moderate number of items n (see Appendix G). Hence, we leave the implementation of the algorithm for the η-unknown case to future work. It is likely that one may need to formulate the ranking problem in an online manner [31] or resort to online methods for performing tensor decompositions [30], [32], [33]. Furthermore, existing online methods will also have to be significantly enhanced for our algorithm to be efficient.

VIII. Conclusion and Further Work

In this paper, we have provided an analytical framework for addressing the problem of recovering the top-K ranked items in an adversarial crowdsourced setting. We considered two scenarios. First, the proportion of adversaries $1-\eta$ is known and the second, more challenging scenario, is when this parameter is unknown. For the first scenario, we adapted

the *SpectralMLE* [10] and *RankCentrality* [7] algorithms to provide an order-wise optimal sample complexity bound for the total number of measurements for recovering the exact top-K set. These results were verified numerically and the dependence of the sample complexity on the reciprocal of $(2\eta-1)^2$ was corroborated. For the second scenario, we adapted Jain and Oh's global optimality result for disambiguating a mixture of discrete distributions [15] to first learn η. Subsequently, we plugged this (inexact) estimate into the known-η algorithm and utilized a sequence of continuity arguments to obtain an upper bound on the sample complexity. This bound is order-wise worse than the case where η is known, showing that the error induced by the estimation of the mixture parameter dominates the overall procedure.

A few natural questions result from our analyses.

1) The foremost concern is to narrow the gap in the sample complexities between the η-known and η-unknown scenarios. This seems challenging given that (i) threshold $\hat{\xi}_t$ in (66) must not be dependent on parameters that are assumed to be unknown such as the weight separation Δ_K and (ii) the fundamental difficulty of obtaining a globally optimal solution for the fraction of adversaries from samples that are drawn from a mixture distribution. Thus, we conjecture that if we adopt a two-step approach—first estimate η, then plug this estimate into the η-known algorithm—such a loss in the order of the sample complexity is unavoidable. This is because the fidelity of the estimate of η in Lemma 5 is natural (cf. Chernoff-Hoeffding bound) and does not seem to be order-wise improvable. We surmise that a new class of algorithms, avoiding the explicit estimation of η, needs to be developed to improve the overall sample complexity performance. Nonetheless, the analysis and results herein might shed light on the fundamental limits of machine learning problems that include a multi-stage estimation and learning procedure.
2) If closing the gap is difficult, can we hope to derive a converse or impossibility result, explicitly taking into account the fact that η is unknown? Our current converse result assumes that η is known, which may be too optimistic for the unknown setting. One possibility to strengthen the lower bound is to leverage converse techniques in *universal* source and channel coding (e.g., Clarke and Barron [34] and Beirami and Fekri [35]). Such techniques are significantly more involved than routine applications of Fano's inequality. Our situation here in which we do not know η is somewhat similar to universal decoding algorithms which do not have any knowledge of the statistical model (source or channel). Another possibility is to place a prior distribution on η (i.e., adopt a Bayesian formulation), and to use this to tighten the converse bound.
3) The tensor decomposition method [15], [16], while being polynomial time in its parameters, incurs high storage and computational costs. As discussed in Appendix G, a tractable implementation to yield meaningful estimates of η is challenging. There has been significant recent progress on large-scale scalable tensor decomposition algorithms in [30], [32], and [33]. In these works, the authors aim to avoid storing and manipulating large tensors directly but the direct adaptation of these works for our problem is still formidable. Since implementation is not the focus of the present work, we leave the development of truly tractable algorithms to future work.
4) Recent work by Shah and Wainwright [24] has shown that simple counting methods for certain observation models (including the BTL model and the mixture BTL model) achieve order-wise optimal sample complexities. In the observation model considered therein, for each pair of items i and j, there is a random number of observations R_{ij} that follows a binomial distribution with parameters $L \in \mathbb{N}$ and probability of success $p \in (0,1)$. Notice that the observation model in [24] differs from ours.
5) Lastly, it would be interesting to consider other choice models (e.g., the Plackett-Luce model [36] studied in [37] and [38]) as well as other comparison graphs not limited to the ER graph, as the comparison graph structure affects the sample complexity significantly, as suggested in [7, Th. 1].

Appendix A
Proof of Lemma 1

For ease of presentation, we will henceforth assume that $w_{\max}=1$ since this simply amounts to a rescaling of all the preference scores.

To prove the lemma, it suffices to show that if τ satisfies $|\tau - w_i| \gtrsim \max\left\{\delta + \frac{\log n}{np}\cdot\xi, \frac{1}{2\eta-1}\sqrt{\frac{\log n}{npL}}\right\}$, then the corresponding likelihood function cannot be the point-wise MLE:

$$\mathcal{L}(\tau, \hat{\boldsymbol{w}}_{\setminus i}; \boldsymbol{y}_i) < \mathcal{L}(w_i, \hat{\boldsymbol{w}}_{\setminus i}; \boldsymbol{y}_i). \tag{A.1}$$

We start by evaluating the likelihood function w.r.t. the ground-truth score vector:

$$\ell^*(\tau) := \frac{1}{L}\log \mathcal{L}(\tau, \boldsymbol{w}_{\setminus i}; \boldsymbol{Y}_i) \tag{A.2}$$

$$= \sum_{j:(i,j)\in\mathcal{E}} \left\{ Y_{ij}\log\left(\eta\frac{\tau}{\tau+w_j} + (1-\eta)\frac{w_j}{\tau+w_j}\right) + (1-Y_{ij})\log\left(\eta\frac{w_j}{\tau+w_j} + (1-\eta)\frac{\tau}{\tau+w_j}\right)\right\}. \tag{A.3}$$

The likelihood loss w.r.t. w_i and τ is then computed as:

$$\ell^*(w_i) - \ell^*(\tau) = \sum_{j:(i,j)\in\mathcal{E}} \left\{ Y_{ij}\log\left(\frac{\eta\frac{w_i}{w_i+w_j} + (1-\eta)\frac{w_j}{w_i+w_j}}{\eta\frac{\tau}{\tau+w_j} + (1-\eta)\frac{w_j}{\tau+w_j}}\right) + (1-Y_{ij})\log\left(\frac{\eta\frac{w_j}{w_i+w_j} + (1-\eta)\frac{w_i}{w_i+w_j}}{\eta\frac{w_j}{\tau+w_j} + (1-\eta)\frac{\tau}{\tau+w_j}}\right)\right\}. \tag{A.4}$$

Taking expectation w.r.t. $\boldsymbol{Y}_i$ conditional on $\mathcal{G}$, we get:

$$\begin{aligned}&\mathbb{E}\left[\ell^*(w_i)-\ell^*(\tau)|\mathcal{G}\right]\\&=\sum_{j:(i,j)\in\mathcal{E}} D\left(\eta\frac{w_i}{w_i+w_j}+(1-\eta)\frac{w_j}{w_i+w_j}\middle\|\right.\\&\qquad\left.\eta\frac{\tau}{\tau+w_j}+(1-\eta)\frac{w_j}{\tau+w_j}\right) \quad (A.5)\\&\overset{(a)}{\gtrsim} np(2\eta-1)^2|w_i-\tau|^2 \quad (A.6)\end{aligned}$$

where (a) follows from Pinsker's inequality ($D(p\|q)\geq 2(p-q)^2$; see [39, Th. 2.33] for example) and using the fact that $d_i \asymp np$ when $p>\frac{\log n}{n}$. Here d_i indicates the degree of node i: the number of edges incident to node i. This suggests that the true point-wise MLE of w_i strictly dominates that of τ in the *mean* sense. We can actually demonstrate that this is the case beyond the mean sense with high probability, as long as $|w_i-\tau|\gtrsim \frac{1}{2\eta-1}\sqrt{\frac{\log n}{npL}}$ (our hypothesis), which is asserted in the following lemma.

Lemma 7: Suppose that $|w_i-\tau|\gtrsim \frac{1}{2\eta-1}\sqrt{\frac{\log n}{npL}}$. Then,

$$\ell^*(w_i)-\ell^*(\tau)\gtrsim np(2\eta-1)^2|w_i-\tau|^2. \quad (A.7)$$

holds with probability approaching one.

Proof: Using Bernstein's inequality formally stated in Lemma 12 (see Appendix F), one can obtain a lower bound on $\ell^*(w_i)-\ell^*(\tau)$ in terms of its expectation $\mathbb{E}\left[\ell^*(w_i)-\ell^*(\tau)|\mathcal{G}\right]$, its variance $\mathsf{Var}\left[\ell^*(w_i)-\ell^*(\tau)|\mathcal{G}\right]$, and the maximum value of individual quantities that we sum over. One can then show that the variance and the maximum value are dominated by the expectation under our hypothesis, thus proving that the lower bound is the order of the desired bound as claimed. For completeness, we include the detailed proof at the end of this appendix; see Appendix A-A. ∎

However, when running our algorithm, we do not have access to the ground truth scores $\boldsymbol{w}_{\setminus i}$. What we can actually compute is

$$\hat{\ell}(\tau):=\frac{1}{L}\log\mathcal{L}(\tau,\hat{\boldsymbol{w}}_{\setminus i};\boldsymbol{Y}_i) \quad (A.8)$$

instead of $\ell^*(\tau)$. Fortunately, such surrogate likelihoods are sufficiently close to the true likelihoods, which we will show in the rest of the proof. From this, we will next demonstrate that (A.1) holds for sufficiently separated τ such that $|\tau-w_i|\gtrsim \max\left\{\delta+\frac{\log n}{np}\cdot\xi, \frac{1}{2\eta-1}\sqrt{\frac{\log n}{npL}}\right\}$.

As seen from (A.31), one can quantify the difference between $\hat{\ell}(w_i)$ and $\hat{\ell}(\tau)$ as

$$\begin{aligned}&\hat{\ell}(w_i)-\hat{\ell}(\tau)\\&=\sum_{j:(i,j)\in\mathcal{E}}\left\{Y_{ij}\log\left\{\frac{(\eta w_i+(1-\eta)\hat{w}_j)(\eta\hat{w}_j+(1-\eta)\tau)}{(\eta\tau+(1-\eta)\hat{w}_j)(\eta\hat{w}_j+(1-\eta)w_i)}\right\}\right.\\&\quad\left.+\log\left(\frac{(\tau+\hat{w}_j)(\eta\hat{w}_j+(1-\eta)w_i)}{(w_i+\hat{w}_j)(\eta\hat{w}_j+(1-\eta)\tau)}\right)\right\}. \quad (A.9)\end{aligned}$$

Using (A.9) and (A.31), we can represent the gap between the surrogate loss and the true loss as

$$\begin{aligned}&\hat{\ell}(w_i)-\hat{\ell}(\tau)-(\ell^*(w_i)-\ell^*(\tau))\\&=\sum_{j:(i,j)\in\mathcal{E}}Y_{ij}\left[\left\{\log\left\{\frac{(\eta w_i+(1-\eta)\hat{w}_j)(\eta\hat{w}_j+(1-\eta)\tau)}{(\eta\tau+(1-\eta)\hat{w}_j)(\eta\hat{w}_j+(1-\eta)w_i)}\right\}\right.\right.\\&\quad\left.-\log\left\{\frac{(\eta w_i+(1-\eta)w_j)(\eta w_j+(1-\eta)\tau)}{(\eta\tau+(1-\eta)w_j)(\eta w_j+(1-\eta)w_i)}\right\}\right\}\\&\quad+\left\{\log\left(\frac{\tau+\hat{w}_j}{w_i+\hat{w}_j}\right)+\log\left(\frac{\eta\hat{w}_j+(1-\eta)w_i}{\eta\hat{w}_j+(1-\eta)\tau}\right)\right.\\&\quad\left.\left.-\log\left(\frac{\tau+w_j}{w_i+w_j}\right)-\log\left(\frac{\eta w_j+(1-\eta)w_i}{\eta w_j+(1-\eta)\tau}\right)\right\}\right]. \quad (A.10)\end{aligned}$$

