Accelerated Spectral Ranking

Arpit Agarwal [1] **Prathamesh Patil** [1] **Shivani Agarwal** [1]

Abstract

The problem of rank aggregation from pairwise and multiway comparisons has a wide range of implications, ranging from recommendation systems to sports rankings to social choice. Some of the most popular algorithms for this problem come from the class of spectral ranking algorithms; these include the rank centrality algorithm for pairwise comparisons, which returns consistent estimates under the Bradley-Terry-Luce (BTL) model for pairwise comparisons (Negahban et al., 2017), and its generalization, the Luce spectral ranking algorithm, which returns consistent estimates under the more general multinomial logit (MNL) model for multiway comparisons (Maystre & Grossglauser, 2015). In this paper, we design a provably faster spectral ranking algorithm, which we call accelerated spectral ranking (ASR), that is also consistent under the MNL/BTL models.

Our accelerated algorithm is achieved by designing a random walk that has a faster mixing time than the random walks associated with previous algorithms. In addition to a faster algorithm, our results yield improved sample complexity bounds for recovery of the MNL/BTL parameters: to the best of our knowledge, we give the first general sample complexity bounds for recovering the parameters of the MNL model from multiway comparisons under any (connected) comparison graph (and improve significantly over previous bounds for the BTL model for pairwise comparisons). We also give a message-passing interpretation of our algorithm, which suggests a decentralized distributed implementation. Our experiments on several real world and synthetic datasets confirm that our new ASR algorithm is indeed orders of magnitude faster than existing algorithms.

[1]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, USA. Correspondence to: Arpit Agarwal <aarpit@seas.upenn.edu>, Shivani Agarwal <ashivani@seas.upenn.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

1. Introduction

The problem of rank aggregation from pairwise or multiway comparisons is a fundamental one in machine learning with applications in recommendation systems, sports, social choice etc. In this problem, given pairwise or multiway comparisons among n items, the goal is to learn a score for each item. These scores can further be used to produce a ranking over these items. For example, in recommendation systems, the goal might be to learn a ranking over items by observing the choices that users make when presented with different subsets of these items; in sports, the goal might be to rank teams/individuals at the end of a tournament based on pairwise or multiway games between these individuals/teams; in social choice, the goal might be to aggregate the choices of individuals when presented with different alternatives such as candidates in an election.

In the case of pairwise comparisons, a popular model is the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Luce, 1959) which posits that given a set of n items, there is a positive weight w_i associated with each item i, and the probability that i is preferred over j in a pairwise comparison between i and j is $\frac{w_i}{w_i+w_j}$. The BTL model is a special case of the multinomial logit (MNL)/Plackett-Luce model (Plackett, 1975; McFadden, 1974) which is defined for more general multiway comparisons. Under the MNL model, the probability that an item i is preferred amongst all items in a set S is $\frac{w_i}{\sum_{j \in S} w_j}$.

Rank aggregation under pairwise comparisons has been an active area of research, and several algorithms have been proposed that are consistent under the BTL model (Negahban et al., 2017; Rajkumar & Agarwal, 2014; Hunter, 2004; Chen & Suh, 2015; Jang et al., 2016; Guiver & Snelson, 2009; Soufiani et al., 2013). The case of multiway comparisons has also received some attention recently (Maystre & Grossglauser, 2015; Jang et al., 2017; Chen et al., 2017). Two popular algorithms are the rank centrality (RC) algorithm (Negahban et al., 2017) for the case of pairwise comparisons, and its generalization to the case of multiway comparisons, called the Luce spectral ranking (LSR) algorithm (Maystre & Grossglauser, 2015). The key idea behind these algorithms is to construct a *random walk* (equivalently a Markov chain) over the *comparison graph* on n items, where there is an edge between two items if they are com-

pared in a pairwise or multiway comparison. This random walk is constructed such that its stationary distribution corresponds to the weights of the MNL/BTL model.

Given the widespread application of these algorithms, understanding their computational aspects is of paramount importance. For random walk based algorithms this amounts to analyzing the mixing/convergence time of their random walks to stationarity. In the case of rank centrality and Luce spectral ranking, ensuring that the stationary distribution of the random walk corresponds to the weights of the underlying model forces their construction to have self loops with large mass. These self loops can lead to a large mixing time of $\Omega\big(\xi^{-1} d_{\max}\big)$, where $d_{\max}$ is the maximum number of unique comparisons that any item participates in; and ξ is the spectral gap of the graph Laplacian. In practical settings $d_{\max}$ can be very large, for example when the graph follows a power-law distribution, and can even be $\Omega(n)$ if one item is compared to a large fraction of the items.

In this paper we show that it is possible to construct a faster mixing random walk whose mixing time is $O\big(\xi^{-1}\big)$. We are able to construct this random walk by relaxing the condition that its stationary distribution should exactly correspond to the weights of the MNL model, and instead imposing a weaker condition that the weights can be recovered through a linear transform of the stationary distribution. We call the resulting algorithm accelerated spectral ranking (ASR).

In addition to computational advantages, the faster mixing property of our random walk also comes with statistical advantages, as it is well understood that faster mixing Markov chains lend themselves to tighter perturbation error bounds (Mitrophanov, 2005). We are able to establish a sample complexity bound of $O\big(\xi^{-2}\, n \,\mathrm{poly}(\log n)\big)$, in terms of the *total variation* distance, for recovering the true weights under the MNL (and BTL) model for almost any comparison graph of practical interest. To our knowledge, these are the first sample complexity bounds for the general case of multiway comparisons under the MNL model. Negahban et al. (2017) show similar results in terms of L_2 error for the special case of BTL model. However, their bounds have an additional dependence on $d_{\max}$, due to the large mixing time of their random walk.