Using Bernstein's inequality under our hypothesis as we did in Lemma 7, one can verify that

$$\begin{aligned}&\hat{\ell}(w_i)-\hat{\ell}(\tau)-(\ell^*(w_i)-\ell^*(\tau))\\&\lesssim\mathbb{E}\left[\hat{\ell}(w_i)-\hat{\ell}(\tau)-(\ell^*(w_i)-\ell^*(\tau))|\mathcal{G}\right]\\&=\sum_{j:(i,j)\in\mathcal{E}}g_\eta(\hat{w}_j) \quad (A.11)\end{aligned}$$

where

$$\begin{aligned}&g_\eta(t)\\&:=\frac{\eta w_i+(1-\eta)w_j}{w_i+w_j}\left\{\log\left\{\frac{(\eta w_i+(1-\eta)t)(\eta t+(1-\eta)\tau)}{(\eta\tau+(1-\eta)t)(\eta t+(1-\eta)w_i)}\right\}\right.\\&\quad\left.-\log\left\{\frac{(\eta w_i+(1-\eta)w_j)(\eta w_j+(1-\eta)\tau)}{(\eta\tau+(1-\eta)w_j)(\eta w_j+(1-\eta)w_i)}\right\}\right\}\\&\quad+\log\left(\frac{\tau+t}{w_i+t}\right)+\log\left(\frac{\eta t+(1-\eta)w_i}{\eta t+(1-\eta)\tau}\right)\\&\quad-\log\left(\frac{\tau+w_j}{w_i+w_j}\right)-\log\left(\frac{\eta w_j+(1-\eta)w_i}{\eta w_j+(1-\eta)\tau}\right). \quad (A.12)\end{aligned}$$

Here the function $g_\eta(t)$ obeys the following two properties: (i) $g_\eta(w_j)=0$ and (ii) the derivative satisfies

$$\begin{aligned}&\left|\frac{\partial g_\eta(t)}{\partial t}\right|\\&=\frac{(2\eta-1)|\tau-w_i|}{(\eta t+(1-\eta)\tau)(\eta t+(1-\eta)w_i)}\\&\quad\times\left|\frac{\eta w_i+(1-\eta)w_j}{w_i+w_j}\cdot\frac{\eta(1-\eta)(t^2-\tau w_i)}{(\eta w_i+(1-\eta)t)(\eta\tau+(1-\eta)t)}\right.\\&\quad\left.-\frac{\eta t^2-(1-\eta)w_i\tau}{(\tau+t)(w_i+t)}\right| \quad (A.13)\\&\overset{(a)}{\lesssim}(2\eta-1)^2|\tau-w_i| \quad (A.14)\end{aligned}$$

where (a) follows from the fact that

$$\begin{aligned}&\left|\frac{\eta w_i+(1-\eta)w_j}{w_i+w_j}\cdot\frac{\eta(1-\eta)(t^2-\tau w_i)}{(\eta w_i+(1-\eta)t)(\eta\tau+(1-\eta)t)}\right.\\&\quad\left.-\frac{\eta t^2-(1-\eta)w_i\tau}{(\tau+t)(w_i+t)}\right|\lesssim(2\eta-1). \quad (A.15)\end{aligned}$$

Notice that the left-hand-side in the above is zero when $\eta=1/2$. This together with the above two properties

demonstrates that

$$|g_\eta(t)| \le |g_\eta(w_j)| + |t - w_j| \cdot \sup_{t\in[w_{\min}, w_{\max}]} \left|\frac{\partial g_\eta(t)}{\partial t}\right| \quad \text{(A.16)}$$

$$\lesssim (2\eta-1)^2|\tau - w_i||t - w_j|. \quad \text{(A.17)}$$

Applying this to the above gap between the surrogate loss and the true loss, we get:

$$\left|\hat{\ell}(w_i) - \hat{\ell}(\tau) - (\ell^*(w_i) - \ell^*(\tau))\right| \lesssim \sum_{j:(i,j)\in\mathcal{E}} (2\eta-1)^2|\tau - w_i||\hat{w}_j - w_j| \quad \text{(A.18)}$$

$$\le (2\eta-1)^2|\tau - w_i| \sum_{j:(i,j)\in\mathcal{E}} |\hat{w}_j^{\text{ub}} - w_j| \quad \text{(A.19)}$$

where the inequality arises from our hypothesis, namely that $|\hat{w}_j - w_j| \le |\hat{w}_j^{\text{ub}} - w_j|$ for all $j \in [n]$.

We now move on to deriving an upper bound on (A.19). From our assumptions on the initial estimate, we have

$$\|\hat{\boldsymbol{w}} - \boldsymbol{w}\|^2 \le \|\hat{\boldsymbol{w}}^{\text{ub}} - \boldsymbol{w}\|^2 \le \|\boldsymbol{w}\|^2\delta^2 \le n\delta^2. \quad \text{(A.20)}$$

Since $\mathcal{G}$ and $\hat{\boldsymbol{w}}^{\text{ub}}$ are statistically independent,

$$\mathbb{E}\left[\sum_{j:(i,j)\in\mathcal{E}} |\hat{w}_j^{\text{ub}} - w_j|\right] = p\|\hat{\boldsymbol{w}}^{\text{ub}} - \boldsymbol{w}\|_1 \le p\sqrt{n}\|\hat{\boldsymbol{w}}^{\text{ub}} - \boldsymbol{w}\| \le np\delta, \quad \text{(A.21)}$$

$$\mathbb{E}\left[\sum_{j:(i,j)\in\mathcal{E}} |\hat{w}_j^{\text{ub}} - w_j|^2\right] = p\|\hat{\boldsymbol{w}}^{\text{ub}} - \boldsymbol{w}\|^2 \le np\delta^2. \quad \text{(A.22)}$$

Recall our assumption that $\max_j |\hat{w}_j^{\text{ub}} - w_j| \le \xi$. Again using Bernstein inequality in Lemma 12 for any fixed $\gamma \ge 3$, with probability at least $1 - 2n^{-\gamma}$, one has

$$\sum_{j:(i,j)\in\mathcal{E}} |\hat{w}_j^{\text{ub}} - w_j| \le \mathbb{E}\left[\sum_{j:(i,j)\in\mathcal{E}} |\hat{w}_j^{\text{ub}} - w_j|\right] + \sqrt{2\gamma\,\log n \cdot \mathbb{E}\left[\sum_{j:(i,j)\in\mathcal{E}} |\hat{w}_j^{\text{ub}} - w_j|^2\right]} + \frac{2\gamma}{3}\xi\log n \quad \text{(A.23)}$$

$$\le np\delta + \sqrt{2\gamma\, np\log n}\delta + \frac{2\gamma}{3}\xi\log n \quad \text{(A.24)}$$

$$\overset{(a)}{\le} np\delta + \sqrt{\gamma}np\delta + \frac{2\gamma}{3}\xi\log n \quad \text{(A.25)}$$

$$\overset{(b)}{\le} \gamma\, np\delta + \gamma\,\xi\log n \quad \text{(A.26)}$$

where (a) follows from our choice on p (we assume $p > \frac{2\log n}{n}$) and (b) follows from the fact that $1 + \sqrt{\gamma} \le \gamma$ for $\gamma \ge 3$. This combined with (A.19) gives us

$$\left|\hat{\ell}(w_i) - \hat{\ell}(\tau) - (\ell^*(w_i) - \ell^*(\tau))\right| \lesssim (2\eta-1)^2|\tau - w_i|np\left(\delta + \frac{\log n}{np}\xi\right). \quad \text{(A.27)}$$

We are now ready to control $\hat{\ell}(w_i) - \hat{\ell}(\tau)$. Putting (A.7) and (A.27) together, with probability approaching one, one has

$$\hat{\ell}(w_i) - \hat{\ell}(\tau) \gtrsim \ell^*(w_i) - \ell^*(\tau) - (2\eta-1)^2|\tau - w_i|np\left(\delta + \frac{\log n}{np}\xi\right) \quad \text{(A.28)}$$

$$\gtrsim np(2\eta-1)^2|w_i - \tau|^2 - (2\eta-1)^2|\tau - w_i|np\left(\delta + \frac{\log n}{np}\xi\right) \quad \text{(A.29)}$$

$$\gtrsim 0 \quad \text{(A.30)}$$

where the last step follows from our hypothesis: $|w_i - \tau| \gtrsim \delta + \frac{\log n}{np}\xi$. This completes the proof of Lemma 1.

A. Proof of Lemma 7

Another representation of the true loss is:

$$\ell^*(w_i) - \ell^*(\tau) = \sum_{j:(i,j)\in\mathcal{E}} \left\{ Y_{ij} \log\left\{\frac{(\eta w_i + (1-\eta)w_j)(\eta w_j + (1-\eta)\tau)}{(\eta\tau + (1-\eta)w_j)(\eta w_j + (1-\eta)w_i)}\right\} + \log\left(\frac{(\tau + w_j)(\eta w_j + (1-\eta)w_i)}{(w_i + w_j)(\eta w_j + (1-\eta)\tau)}\right)\right\} \quad \text{(A.31)}$$

This gives

$$\mathsf{Var}\left[\ell^*(w_i) - \ell^*(\tau)|\mathcal{G}\right] = \mathsf{Var}\left[\sum_{j:(i,j)\in\mathcal{E}} Y_{ij} \times \log\left\{\frac{(\eta w_i + (1-\eta)w_j)(\eta w_j + (1-\eta)\tau)}{(\eta\tau + (1-\eta)w_j)(\eta w_j + (1-\eta)w_i)}\right\}\right] \quad \text{(A.32)}$$

$$\overset{(a)}{\lesssim} |w_i - \tau|^2(2\eta-1)^2 \sum_{j:(i,j)\in\mathcal{E}} \mathsf{Var}[Y_{ij}] \quad \text{(A.33)}$$

$$= |w_i - \tau|^2(2\eta-1)^2 \times \sum_{j:(i,j)\in\mathcal{E}} \frac{(\eta w_i + (1-\eta)w_j)(\eta w_j + (1-\eta)w_i)}{L(w_i + w_j)^2} \quad \text{(A.34)}$$

$$\lesssim |w_i - \tau|^2(2\eta-1)^2\frac{np}{L} \quad \text{(A.35)}$$

where (a) follows from the fact that $\log\frac{\beta}{\alpha} \le \frac{\beta-\alpha}{\alpha}$ for $\beta > \alpha > 0$. Also note that the maximum value of individual quantities $\frac{1}{L}Y_{ij}^{(\ell)}$ that we sum over is given by

$$\frac{1}{L}Y_{ij}^{(\ell)}\left|\log\left\{\frac{(\eta w_i + (1-\eta)w_j)(\eta w_j + (1-\eta)\tau)}{(\eta\tau + (1-\eta)w_j)(\eta w_j + (1-\eta)w_i)}\right\}\right| \lesssim \frac{|w_i - \tau|(2\eta-1)}{L}. \quad \text{(A.36)}$$

Making use of Bernstein's inequality together with (A.5),

(A.35) and (A.36) implies that conditional on $\mathcal{G}$,

$$\begin{aligned}
&\ell^*(w_i) - \ell^*(\tau) \\
&\geq \mathbb{E}\left[\ell^*(w_i) - \ell^*(\tau)|\mathcal{G}\right] \\
&\quad -\sqrt{2\gamma\,\log n \cdot \mathsf{Var}\left[\ell^*(w_i) - \ell^*(\tau)|\mathcal{G}\right]} - \frac{2\gamma}{3} B \log n \quad \text{(A.37)} \\
&\gtrsim np(2\eta-1)^2|w_i-\tau|^2 - \sqrt{2\gamma}\sqrt{\frac{np\log n}{L}}|w_i-\tau|(2\eta-1) \\
&\quad -\frac{2\gamma}{3}\frac{|w_i-\tau|(2\eta-1)}{L}\log n \quad \text{(A.38)} \\
&\geq np(2\eta-1)^2|w_i-\tau|^2 \\
&\quad -\left(\sqrt{2\gamma}+\frac{2\gamma}{3}\right)\sqrt{\frac{np\log n}{L}}|w_i-\tau|(2\eta-1) \quad \text{(A.39)} \\
&\overset{(a)}{\gtrsim} np(2\eta-1)^2|w_i-\tau|^2 \quad \text{(A.40)}
\end{aligned}$$

holds with probability at least $1-2n^{-\gamma}$. Here (a) follows from our hypothesis: $|w_i-\tau| \gtrsim \frac{1}{2\eta-1}\sqrt{\frac{\log n}{npL}}$.

Appendix B
Proof of Lemma 2

As mentioned earlier, the proof builds upon the analysis structured by [7, Lemma 2], which bounds the deviation of the Markov chain w.r.t. the transition matrix $\hat{P}$ (defined in Algorithm 2) after t steps:

$$\frac{\|\hat{p}_t - \boldsymbol{w}\|}{\|\boldsymbol{w}\|} \leq \rho^t \frac{\|\hat{p}_0 - \boldsymbol{w}\|}{\|\boldsymbol{w}\|}\sqrt{\frac{w_{\max}}{w_{\min}}} + \frac{1}{1-\rho}\|\Delta\|\sqrt{\frac{w_{\max}}{w_{\min}}} \tag{B.1}$$

where $\hat{p}_t$ denotes the distribution w.r.t. $\hat{P}$ at time t seeded by an arbitrary initial distribution $\hat{p}_0$, the matrix $\Delta := \hat{P} - P$ indicates the fluctuation of the transition probability matrix around its mean $P := \mathbb{E}[\hat{P}]$, and $\rho := \lambda_{\max} + \|\Delta\|\sqrt{\frac{w_{\max}}{w_{\min}}}$. Here $\lambda_{\max} = \max\{\lambda_2, -\lambda_n\}$ and λ_i indicates the i-th eigenvalue of P.

For an arbitrary η case, a bound on $\|\Delta\|$ is:

$$\|\Delta\| \lesssim \frac{1}{2\eta-1}\sqrt{\frac{\log n}{npL}} \tag{B.2}$$

which will be proved in the sequel. On the other hand, adapting the analysis in [7] (particularly see Lemma 4 in the reference), one can easily verify that $\rho < 1$ under our assumption that $Lnp \gtrsim \frac{\log n}{(2\eta-1)^2}$. Applying the bound on $\|\Delta\|$ and $\rho < 1$ to the above gives the claimed bound, which completes the proof.

Let us now prove the bound on $\|\Delta\|$, which is a generalization of the proof in [7]. Let D be a diagonal matrix with $D_{ii} := \Delta_{ii}$. Let $\bar{\Delta} := \Delta - D$. Note that

$$\|\Delta\| \leq \|D\| + \|\bar{\Delta}\| = \max_i |\Delta_{ii}| + \|\bar{\Delta}\|. \tag{B.3}$$

We will use Hoeffding inequality to bound $|\Delta_{ii}|$. As for $\|\bar{\Delta}\|$, we will focus on bounds of $\mathbb{E}[|\Delta_{ij}|^p]$, since Tropp inequality in [27] turns out to relate the bound of $\mathbb{E}[|\Delta_{ij}|^p]$ to that of $\|\bar{\Delta}\|$, as pointed out in [7]. Hence, here we provide derivations mainly for the bounds on $|\Delta_{ii}|$ and $\mathbb{E}[|\Delta_{ij}|^p]$. Later we will appeal to a relationship between $\|\bar{\Delta}\|$ and $\mathbb{E}[|\Delta_{ij}|^p]$, formally stated in Lemma 8 (see below), to prove the desired bound on $\|\bar{\Delta}\|$.

Bounding $|\Delta_{ii}|$: Observe that

$$\begin{aligned}
Ld_{\max}\Delta_{ii} &= -Ld_{\max}\sum_{k\neq i}\Delta_{ik} \\
&= -\sum_{k\neq i}\sum_{\ell=1}^{L}\left(\frac{Y_{ki}^{(\ell)} - (1-\eta)}{2\eta-1} - \frac{w_k}{w_i+w_k}\right). \quad \text{(B.4)}
\end{aligned}$$

Let $X_{k\ell} := \frac{Y_{ki}^{(\ell)}-(1-\eta)}{2\eta-1} - \frac{w_k}{w_i+w_k}$. Then, we have $\mathbb{E}[X_{k\ell}] = 0$ and $-\frac{\eta+1}{2\eta-1} \leq X_{k\ell} \leq \frac{\eta}{2\eta-1}$. Using Hoeffding inequality, we obtain:

$$\begin{aligned}
\Pr\left[|Ld_{\max}\Delta_{ii}| \geq t\right] &\leq 2\exp\left(-\frac{2(t\frac{2\eta-1}{2\eta+1})^2}{Ld_i}\right) \\
&\leq 2\exp\left(-\frac{2(t\frac{2\eta-1}{2\eta+1})^2}{Ld_{\max}}\right). \quad \text{(B.5)}
\end{aligned}$$

Choosing $t = c\sqrt{Ld_{\max}\log n}\left(\frac{2\eta+1}{2\eta-1}\right)$ for some $c > 0$, one can make the tail bound arbitrarily close to zero in the limit of large n. Also $d_{\max} \asymp np$ when $p > \frac{\log n}{n}$. Hence, with probability approaching one, one has $\|D\| \lesssim \frac{1}{2\eta-1}\sqrt{\frac{\log n}{npL}}$.

Bounding $\|\bar{\Delta}\|$: A careful inspection reveals that

$$\bar{\Delta} = \sum_{i<j:(i,j)\in\mathcal{E}}(e_ie_j^T - e_je_i^T)(\hat{p}_{ij} - p_{ij}) \tag{B.6}$$

where e_i denotes the standard basis vector in which only the i-th entry is 1 while the others are zeros. Here with a slight abuse of notation, we use $\mathcal{E}$ to indicate $\mathcal{E}^{\mathsf{init}}$. As mentioned earlier, we intend to make use of the concentration result by Tropp [27] for sum of independent self-adjoint matrices. To this end, we apply the dilation idea in [27] for symmetrization:

$$Z_{ij} := A_{ij}\Delta_{ij} := \begin{bmatrix} 0 & e_ie_j^T - e_je_i^T \\ e_je_i^T - e_ie_j^T & 0 \end{bmatrix}\Delta_{ij}. \tag{B.7}$$

Note that

$$\begin{aligned}
\|\bar{\Delta}\| &= \left\|\sum_{i<j:(i,j)\in\mathcal{E}}(e_ie_j^T - e_je_i^T)(\hat{p}_{ij} - p_{ij})\right\| \\
&= \left\|\sum_{i<j:(i,j)\in\mathcal{E}}A_{ij}\Delta_{ij}\right\| = \left\|\sum_{i<j:(i,j)\in\mathcal{E}}Z_{ij}\right\|. \quad \text{(B.8)}
\end{aligned}$$

We now invoke Tropp's inequality formally stated in the following lemma.

Lemma 8: Consider a sequence Z_{ij} of independent random self-adjoint matrices. Assume that

$$\mathbb{E}[Z_{ij}] = 0 \quad \textit{and} \quad \mathbb{E}[Z_{ij}^p] \preceq \frac{p!}{2}R^{p-2}\tilde{A}_{ij}^2, \quad p \geq 2. \tag{B.9}$$

Define $\sigma^2 := \left\|\sum_{i,j}\tilde{A}_{ij}^2\right\|$. *Then, for all* $t \geq 0$,

$$\Pr\left[\left\|\sum_{i,j}Z_{ij}\right\| \geq t\right] \leq \exp\left(-\frac{t^2/2}{\sigma^2 + Rt}\right). \tag{B.10}$$

To figure out what $\tilde{A}_{ij}$, σ^2 and R are, we consider

$$\mathbb{E}[Z_{ij}^p] \overset{(a)}{\leq} \mathbb{E}[\Delta_{ij}^p] A_{ij}^2 \tag{B.11}$$

$$\overset{(b)}{\leq} \frac{p!}{2}\left(\frac{2\eta+1}{2\eta-1}\frac{1}{\sqrt{Ld_{\max}^2}}\right)^{p-2}\frac{2\eta-1}{2\eta+1}\frac{1}{Ld_{\max}^2}A_{ij}^2. \tag{B.12}$$

To see (a), note that A_{ij}^p is equal to A_{ij}^2 when p is even; A_{ij} otherwise. Also one can verify that the eigenvalues of A_{ij} are either 1 or -1. Hence, $A_{ij}^p \preceq A_{ij}^2$. To see (b), observe that

$$Ld_{\max}\Delta_{ij} = \sum_{\ell=1}^{L}\left(\frac{Y_{ji}^{(\ell)} - (1-\eta)}{2\eta-1} - \frac{w_j}{w_i+w_j}\right). \tag{B.13}$$

Applying Hoeffding inequality into the term inside the summation, we get

$$\Pr\left[|Ld_{\max}\Delta_{ij}| \geq t\right] \leq 2\exp\left(-\frac{2(t\frac{2\eta-1}{2\eta+1})^2}{L}\right), \tag{B.14}$$

which yields

$$\Pr\left[|\Delta_{ij}| \geq t\right] \leq 2\exp\left(-2t^2\left(\frac{2\eta-1}{2\eta+1}\right)^2 Ld_{\max}^2\right). \tag{B.15}$$

This implies that Δ_{ij} is a sub-Gaussian random variable. Hence, we obtain the bound

$$\mathbb{E}\left[|\Delta_{ij}|^p\right] \leq \frac{p!}{2}\left(\frac{2\eta+1}{2\eta-1}\frac{1}{\sqrt{Ld_{\max}^2}}\right)^p, \tag{B.16}$$

which yields (b) in (B.12).

We now see that $R = \frac{2\eta+1}{2\eta-1}\frac{1}{\sqrt{Ld_{\max}^2}}$ and $\tilde{A}_{ij}^2 = \frac{2\eta-1}{2\eta+1}\frac{1}{Ld_{\max}^2}A_{ij}^2$. Some calculations yield

$$\sigma_2 := \left\|\sum_{i<j:(i,j)\in\mathcal{E}}\tilde{A}_{ij}^2\right\| \tag{B.17}$$

$$= \frac{2\eta-1}{2\eta+1}\frac{1}{Ld_{\max}^2}\left\|\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathbf{1}\{(i,j)\in\mathcal{E}\}\times\begin{bmatrix} e_ie_i^T + e_je_j^T & 0 \\ 0 & e_ie_i^T + e_je_j^T\end{bmatrix}\right\| \tag{B.18}$$

$$= \frac{2\eta-1}{2\eta+1}\frac{1}{Ld_{\max}^2}\left\|\sum_{i=1}^{n} d_i\begin{bmatrix} e_ie_i^T & 0 \\ 0 & e_ie_i^T\end{bmatrix}\right\| \tag{B.19}$$

$$= \frac{2\eta-1}{2\eta+1}\frac{1}{Ld_{\max}}. \tag{B.20}$$

Now applying Lemma 8 and using the fact that $\|\bar{\Delta}\| = \|\sum_{i,j} Z_{ij}\|$, we get:

$$\Pr\left[\|\bar{\Delta}\| \geq t\right] \leq 2n\exp\left(\frac{-t^2/2}{\frac{(2\eta+1)^2}{Ld_{\max}(2\eta-1)^2} + \frac{(2\eta+1)t}{\sqrt{Ld_{\max}^2(2\eta-1)^2}}}\right). \tag{B.21}$$

Under the assumption that $d_{\max} \asymp np \geq \log n$ and choosing $t = \frac{c_1}{2\eta-1}\sqrt{\log n/(npL)}$, the tail probability is bounded by $2n\exp\{-c_2^2\log n\}$ for some constants c_1 and c_2. Hence, with probability approaching one, we get the desired bound:

$$\|\bar{\Delta}\| \lesssim \frac{1}{2\eta-1}\sqrt{\frac{\log n}{npL}}. \tag{B.22}$$

Appendix C
Proof of Lemma 3

Proof: In this proof, for the sake of brevity, we only highlight the parts of the proof of Lemma 1 that have to be modified when we use $\hat{\eta}$ in place of η in the likelihood function.

Define

$$\kappa^*(\tau) := \frac{1}{L}\log\hat{\mathcal{L}}(\tau, \boldsymbol{w}_{\setminus i}; \boldsymbol{Y}_i^{\text{iter}}) \tag{C.1}$$

$$= \sum_{j:(i,j)\in\mathcal{E}}\left\{Y_{ij}\log\left(\hat{\eta}\frac{\tau}{\tau+w_j} + (1-\hat{\eta})\frac{w_j}{\tau+w_j}\right) + (1-Y_{ij})\log\left(\hat{\eta}\frac{w_j}{\tau+w_j} + (1-\hat{\eta})\frac{\tau}{\tau+w_j}\right)\right\}. \tag{C.2}$$

Notice that κ^* is similar to ℓ^* in (A.3) except that η in the latter is replaced by its surrogate $\hat{\eta}$ in the former because we only have access to this estimate.

Consider the difference

$$\kappa^*(w_i) - \kappa^*(\tau) = \sum_{j:(i,j)\in\mathcal{E}}\left\{Y_{ij}\log\left(\frac{\hat{\eta}\frac{w_i}{w_i+w_j} + (1-\hat{\eta})\frac{w_j}{w_i+w_j}}{\hat{\eta}\frac{\tau}{\tau+w_j} + (1-\hat{\eta})\frac{w_j}{\tau+w_j}}\right) + (1-Y_{ij})\log\left(\frac{\hat{\eta}\frac{w_j}{w_i+w_j} + (1-\hat{\eta})\frac{w_i}{w_i+w_j}}{\hat{\eta}\frac{w_j}{\tau+w_j} + (1-\hat{\eta})\frac{\tau}{\tau+w_j}}\right)\right\}. \tag{C.3}$$

Now when we take expectation

$$\mathbb{E}[Y_{ij}] = \eta\frac{w_i}{w_i+w_j} + (1-\eta)\frac{w_j}{w_i+w_j}. \tag{C.4}$$

Note that this is in terms of η and not $\hat{\eta}$ as in the difference of the empirical log-likelihoods in (C.3). In particular, $\mathbb{E}[\kappa^*(w_i) - \kappa^*(\tau)\,|\,\mathcal{G}]$ is not a sum of KL divergences but instead there is some "mismatch". However, by some basic approximations, we have

$$\mathbb{E}\left[\kappa^*(w_i) - \kappa^*(\tau)\,\middle|\,\mathcal{G}\right] = \sum_{j:(i,j)\in\mathcal{E}}\left\{\left(\eta\frac{w_i}{w_i+w_j} + (1-\eta)\frac{w_j}{w_i+w_j}\right)\times\log\left(\frac{\hat{\eta}\frac{w_i}{w_i+w_j} + (1-\hat{\eta})\frac{w_j}{w_i+w_j}}{\hat{\eta}\frac{\tau}{\tau+w_j} + (1-\hat{\eta})\frac{w_j}{\tau+w_j}}\right) + \left(\eta\frac{w_j}{w_i+w_j} + (1-\eta)\frac{w_i}{w_i+w_j}\right)\times\log\left(\frac{\hat{\eta}\frac{w_j}{w_i+w_j} + (1-\hat{\eta})\frac{w_i}{w_i+w_j}}{\hat{\eta}\frac{w_j}{\tau+w_j} + (1-\hat{\eta})\frac{\tau}{\tau+w_j}}\right)\right\} \tag{C.5}$$

$$\gtrsim \sum_{j:(i,j)\in\mathcal{E}} \Bigg\{ \left(\hat{\eta}\frac{w_i}{w_i+w_j} + (1-\hat{\eta})\frac{w_j}{w_i+w_j}\right) \times \log\left(\frac{\hat{\eta}\frac{w_i}{w_i+w_j} + (1-\hat{\eta})\frac{w_j}{w_i+w_j}}{\hat{\eta}\frac{\tau}{\tau+w_j} + (1-\hat{\eta})\frac{w_j}{\tau+w_j}}\right) + \left(\hat{\eta}\frac{w_j}{w_i+w_j} + (1-\hat{\eta})\frac{w_i}{w_i+w_j}\right) \times \log\left(\frac{\hat{\eta}\frac{w_j}{w_i+w_j} + (1-\hat{\eta})\frac{w_i}{w_i+w_j}}{\hat{\eta}\frac{w_j}{\tau+w_j} + (1-\hat{\eta})\frac{\tau}{\tau+w_j}}\right)\Bigg\} \tag{C.6}$$

$$= \sum_{j:(i,j)\in\mathcal{E}} D\left(\hat{\eta}\frac{w_i}{w_i+w_j} + (1-\hat{\eta})\frac{w_j}{w_i+w_j} \,\middle\|\, \hat{\eta}\frac{\tau}{\tau+w_j} + (1-\hat{\eta})\frac{w_j}{\tau+w_j}\right) \tag{C.7}$$

$$\gtrsim np(2\hat{\eta}-1)^2|w_i-\tau|^2 \tag{C.8}$$

where

1) (C.5) follows from the difference of κ^*'s in (C.3) and the expectation in (C.4);
2) (C.6) holds with high probability (guaranteed by the sample complexity bound in Theorem 2) by multiplicatively and uniformly approximating $\eta w_i + (1-\eta)w_j$ by $\hat{\eta}w_i + (1-\hat{\eta})w_j$ and $\eta w_j + (1-\eta)w_i$ by $\hat{\eta}w_j + (1-\hat{\eta})w_i$ using Lemma 9 (in Appendix C-A at the end of this appendix) with constant $\nu = 0.1$ (say);
3) (C.8) is an application of Pinsker's inequality [39, Th. 2.33].

The punchline in this calculation is that with our choice of parameters, the scaling of the lower bound of $\mathbb{E}[\kappa^*(w_i) - \kappa^*(\tau)\,|\,\mathcal{G}]$ is the same as that for the known η case in (A.6).

Now we bound the conditional variance. We have

$$\mathsf{Var}\left[\kappa^*(w_i) - \kappa^*(\tau)\,\middle|\,\mathcal{G}\right] = \mathsf{Var}\left[\sum_{j:(i,j)\in\mathcal{E}} Y_{ij} \times \log\left\{\frac{(\hat{\eta}w_i + (1-\hat{\eta})w_j)(\hat{\eta}w_j + (1-\hat{\eta})\tau)}{(\hat{\eta}\tau + (1-\hat{\eta})w_j)(\hat{\eta}w_j + (1-\hat{\eta})w_i)}\right\}\right] \tag{C.9}$$

$$\lesssim |w_i-\tau|^2(2\hat{\eta}-1)^2 \sum_{j:(i,j)\in\mathcal{E}} \mathsf{Var}[Y_{ij}] \tag{C.10}$$

$$\leq |w_i-\tau|^2(2\hat{\eta}-1)^2 \sum_{j:(i,j)\in\mathcal{E}} \frac{1}{4L} \tag{C.11}$$

$$\lesssim |w_i-\tau|^2(2\hat{\eta}-1)^2\frac{np}{L}. \tag{C.12}$$

where

1) (C.10) follows from the original argument as in the proof of Lemma 7 in Appendix A-A;
2) (C.11) follows from the fact that the variance of any Bernoulli random variable is upper bounded by $1/4$;
3) and (C.12) holds with high probability due to the nature of the Erdős-Rényi graph.

Thus, by using the bounds in (C.8), (C.12) and Bernstein's inequality (Lemma 12), and mimicking the proof of Lemma 7 in Appendix A-A with $\hat{\eta}$ in place of η, we may conclude that

$$\kappa^*(w_i) - \kappa^*(\tau) \gtrsim np(2\hat{\eta}-1)^2|w_i-\tau|^2. \tag{C.13}$$

By Lemma 10 which allows us to multiplicatively approximate $(2\hat{\eta}-1)^2$ with $(2\eta-1)^2$ (to within a constant factor of $(1-\nu)^2$), we also have

$$\kappa^*(w_i) - \kappa^*(\tau) \gtrsim np(2\eta-1)^2|w_i-\tau|^2 \tag{C.14}$$

with probability tending to one polynomially fast.

Just as in the proof of Lemma 1, we do not have access to the true ground truth scores $\boldsymbol{w}_{\setminus i}$. We instead analyze the behavior of surrogate log-likelihoods $\hat{\kappa}$ with the true score vectors $\boldsymbol{w}_{\setminus i}$ replaced by their estimates $\hat{\boldsymbol{w}}_{\setminus i}$. We have

$$\hat{\kappa}(w_i) - \hat{\kappa}(\tau) = \sum_{j:(i,j)\in\mathcal{E}} \left\{ Y_{ij} \log\left\{\frac{(\hat{\eta}w_i + (1-\hat{\eta})\hat{w}_j)(\hat{\eta}\hat{w}_j + (1-\hat{\eta})\tau)}{(\hat{\eta}\tau + (1-\hat{\eta})\hat{w}_j)(\hat{\eta}\hat{w}_j + (1-\hat{\eta})w_i)}\right\} + \log\left\{\frac{(\tau+\hat{w}_j)(\hat{\eta}\hat{w}_j + (1-\hat{\eta})w_i)}{(w_i+\hat{w}_j)(\hat{\eta}\hat{w}_j + (1-\hat{\eta})\tau)}\right\}\right\}. \tag{C.15}$$

In a similar way to the case where η is known (cf. (A.10)), we can quantify the gap between the difference of surrogate log-likelihoods $\hat{\kappa}(w_i) - \hat{\kappa}(\tau)$ and difference of true log-likelihoods $\kappa^*(w_i) - \kappa^*(\tau)$ as follows:

$$\hat{\kappa}(w_i) - \hat{\kappa}(\tau) - \left(\kappa^*(w_i) - \kappa^*(\tau)\right) \lesssim \sum_{j:(i,j)\in\mathcal{E}} g_{\eta,\hat{\eta}}(\hat{w}_j), \tag{C.16}$$

where now

$$g_{\eta,\hat{\eta}}(t) := \frac{\eta w_i + (1-\eta)w_j}{w_i+w_j}\left\{\log\left(\frac{(\hat{\eta}w_i + (1-\hat{\eta})t)(\hat{\eta}t + (1-\hat{\eta})\tau)}{(\hat{\eta}\tau + (1-\hat{\eta})t)(\hat{\eta}t + (1-\hat{\eta})w_i)}\right) - \log\left(\frac{(\hat{\eta}w_i + (1-\hat{\eta})w_j)(\hat{\eta}w_j + (1-\hat{\eta})\tau)}{(\hat{\eta}\tau + (1-\hat{\eta})w_j)(\hat{\eta}w_j + (1-\hat{\eta})w_i)}\right)\right\} + \log\left(\frac{\tau+t}{w_i+t}\right) + \log\left(\frac{\hat{\eta}t + (1-\hat{\eta})w_i}{\hat{\eta}t + (1-\hat{\eta})\tau}\right) - \log\left(\frac{\tau+w_j}{w_i+w_j}\right) - \log\left(\frac{\hat{\eta}w_j + (1-\hat{\eta})w_i}{\hat{\eta}w_j + (1-\hat{\eta})\tau}\right). \tag{C.17}$$

Note that $g_{\eta,\eta}(t) = g_\eta(t)$ in (A.12) in the proof of Lemma 1. The reason why η appears in the leading factor in (C.17) is because we are taking expectation of Y_{ij} which is generated from the *true* model with parameter η (cf. (C.4)). The parameter $\hat{\eta}$ appears in $\{\ldots\}$ in (C.17) because the log-likelihood function $\kappa^*(\cdot)$ (cf. (C.2)) is defined with respect to the surrogate $\hat{\eta}$ since here we assume we have no knowledge of the true η.

Several properties of $g_\eta(t)$ were studied in the proof of Lemma 1. Here we need to study $g_{\eta,\hat{\eta}}(t)$. In fact, by using Lemma 9 to approximate $\eta w_i + (1-\eta)w_j$ with $\hat{\eta}w_i + (1-\hat{\eta})w_j$,

we see that with probability tending to one polynomially fast,

$$\begin{aligned} & g_{\eta,\hat{\eta}}(t) \\ & \lesssim \frac{\hat{\eta}w_i+(1-\hat{\eta})w_j}{w_i+w_j}\left\{\log\left\{\frac{(\hat{\eta}w_i+(1-\hat{\eta})t)(\hat{\eta}t+(1-\hat{\eta})\tau)}{(\hat{\eta}\tau+(1-\hat{\eta})t)(\hat{\eta}t+(1-\hat{\eta})w_i)}\right\}\right. \\ & \quad \left.-\log\left\{\frac{(\hat{\eta}w_i+(1-\hat{\eta})w_j)(\hat{\eta}w_j+(1-\hat{\eta})\tau)}{(\hat{\eta}\tau+(1-\hat{\eta})w_j)(\hat{\eta}w_j+(1-\hat{\eta})w_i)}\right\}\right\} \\ & \quad +\log\left(\frac{\tau+t}{w_i+t}\right)+\log\left(\frac{\hat{\eta}t+(1-\hat{\eta})w_i}{\hat{\eta}t+(1-\hat{\eta})\tau}\right) \\ & \quad -\log\left(\frac{\tau+w_j}{w_i+w_j}\right)-\log\left(\frac{\hat{\eta}w_j+(1-\hat{\eta})w_i}{\hat{\eta}w_j+(1-\hat{\eta})\tau}\right) \qquad \text{(C.18)} \\ & = g_{\hat{\eta}}(t) \qquad \text{(C.19)} \end{aligned}$$

where $g_{\hat{\eta}}(t)$ is $g(t)$ in (A.12) with η replaced by $\hat{\eta}$. Basically, we replaced the factor $\hat{\eta}w_i + (1-\hat{\eta})w_j$ with (a constant multiplied by) $\eta w_i + (1-\eta)w_j$ in (C.18). Now, the bound in (C.16) can be further upper bounded as

$$\hat{\kappa}(w_i) - \hat{\kappa}(\tau) - \big(\kappa^*(w_i) - \kappa^*(\tau)\big) \lesssim \sum_{j:(i,j)\in\mathcal{E}} g_{\hat{\eta}}(\hat{w}_j). \qquad \text{(C.20)}$$

The rest of the proof of Lemma 1, in particular the steps in (A.37)–(A.40), goes through verbatim with η replaced by $\hat{\eta}$. Finally, we can use Lemma 10 to multiplicatively approximate $(2\hat{\eta}-1)$ with $(2\eta-1)$ to complete the proof of Lemma 3. ■

A. Approximation Lemmas and Their Proofs

Lemma 9: For any pair of weights (w_i, w_j) and any constant $\nu > 0$, if

$$L \gtrsim \left(\frac{w_{\max}}{\nu w_{\min}}\right)^2 \log\frac{n}{\delta}, \qquad \text{(C.21)}$$

we have that

$$\left|\left(\frac{\eta w_i+(1-\eta)w_j}{\hat{\eta}w_i+(1-\hat{\eta})w_j}\right)-1\right| \le \nu \qquad \text{(C.22)}$$

with probability exceeding $1-\delta$.

The important point here is that this approximation is *uniform* over $(i,j) \in [n]^2$ and as $n \to \infty$; cf. the lower bound on L in (C.21) and the threshold ν in (C.22) do not depend on (i,j). This bound implies that, with high probability, we can readily approximate $\eta w_i + (1-\eta)w_j$ with $(1\pm\nu)(\hat{\eta}w_i + (1-\hat{\eta})w_j)$ for any constant $\nu > 0$. Also note that since $w_{\min}, w_{\max} = \Theta(1)$ and $\nu > 0$ is also a constant, the bound in (C.21) is in fact $L \gtrsim \log\frac{n}{\delta} \asymp \log n$ (with $\delta = 1/\operatorname{poly}(n)$). This is clearly satisfied by the assumption in (16) in Theorem 2.

Proof: [Proof of Lemma 9] Assume without loss of generality that $w_i > w_j$ (the expression in (C.22) is symmetric in w_i and w_j). Consider

$$\begin{aligned} & \Pr\left[\frac{\eta w_i+(1-\eta)w_j}{\hat{\eta}w_i+(1-\hat{\eta})w_j} > 1+\nu\right] \\ & = \Pr\left[(\eta-\hat{\eta})(w_i-w_j) > \nu\hat{\eta}w_i + \nu(1-\hat{\eta})w_j\right] \qquad \text{(C.23)} \\ & \le \Pr\left[(\eta-\hat{\eta})(w_i-w_j) > \nu w_{\min}\right] \qquad \text{(C.24)} \\ & = \Pr\left[\eta-\hat{\eta} > \nu\frac{w_{\min}}{w_i-w_j}\right] \qquad \text{(C.25)} \\ & \le \Pr\left[\eta-\hat{\eta} > \nu\frac{w_{\min}}{w_{\max}}\right] \qquad \text{(C.26)} \\ & \le \Pr\left[|\eta-\hat{\eta}| > \nu\frac{w_{\min}}{w_{\max}}\right] \qquad \text{(C.27)} \end{aligned}$$

where in (C.24), we lower bounded w_i, w_j by $w_{\min}$, (C.25) assumes that $w_i > w_j$ and (C.26) follows because $w_i - w_j \le w_i \le w_{\max}$. A bound for the other inequality $\Pr\left(\frac{\eta w_i+(1-\eta)w_j}{\hat{\eta}w_i+(1-\hat{\eta})w_j} < 1-\nu\right)$ proceeds in a completely analogous way. Since $w_{\min}, w_{\max} = \Theta(1)$, the result follows immediately from the union bound and the probabilistic bound on $|\hat{\eta}-\eta|$ (Lemma 5). ■

Lemma 10: For any constant $\nu > 0$, if

$$L \gtrsim \frac{1}{\nu^2(2\eta-1)^2}\log\frac{n}{\delta}, \qquad \text{(C.28)}$$

we have that

$$\left|\left(\frac{2\hat{\eta}-1}{2\eta-1}\right)-1\right| \le \nu \qquad \text{(C.29)}$$

with probability exceeding $1-\delta$.

Here, in contrast to Lemma 9, $(2\eta-1)$ in (C.29) may be vanishingly small, so the lower bound on L in (C.28) contains the additional term $(2\eta-1)^2$. *Proof:* [Proof of Lemma 10] Consider

$$\begin{aligned} \Pr\left[\left|\left(\frac{2\hat{\eta}-1}{2\eta-1}\right)-1\right| > \nu\right] &= \Pr\left[\left|\frac{\hat{\eta}-\eta}{2\eta-1}\right| > \frac{\nu}{2}\right] \qquad \text{(C.30)} \\ &= \Pr\left[|\hat{\eta}-\eta| > \frac{\nu}{2}(2\eta-1)\right]. \qquad \text{(C.31)} \end{aligned}$$

But we know from Lemma 5 that if

$$L \gtrsim \frac{1}{\left(\frac{\nu}{2}(2\eta-1)\right)^2}\log\frac{n}{\delta} \asymp \frac{1}{\nu^2(2\eta-1)^2}\log\frac{n}{\delta}, \qquad \text{(C.32)}$$

then the probability in (C.31) is no larger than δ. ■

APPENDIX D
PROOF OF LEMMA 4

From the proof sketch in Section VI-D, we see that it suffices to prove the upper bound on $\|\hat{\Delta}\|$ in (71). The entries of $\hat{\Delta}$ are denoted in the usual way as $\hat{\Delta}_{ij}$ where $i, j \in [n]$. When η was known, it was imperative to understand the probability that

$$F_{ij} := Ld_{\max}\Delta_{ij} = \frac{\left(\sum_{\ell=1}^{L} Y_{ij}^{(\ell)}\right) - L(1-\eta)}{2\eta-1} - L\frac{w_i}{w_i+w_j} \qquad \text{(D.1)}$$

deviates from zero. See the corresponding bound in (B.14). When one only has an estimate of η, namely $\hat{\eta}$, it is then imperative to do the same for

$$\hat{F}_{ij} := \frac{\left(\sum_{\ell=1}^{L} Y_{ij}^{(\ell)}\right) - L(1-\hat{\eta})}{2\hat{\eta}-1} - L\frac{w_i}{w_i+w_j}. \qquad \text{(D.2)}$$

Our overarching strategy is to bound $\hat{F}_{ij}$ in terms of F_{ij} and then use the concentration bound we had established for F_{ij} in (B.14) to then understand the stochastic behavior of $\hat{F}_{ij}$.

To simplify notation, define the sum $U := LY_{ij} = \sum_{\ell=1}^{L} Y_{ij}^{(\ell)}$. Consequently,

$$\begin{aligned}
&\left|\hat{F}_{ij} - F_{ij}\right| \\
&\quad = \left|\frac{U - L(1-\hat{\eta})}{2\hat{\eta}-1} - \frac{U - L(1-\eta)}{2\eta - 1}\right| && \text{(D.3)}\\
&\quad \le L\left|\frac{1-\hat{\eta}}{2\hat{\eta}-1} - \frac{1-\eta}{2\eta-1}\right| + U\left|\frac{1}{2\hat{\eta}-1} - \frac{1}{2\eta-1}\right| && \text{(D.4)}\\
&\quad \le L\left[\left|\frac{1-\hat{\eta}}{2\hat{\eta}-1} - \frac{1-\eta}{2\eta-1}\right| + \left|\frac{1}{2\hat{\eta}-1} - \frac{1}{2\eta-1}\right|\right] && \text{(D.5)}
\end{aligned}$$

where the final bound follows from the fact that $|U| \le L$ almost surely (since $Y_{ij}^{(\ell)} \in \{0,1\}$). Now we make use of the following lemma that uses the sample complexity result in Lemma 5 to quantify the Lipschitz constant of the maps $t \mapsto \frac{1}{2t-1}$ and $t \mapsto \frac{1-t}{2t-1}$ in the vicinity of $t = (1/2)^{+}$.

Lemma 11: Let $\lambda_1 : (1/2, 1] \to \mathbb{R}_+$ and $\lambda_2 : (1/2, 1] \to \mathbb{R}_+$ be defined as

$$\lambda_1(t) := \frac{1-t}{2t-1}, \quad \textit{and} \quad \lambda_2(t) := \frac{1}{2t-1}. \tag{D.6}$$

Then if

$$L \gtrsim \frac{1}{(2\eta-1)^2}\log\frac{n}{\delta} \tag{D.7}$$

with probability exceeding $1-\delta$ (over the random variable $\hat{\eta}$ which depends on the samples drawn from the mixture distribution (5)*), we have for each $j = 1, 2$,*

$$|\lambda_j(\hat{\eta}) - \lambda_j(\eta)| \le \frac{8}{(2\eta-1)^2}|\hat{\eta} - \eta|. \tag{D.8}$$

The proof of this lemma is deferred to Appendix D-A at the end of this appendix. We take $\delta = 1/\operatorname{poly}(n)$ in the sequel so (D.7) is equivalently

$$L \gtrsim \frac{\log n}{(2\eta-1)^2} \tag{D.9}$$

which when combined with $S = \binom{n}{2} pL$ is less stringent than the statement of Theorem 2. Thus, under the condition (D.9), Lemma 11 yields that

$$\left|\hat{F}_{ij} - F_{ij}\right| \le \frac{16L}{(2\eta-1)^2}|\hat{\eta} - \eta| \tag{D.10}$$

with probability exceeding $1 - 1/\operatorname{poly}(n)$. By the reverse triangle inequality, we obtain

$$\left|\hat{F}_{ij} - F_{ij}\right| \ge \left||\hat{F}_{ij}| - |F_{ij}|\right|. \tag{D.11}$$

To make the dependence of $|\hat{\eta} - \eta|$ on the number of samples L explicit, we define

$$\varepsilon_L := |\hat{\eta} - \eta|. \tag{D.12}$$

By uniting (D.10)–(D.12), we obtain

$$|F_{ij}| - \varepsilon_L' \le \left|\hat{F}_{ij}\right| \le |F_{ij}| + \varepsilon_L' \tag{D.13}$$

where

$$\varepsilon_L' := \frac{16L}{(2\eta-1)^2}\varepsilon_L. \tag{D.14}$$

For later reference, define

$$\varepsilon_L'' := \frac{16L}{(2\eta-1)^2} d_{\max}\varepsilon_L. \tag{D.15}$$

With the estimate in (D.13), we observe that for any $t > 0$, one has

$$\Pr\left[\left|\hat{F}_{ij}\right| \ge t\right] \le \Pr\left[|F_{ij}| + \varepsilon_L' \ge t\right] = \Pr\left[|F_{ij}| \ge t - \varepsilon_L'\right] \tag{D.16}$$

where the randomness in the probability on the left is over both $\hat{\eta}$ and $\boldsymbol{Y} := \{Y_{ij}^{(\ell)} : \ell \in [L], (i,j) \in \mathcal{E}\}$ (the former is a function of the latter) whereas the randomness in the probability on the right is only over $\boldsymbol{Y}$. Thus, by using the equality $F_{ij} = Ld_{\max}\Delta_{ij}$ and applying Hoeffding's inequality to (D.16) (cf. the bound in (B.14)), we obtain

$$\Pr\left[\left|Ld_{\max}\hat{\Delta}_{ij}\right| \ge t\right] \le 2\exp\left(-\frac{2((t-\varepsilon_L')\frac{2\eta-1}{2\eta+1})^2}{L}\right). \tag{D.17}$$

Now by the same argument as in (B.4), $Ld_{\max}\hat{\Delta}_{ii} = -\sum_{k\neq i} Ld_{\max}\hat{\Delta}_{ik} = -\sum_{k\neq i}\hat{F}_{ik}$ so we have

$$|Ld_{\max}\hat{\Delta}_{ii}| - \varepsilon_L'' \le |Ld_{\max}\hat{\Delta}_{ii}| \le |Ld_{\max}\hat{\Delta}_{ii}| + \varepsilon_L''. \tag{D.18}$$

As a result, similarly to the calculation that led to (D.17), we obtain

$$\Pr\left[\left|Ld_{\max}\hat{\Delta}_{ii}\right| \ge t\right] \le 2\exp\left(-\frac{2((t-\varepsilon_L'')\frac{2\eta-1}{2\eta+1})^2}{Ld_{\max}}\right). \tag{D.19}$$

From the Hoeffding bound analysis leading to the non-asymptotic bound in (D.19), we know that by choosing

$$t := c\sqrt{Ld_{\max}\log n}\left(\frac{2\eta+1}{2\eta-1}\right) + \varepsilon_L'', \tag{D.20}$$

for some sufficiently large constant $c > 0$,

$$\Pr\left[\left|Ld_{\max}\hat{\Delta}_{ii}\right| \ge t\right] = O\left(\frac{1}{\operatorname{poly}(n)}\right). \tag{D.21}$$

In other words,

$$|\hat{\Delta}_{ii}| \lesssim \frac{1}{2\eta-1}\sqrt{\frac{\log n}{Ld_{\max}}} + \frac{\varepsilon_L''}{Ld_{\max}} \tag{D.22}$$

with probability at least $1 - 1/\operatorname{poly}(n)$. Recall the definition of ε_L'' in (D.15). We now design $(\varepsilon_L, \varepsilon_L'')$ such that

$$\frac{\varepsilon_L''}{Ld_{\max}} = \frac{16}{(2\eta-1)^2}\varepsilon_L = \frac{1}{2\eta-1}\sqrt[4]{\frac{\log^2 n}{Ld_{\max}}}. \tag{D.23}$$

Now note $d_{\max} = \Theta(\log n)$ with high probability. This implies that the second term in (D.22) dominates the first term. Thus,

$$|\hat{\Delta}_{ii}| \lesssim \frac{1}{2\eta-1}\sqrt[4]{\frac{\log^2 n}{Ld_{\max}}}, \tag{D.24}$$

with probability at least $1 - 1/\operatorname{poly}(n)$. A similar high probability bound, of course, holds for $|\hat{\Delta}_{ij}|$ if we choose t in (D.17) similarly to the choice made in (D.20). We may rearrange (D.23) to yield

$$\varepsilon_L \asymp (2\eta-1)\sqrt[4]{\frac{\log^2 n}{Ld_{\max}}}. \tag{D.25}$$

Given the bound on the diagonal elements $\hat{\Delta}_{ii}$ in (D.24) and a similar bound on the off-diagonal elements $\hat{\Delta}_{ij}$, similarly to

the proof of Lemma 2 in Appendix B, the spectral norm of $\hat{\Delta}$ can be bounded as

$$\|\hat{\Delta}\| \lesssim \frac{1}{2\eta-1}\sqrt[4]{\frac{\log^2 n}{Ld_{\max}}}. \tag{D.26}$$

Now we check that the lower bound on L is satisfied when we choose ε_L according to (D.25). Using the sample complexity bound in (70) and rearranging, we obtain

$$L \gtrsim \frac{\log n}{(2\eta-1)^4} \tag{D.27}$$

which when combined with $S = \binom{n}{2}pL$ is less stringent than the statement of Theorem 2. This completes the proof of the upper bound of $\|\hat{\Delta}\|$ in (71).

As a final remark, let us mention that we could have improved the sample complexity S_{Δ_K} from being dependent on the reciprocal of Δ_K^4 in (16) to being dependent on the reciprocal of Δ_K^2 if the value of Δ_K is known *a priori*. If it were known, we could have incorporated it into the choice of ε_L'' in (D.23).

A. Proof of Lemma 11

Consider the functions $\lambda_1 : (1/2, 1] \to \mathbb{R}$ and $\lambda_2 : (1/2, 1] \to \mathbb{R}$ given by (D.6). By direct differentiation, we have

$$\lambda_1'(t) = \frac{-1}{(2t-1)^2}, \quad \text{and} \quad \lambda_2'(t) = \frac{-2}{(2t-1)^2}. \tag{D.28}$$

We note that an everywhere differentiable function g is Lipschitz continuous with Lipschitz constant $\sup|g'|$. We now assume that $\eta, \hat{\eta} \in [\eta^*, 1]$ for some $\eta^* > 1/2$. By using the fact that $2/(2\eta^*-1)^2$ is an upper bound of the derivative of $\lambda_j|_{[\eta^*,1]}$ (i.e., λ_j restricted to the domain $[\eta^*, 1]$), one has

$$|\lambda_j(\hat{\eta}) - \lambda_j(\eta)| \le \frac{2}{(2\eta^*-1)^2}|\hat{\eta}-\eta| \tag{D.29}$$

for $j = 1, 2$. We now put

$$\eta^* := \frac{1}{2}\Big(\eta + \frac{1}{2}\Big). \tag{D.30}$$

This quantity is the average of $1/2$ and η and so is greater than $1/2$ as required. Also, $\eta - \eta^* = \eta/2 - 1/4$. Now, (D.29) becomes

$$|\lambda_j(\hat{\eta}) - \lambda_j(\eta)| \le \frac{2}{(\eta-1/2)^2}|\hat{\eta}-\eta| = \frac{8}{(2\eta-1)^2}|\hat{\eta}-\eta| \tag{D.31}$$

for $j = 1, 2$ if $\hat{\eta} \in [\eta^*, 2\eta-\eta^*] \subset [\eta^*, 1]$. The probability that this happens (recalling that $\hat{\eta}$ is the random variable in question) is

$$\Pr\left[\eta^* \le \hat{\eta} \le 2\eta - \eta^*\right] = 1 - \Pr\left[|\hat{\eta}-\eta| > \frac{\eta}{2} - \frac{1}{4}\right]. \tag{D.32}$$

From Lemma 5, we know that if

$$L \gtrsim \frac{1}{\varepsilon^2}\log\frac{n}{\delta}, \tag{D.33}$$

then we have $|\hat{\eta}-\eta| \le \varepsilon$ with probability at least $1-\delta$. Hence, if

$$L \gtrsim \frac{1}{(\frac{\eta}{2}-\frac{1}{4})^2}\log\frac{n}{\delta} \asymp \frac{1}{(2\eta-1)^2}\log\frac{n}{\delta} \tag{D.34}$$

then (D.31) holds with probability at least $1-\delta$. This completes the proof of Lemma 11.

Appendix E
Proof of Lemma 6

A. The Scaling of Singular Values $\sigma_i(M_2)$

Since M_2 is symmetric and positive semidefinite, its eigenvalues (which are all non-negative) are the same as its singular values. Since the eigenvectors are invariant to scaling, let us assume that

$$v = \pi_0 + b\pi_1 \tag{E.1}$$

is an eigenvector. Then by uniting the definition of M_2 in (56) and (E.1), we have

$$\begin{aligned} M_2 v = {} & (\eta\|\pi_0\|^2 + \eta b\langle\pi_0,\pi_1\rangle)\pi_0 \\ & + ((1-\eta)a\langle\pi_0,\pi_1\rangle + b(1-\eta)\|\pi_1\|^2)\pi_1. \end{aligned} \tag{E.2}$$

Since v is assumed to be an eigenvector, $M_2 v$ satisfies that

$$M_2 v = \sigma v \tag{E.3}$$

where σ is some eigenvalue or singular value. Since π_0 is linearly independent of π_1, this equates to

$$\eta\|\pi_0\|^2 + \eta b\langle\pi_0,\pi_1\rangle = \sigma \tag{E.4}$$

$$(1-\eta)a\langle\pi_0,\pi_1\rangle + b(1-\eta)\|\pi_1\|^2 = \sigma b. \tag{E.5}$$

Now note from the definitions of π_0 and π_1 that

$$\|\pi_0\|^2 = \|\pi_1\|^2 \tag{E.6}$$

because the elements are the same and π_1 is simply a permuted version of π_0. So we will replace $\|\pi_1\|^2$ with $\|\pi_0\|^2$ henceforth. Eliminating σ from the simultaneous equations in (E.4) and (E.5), we obtain the quadratic equation in the unknown b:

$$\eta\langle\pi_0,\pi_1\rangle b^2 + (2\eta-1)\|\pi_0\|^2 b - (1-\eta)\langle\pi_0,\pi_1\rangle = 0 \tag{E.7}$$

which implies that

$$b^* = \frac{-(2\eta-1)\|\pi_0\|^2 \pm \sqrt{(2\eta-1)^2\|\pi_0\|^4 + 4\eta(1-\eta)\langle\pi_0,\pi_1\rangle^2}}{2\eta\langle\pi_0,\pi_1\rangle}. \tag{E.8}$$

Now, we observe that

$$\langle\pi_0,\pi_1\rangle = \sum_{(i,j)\in\mathcal{E}} 2\frac{w_i w_j}{w_i + w_j} \tag{E.9}$$

$$\|\pi_0\|^2 = \sum_{(i,j)\in\mathcal{E}} \frac{w_i^2 + w_j^2}{(w_i+w_j)^2}. \tag{E.10}$$

so by the fact that $w_{\min}$ and $w_{\max}$ are bounded, we see that $\langle\pi_0,\pi_1\rangle = \Theta(|\mathcal{E}|)$ and $\|\pi_0\|^2 = \Theta(|\mathcal{E}|)$. Plugging these estimates into b^*, we see that $b^* = \Theta(1)$. Thus, by (E.4),

we see that with high probability over the realization of the Erdős-Rényi graph,

$$\sigma = \Theta(\eta|\mathcal{E}|) = \Theta\left(\eta n^2 p\right). \tag{E.11}$$

This scaling holds for both singular values $\sigma_1(M_2)$ and $\sigma_2(M_2)$ so this proves (76). Two distinct values for the singular values due to the $\pm$ sign in b^* in (E.8). This completes the proof of (76).

B. The Scaling of Block-Incoherence Parameter $\mu(M_2)$

Now let us evaluate the scaling of $\mu(M_2)$. From (E.1) and (E.8), we know the form of the eigenvectors of M_2. The singular vectors must be normalized so they can be written as

$$\hat{v} := \frac{v}{\|v\|_2}. \tag{E.12}$$

Since the length of v is $2|\mathcal{E}|$, and the values (elements) of v are uniformly upper and lower bounded, it is easy to see that $\|v\|_2 = \Theta(\sqrt{|\mathcal{E}|})$. As a result, one has

$$\hat{v} = \Theta\left(\frac{1}{\sqrt{|\mathcal{E}|}}\right)v. \tag{E.13}$$

Thus, each subblock of U has entries that scale as $O(|\mathcal{E}|^{-1/2})$ and so

$$\left\|U^{(k)}\right\|_2 = \Theta\left(\frac{1}{\sqrt{|\mathcal{E}|}}\right). \tag{E.14}$$

As a result, from the definition of $\mu(M_2)$ in (72), we see that $\mu(M_2)$ is of constant order, i.e.,

$$\mu(M_2) = \Theta(1), \tag{E.15}$$

which completes the proof of (77).

Appendix F
Bernstein Inequality

Lemma 12: Consider n independent random variables X_i with $|X_i| \le B$. For any $\gamma \ge 2$, one has

$$\left|\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right| \le \sqrt{2\gamma \log n \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right]} + \frac{2\gamma}{3} B \log n \tag{F.1}$$

with probability at least $1 - 2n^{-\gamma}$.

Appendix G
STGD Method to Solve (60)

In this appendix, we describe how to solve (60) in Algorithm 3 using the STGD method (cf. Huang *et al.* [30]). We also discuss some difficulties in directly using [30] on our problem in which $|\mathcal{E}|$ is assumed to be large (for Theorem 2 to be valid).

First, define a linear operator $\tau : \mathbb{R}^{2|\mathcal{E}|\times 2|\mathcal{E}|\times 2|\mathcal{E}|} \to \mathbb{R}^{2\times 2\times 2}$ given by the recipe

$$\tau(W) := \mathcal{P}_{\Omega_3}(W)\left[Q_{\widehat{M}_2}\right]_3. \tag{G.1}$$

Then (60) becomes

$$\underset{Z\in\mathbb{R}^{2\times 2\times 2}}{\operatorname{argmin}} \left\{ f(Z) := \left\| \tau\left(Z\left[P_{\widehat{M}_2}\right]_3 - \sum_{t\in\mathcal{I}_2} \frac{\otimes^3 \underline{Y}^{(t)}}{|\mathcal{I}_2|}\right)\right\|_F^2 \right\}. \tag{G.2}$$

Next, we rewrite f as follows

$$\begin{aligned} f(Z) &\overset{c}{=} \left\|\tau\left(Z\left[P_{\widehat{M}_2}\right]_3\right)\right\|_F^2 \\ &\quad -2\left\langle \tau\left(Z\left[P_{\widehat{M}_2}\right]_3\right), \tau\left(\frac{1}{|\mathcal{I}_2|}\sum_{t\in\mathcal{I}_2}\otimes^3\underline{Y}^{(t)}\right)\right\rangle \qquad \text{(G.3)} \\ &= \left\|\tau\left(Z\left[P_{\widehat{M}_2}\right]_3\right)\right\|_F^2 \\ &\quad -2\left\langle \tau(\boldsymbol{Z}[\boldsymbol{X},\boldsymbol{X},\boldsymbol{X}]), \frac{1}{|\mathcal{I}_2|}\sum_{t\in\mathcal{I}}\tau(\boldsymbol{x}_t\otimes\boldsymbol{x}_t\otimes\boldsymbol{x}_t)\right\rangle \qquad \text{(G.4)} \\ &= \frac{1}{|\mathcal{I}_2|}\sum_{t\in\mathcal{I}_2}\left\|\tau\left(Z\left[P_{\widehat{M}_2}\right]_3\right)\right\|_F^2 \\ &\quad -2\left\langle \tau\left(Z\left[P_{\widehat{M}_2}\right]_3\right), \tau\left(\otimes^3\underline{Y}^{(t)}\right)\right\rangle, \qquad \text{(G.5)} \end{aligned}$$

where '$\overset{c}{=}$' omits constants that are independent of Z, i.e., $f(Z) \overset{c}{=} g(Z)$ if and only if $f(Z) = g(Z) + c$ where c does not depend on Z. Now, define

$$f_t(Z) := \left\|\tau\left(Z\left[P_{\widehat{M}_2}\right]_3\right)\right\|_F^2 - 2\left\langle \tau\left(Z\left[P_{\widehat{M}_2}\right]_3\right), \tau\left(\otimes^3\underline{Y}^{(t)}\right)\right\rangle, \tag{G.6}$$

then $f(Z) = \frac{1}{|\mathcal{I}_2|}\sum_{t\in\mathcal{I}_2} f_t(Z)$. Therefore we can use the STGD method [30] to minimize f over tensors $Z \in \mathbb{R}^{2\times 2\times 2}$. It remains to find the gradient of f_t. For any (small perturbation matrix) $\Delta Z \in \mathbb{R}^{2\times 2\times 2}$, we have

$$\begin{aligned} f_t(Z+\Delta Z) &= \left\|\tau\left(\left[P_{\widehat{M}_2}\right]_3 + \Delta Z\left[P_{\widehat{M}_2}\right]_3\right)\right\|_F^2 \\ &\quad -2\left\langle \tau(Z\left[P_{\widehat{M}_2}\right]_3 + \Delta Z\left[P_{\widehat{M}_2}\right]_3), \tau\left(\otimes^3\underline{Y}^{(t)}\right)\right\rangle \qquad \text{(G.7)} \\ &= f_t(Z) + 2\left\langle \tau\left(\Delta Z\left[P_{\widehat{M}_2}\right]_3\right), \tau\left(Z\left[P_{\widehat{M}_2}\right]_3 - \otimes^3\underline{Y}^{(t)}\right)\right\rangle \\ &\quad + o\left(\|\Delta Z\|_F\right). \qquad \text{(G.8)} \end{aligned}$$

By definition, the Fréchet derivative [40] of f_t at Z, $\mathcal{D}f_t(Z) : \mathbb{R}^{2\times 2\times 2} \to \mathbb{R}$ is defined as

$$\begin{aligned} &[\mathcal{D}f_t(Z)](\Delta Z) \\ &\quad := 2\left\langle \tau\left(\Delta Z\left[P_{\widehat{M}_2}\right]_3\right), \tau\left(Z\left[P_{\widehat{M}_2}\right]_3 - \otimes^3\underline{Y}^{(t)}\right)\right\rangle. \qquad \text{(G.9)} \end{aligned}$$

To find an explicit form of the gradient of f_t at Z, denoted as $\nabla f_t(Z)$, we write $[\mathcal{D}f_t(Z)](\Delta Z)$ in the form $\langle \nabla f_t(Z), \Delta Z\rangle$. Define $B_t(Z) := \tau\left(Z\left[P_{\widehat{M}_2}\right]_3 - \otimes^3\underline{Y}^{(t)}\right)$. Then, for any $\Delta Z \in \mathbb{R}^{2\times 2\times 2}$, we have (G.10)–(G.16) at the top of the next page. Therefore, we find that the gradient of f_t is

$$\nabla f_t(Z) = 2\mathcal{P}_{\Omega_3}\left(B_t(Z)\left[Q_{\widehat{M}_2}^T\right]_3\right)\left[P_{\widehat{M}_2}^T\right]_3. \tag{G.17}$$

However, from (G.13), we observe that the complexity for computing the gradient $\nabla f_t(Z)$ is $\Theta(|\mathcal{E}|^3)$. Since Theorem 2

$$[\mathcal{D} f_t(Z)](\Delta Z) = 2\left\langle \tau\left(\Delta Z\left[P_{\widehat{M}_2}\right]_3\right), B_t(Z)\right\rangle \tag{G.10}$$

$$= 2\sum_{j_1,j_2,j_3\in[2]}\sum_{i_1,i_2,i_3\in\Omega_3}\left(\Delta Z\left[P_{\widehat{M}_2}\right]_3\right)_{i_1,i_2,i_3}\prod_{k=1}^{3}\left(Q_{\widehat{M}_2}\right)_{i_k,j_k}\left(B_t(Z)\right)_{j_1,j_2,j_3} \tag{G.11}$$

$$= 2\sum_{j_1,j_2,j_3\in[2]}\sum_{i_1,i_2,i_3\in\Omega_3}\sum_{a_1,a_2,a_3\in[2]}(\Delta Z)_{a_1,a_2,a_3}\prod_{k=1}^{3}\left(P_{\widehat{M}_2}\right)_{a_k,i_k}\prod_{k=1}^{3}\left(Q_{\widehat{M}_2}\right)_{i_k,j_k}\left(B_t(Z)\right)_{j_1,j_2,j_3} \tag{G.12}$$

$$= 2\sum_{a_1,a_2,a_3\in[2]}(\Delta Z)_{a_1,a_2,a_3}\sum_{i_1,i_2,i_3\in\Omega_3}\left(\sum_{j_1,j_2,j_3\in[2]}\left(B_t(Z)\right)_{j_1,j_2,j_3}\prod_{k=1}^{3}\left(Q^T_{\widehat{M}_2}\right)_{j_k,i_k}\right)\prod_{k=1}^{3}\left(P^T_{\widehat{M}_2}\right)_{i_k,a_k} \tag{G.13}$$

$$= 2\sum_{a_1,a_2,a_3\in[2]}(\Delta Z)_{a_1,a_2,a_3}\sum_{i_1,i_2,i_3\in\Omega_3}\left(B_t(Z)\left[Q^T_{\widehat{M}_2}\right]_3\right)_{i_1,i_2,i_3}\prod_{k=1}^{3}\left(P^T_{\widehat{M}_2}\right)_{i_k,a_k} \tag{G.14}$$

$$= 2\sum_{a_1,a_2,a_3\in[2]}(\Delta Z)_{a_1,a_2,a_3}\left(\mathcal{P}_{\Omega_3}\left(B_t(Z)\left[Q^T_{\widehat{M}_2}\right]_3\right)\left[P^T_{\widehat{M}_2}\right]_3\right)_{a_1,a_2,a_3} \tag{G.15}$$

$$= 2\left\langle \mathcal{P}_{\Omega_3}\left(B_t(Z)\left[Q^T_{\widehat{M}_2}\right]_3\right)\left[P^T_{\widehat{M}_2}\right]_3, \Delta Z\right\rangle. \tag{G.16}$$

requires $|\mathcal{E}|$ (which is close to $\binom{n}{2}p = \Theta(n\log n)$ w.h.p.) to be large, the complexity for computing $\nabla f_t(Z)$ dramatically increases as $|\mathcal{E}|$ grows. Moreover, note that f_t in (G.6) is *not separable* across the triple of indices $(i_1, i_2, i_3) \in \Omega_3$ so a further stochastic gradient descent-like algorithm to minimize f_t may not be easy to derive. The same problem also arises in the computation of $B_t(Z)$. That said, the STGD algorithm described herein can at least deal with dataset with a large L, thereby providing only a partial remedy to the scalability issue pertaining to Algorithm 3.

ACKNOWLEDGMENTS

The authors would like to sincerely thank the Associate Editor Prof. Constantine Caramanis and the anonymous reviewers for their extensive and useful comments during the revision process.

REFERENCES

[1] A. Caplin and B. Nalebuff, "Aggregation and social choice: A mean voter theorem," *Econometrica*, vol. 59, no. 1, pp. 1–23, 1991.
[2] H. A. Soufiani, D. Parkes, and L. Xia, "Computing parametric ranking models via rank-breaking," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 360–368.
[3] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the Web," in *Proc. 10th Int. World Wide Web Conf.*, May 2001, pp. 613–622.
[4] L. Baltrunas, T. Makcinskas, and F. Ricci, "Group recommendations with rank aggregation and collaborative filtering," in *Proc. ACM Conf. Recommender Syst.*, 2010, pp. 119–126.
[5] T.-K. Huang, C.-J. Lin, and R. C. Weng, "Ranking individuals by group comparisons," *J. Mach. Learn. Res.*, vol. 9, no. 10, pp. 2187–2216, 2008.
[6] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz, "Pairwise ranking aggregation in a crowdsourced setting," in *Proc. 6th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2013, pp. 193–202.
[7] S. Negahban, S. Oh, and D. Shah. (Nov. 2015). "Rank centrality: Ranking from pair-wise comparisons." [Online]. Available: https://arxiv.org/abs/1209.1688
[8] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, nos. 1–7, pp. 107–117, Apr. 1998.
[9] L. R. Ford, "Solution of a ranking problem from binary comparisons," *Amer. Math. Monthly*, vol. 64, no. 8, pp. 28–33, 1957.
[10] Y. Chen and C. Suh, "Spectral MLE: Top-K rank aggregation from pairwise comparisons," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 371–380.
[11] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, vol. 41, nos. 3–4, pp. 324–345, 1952.
[12] R. D. Luce, *Individual Choice Behavior: A Theoretical Analysis*. Hoboken, NJ, USA: Wiley, 1959.
[13] C. Castillo and B. D. Davison, "Adversarial Web search," *Found. Trends Inf. Retr.*, vol. 4, no. 5, pp. 377–486, 2010.
[14] A. Broder, "A taxonomy of Web search," *ACM SIGIR Forum*, vol. 36, no. 2, pp. 3–10, 2002.
[15] P. Jain and S. Oh, "Learning mixtures of discrete product distributions using spectral decompositions," in *Proc. Conf. Learn. Theory (COLT)*, 2014, pp. 824–856.
[16] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2773–2832, 2014.
[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.
[18] J. Yi, R. Jin, S. Jain, and A. K. Jain, "Inferring users' preferences from crowdsourced pairwise comparisons: A matrix completion approach," in *Proc. 1st AAAI Conf. Human Comput. Crowdsourcing (HCOMP)*, 2013, pp. 207–215.
[19] P. Ye and D. Doermann, "Combining preference and absolute judgements in a crowd-sourced setting," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2013, pp. 1–7.
[20] Y. Kim, W. Kim, and K. Shim, "Latent ranking analysis using pairwise comparisons," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Dec. 2014, pp. 869–874.
[21] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. E. Schapire, and L. Sellie, "On the learnability of discrete distributions," in *Proc. ACM Symp. Theory Comput. (STOC)*, May 1994, pp. 273–282.
[22] Y. Freund and Y. Mansour, "Estimating a mixture of two product distributions," in *Proc. Conf. Learn. Theory (COLT)*, 1999, pp. 53–62.
[23] J. Feldman, R. O'Donnell, and R. A. Servedio, "Learning mixtures of product distributions over discrete domains," *SIAM J. Comput.*, vol. 37, no. 5, pp. 1536–1564, 2008.
[24] N. Shah and M. Wainwright. (Dec. 2015). "Simple, robust and optimal ranking from pairwise comparisons." [Online]. Available: https://arxiv.org/abs/1512.08949
[25] J.-C. d. Borda, "Mémoire sur les élections au scrutin," in *Histoire de l'Académie Royale des Sciences*. Paris, France: Histoire de l'Academie Royale des Sci., 1781.

[26] A. Ammar and D. Shah, "Ranking: Compare, don't score," in *Proc. Allerton Conf.*, Sep. 2011, pp. 776–783.
[27] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Found. Comput. Math.*, vol. 12, no. 4, pp. 389–434, Aug. 2012.
[28] I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1007–1016, Mar. 2006.
[29] V. Y. F. Tan and G. Atia, "Strong impossibility results for sparse signal processing," *IEEE Signal Process. Lett.*, vol. 21, no. 3, pp. 260–264, Mar. 2014.
[30] F. Huang, U. N. Niranjan, M. U. Hakeem, and A. Anandkumar, "Online tensor methods for learning latent variable models," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 2797–2835, 2015.
[31] R. C. Weng and C.-J. Lin, "A Bayesian approximation method for online ranking," *J. Mach. Learn. Res.*, vol. 12, pp. 267–300, Feb. 2011.
[32] A. Cichocki. "Tensor networks for big data analytic and large-scale optimization problems," in *Proc. 2nd Int. Conf. Eng. Comput. Schematics (ECM)*, 2013. [Online]. Available: https://arxiv.org/abs/1407.3124
[33] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—Online stochastic gradient for tensor decomposition," in *Proc. Conf. Learn. Theory (COLT)*, 2015, pp. 797–842.
[34] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inf. Theory*, vol. 36, no. 3, pp. 453–471, May 1990.
[35] A. Beirami and F. Fekri, "Results on the redundancy of universal compression for finite-length sequences," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aug. 2011, pp. 1504–1508.
[36] R. L. Plackett, "The analysis of permutations," *J. Roy. Statist. Soc. C (Appl. Statist.)*, vol. 24, no. 2, pp. 193–202, 1975.
[37] B. Hajek, S. Oh, and J. Xu, "Minimax-optimal inference from partial rankings," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 1475–1483.
[38] L. Maystre and M. Grossglauser, "Fast and accurate inference of Plackett-Luce models," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 172–180.
[39] R. W. Yeung, *Information Theory and Network Coding*. New York, NY, USA: Springer, 2008.
[40] B. S. Mordukhovich, "Variational analysis and generalized differentiation I," in *Basic Theory*. Berlin, Germany: Springer, 2006.

Changho Suh (S'10–M'12) is an Ewon Associate Professor in the School of Electrical Engineering at Korea Advanced Institute of Science and Technology (KAIST). He recevied the B.S. and M.S. degrees in Electrical Engineering from KAIST in 2000 and 2002 respectively, and the Ph.D. degree in Electrical Engineering and Computer Sciences from UC-Berkeley in 2011. From 2011 to 2012, he was a postdoctoral associate at the Research Laboratory of Electronics in MIT. From 2002 to 2006, he had been with the Telecommunication R&D Center, Samsung Electronics.

Dr. Suh received the 2015 IEIE Hadong Young Engineer Award, a 2015 Bell Labs Prize finalist, the 2013 IEEE Communications Society Stephen O. Rice Prize, the 2011 David J. Sakrison Memorial Prize (top research award in the UC-Berkeley EECS Department), and the 2009 IEEE ISIT Best Student Paper Award.

Vincent Y. F. Tan (S'07–M'11–SM'15) is currently an Assistant Professor in the Department of Electrical and Computer Engineering (ECE) and the Department of Mathematics at the National University of Singapore (NUS). He received the B.A. and M.Eng. degrees in Electrical and Information Sciences from Cambridge University in 2005 and the Ph.D. degree in Electrical Engineering and Computer Science (EECS) from the Massachusetts Institute of Technology in 2011. He was a postdoctoral researcher in the Department of ECE at the University of Wisconsin-Madison and a research scientist at the Institute for Infocomm (I2R) Research, A*STAR, Singapore. His research interests include information theory and machine learning.

Dr. Tan received the MIT EECS Jin-Au Kong outstanding doctoral thesis prize in 2011 and the NUS Young Investigator Award in 2014. He has authored a monograph on *Asymptotic Estimates in Information Theory with Non-Vanishing Error Probabilities* in the Foundations and Trends in Communications and Information Theory. He is currently an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS and the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING.

Renbo Zhao (S'14–M'16) received the B.Eng. degree in electrical engineering from the National University of Singapore (NUS), Singapore, in 2015. He is currently working toward the Master's degree in the Department of Mathematics, NUS. He is also currently a Research Engineer in the Department of Electrical and Computer Engineering and the Department of Industrial and Systems Engineering, NUS. His research interests include large-scale optimization methods, with a focus on their applications in signal processing, machine learning, control theory, and operations research.