We also show that our algorithm can be viewed as a message passing algorithm. This connection provides a very attractive property to our algorithm – it can be implemented in a distributed manner with decentralized communication and comparison data being stored in different machines.

We finally conduct several experiments on synthetic and real world datasets to compare the convergence time of our algorithm with the previous algorithms. These experiments confirm the behavior predicted by our theoretical analysis of mixing times– the convergence of our algorithm is in fact orders of magnitude faster than existing algorithms.

1.1. Our Contributions

We summarize our contributions as follows:

1. **Faster Algorithm:** We present an algorithm for aggregating pairwise comparisons under the BTL model, and more general multiway comparisons under the MNL model, that is provably faster than the previous algorithms of Negahban et al. (2017); Maystre & Grossglauser (2015). We also give experimental evidence supporting this fact.

2. **New and Improved Error Bounds:** We present the first error bounds for parameter recovery by spectral ranking algorithms under the general MNL model for any general (connected) comparison graph. These bounds improve upon the existing bounds of Negahban et al. (2017) for the special case of the BTL model.

3. **Message Passing Interpretation:** We provide an interpretation of our algorithm as a message passing/belief propagation algorithm. This connection can be used to design a decentralized distributed algorithm, which can work with distributed data storage.

1.2. Organization

In Section 2 we describe the problem formally. In Section 3 we present our algorithm for rank aggregation under the MNL/BTL model. In Section 4 we analyze the mixing time of our random walk, showing that our random walk converges much faster than existing approaches. In Section 5 we give bounds on sample complexity for recovery of MNL parameters with respect to the total variation distance. In Section 6 we give a message passing view of our algorithm. In Section 7 we provide experimental results on synthetic and real world datasets.

2. Problem Setting and Preliminaries

We consider a setting where there are n items, and one observes outcomes of noisy pairwise or multiway comparisons between these items. We will assume that the outcome of these comparisons is generated according to the multinomial logit (MNL) model, which posits that each item $i \in [n]$ is associated with a (unknown) weight/score $w_i > 0$, and the probability that item i wins a comparison is proportional to its weight w_i. More formally, when there is a (multiway) comparison between items of a set $S \subseteq [n]$, for $i \in S$, we have

$$p_{i|S} := \Pr(i \text{ is the most preferred item in } S) = \frac{w_i}{\sum_{j \in S} w_j}\,.$$

This model is also referred to as the Plackett-Luce model, and it reduces to the Bradley-Terry-Luce (BTL) model in the special case of pairwise comparisons, i.e. $|S| = 2$. Let

$\mathbf{w} \in \mathbb{R}^n_+$ be the vector of weights, i.e. $\mathbf{w} = (w_1, \cdots, w_n)^\top$. Note that this model is invariant to any scaling of $\mathbf{w}$, so for uniqueness we will assume that $\sum_{i=1}^n w_i = 1$, i.e. $\mathbf{w} \in \Delta_n$ where Δ_n is the n-dimensional probability simplex.

The comparison data is of the following form: there are d different *comparison sets* $S_1, \cdots, S_d \subseteq [n]$, with $|S_a| = m$ for all $a \in [d]$ and some constant $m < n$. For each set S_a, for $a \in [d]$, one observes the outcomes of L independent m-way comparisons between items in S_a, drawn according to the MNL model. The assumptions that each comparison set is of the same size m, and each set is compared an equal L number of times, are only for simplicity of exposition, and we give a generalization in the supplementary material. We will denote by y_a^l the winner of the l-th comparison amongst items of S_a, for $l \in [L]$ and $a \in [d]$.

Given comparison data $\mathbf{Y} = \{(S_a, \mathbf{y}_a)\}_{a=1}^d$, where $\mathbf{y}_a = (y_a^1, \cdots, y_a^L)$, the problem is to find a weight vector $\hat{\mathbf{w}} \in \Delta_n$, which is close to the true weight vector $\mathbf{w}$ under some notion of error/distance. More formally, the problem is to find $\hat{\mathbf{w}} \in \Delta_n$, such that $\|\hat{\mathbf{w}} - \mathbf{w}\|$ can be bounded in terms of the parameters n, L, and m, for some norm $\|\cdot\|$. We will give results in terms of the total variation distance, which for two vectors $\mathbf{u}, \hat{\mathbf{u}} \in \Delta_n$ is defined as

$$\|\mathbf{u} - \hat{\mathbf{u}}\|_{\text{TV}} = \frac{1}{2}\|\mathbf{u} - \hat{\mathbf{u}}\|_1 = \frac{1}{2}\sum_{i \in [n]} |u_i - \hat{u}_i| \,.$$

In the following sections, we will present an algorithm for recovering an estimate $\hat{\mathbf{w}}$ of $\mathbf{w}$, and give bounds on the error $\|\hat{\mathbf{w}} - \mathbf{w}\|_{\text{TV}}$ in terms of the problem parameters under natural assumptions on the comparison data.

3. Accelerated Spectral Ranking Algorithm

In this section, we will describe our algorithm, which we term as accelerated spectral ranking (ASR). Our algorithm is based on the idea of constructing a *random walk*[1] on the *comparison graph* with n vertices, which has an edge between nodes i and j if items i and j are compared in any m-way comparison. The key idea is to construct the random walk such that the probability of transition from node i to node j is proportional to w_j. If w_j is larger than w_i, then with other quantities being equal, one would expect the random walk to spend more time in node j than node i in its steady state distribution. Hence, if we can calculate the stationary distribution of this random walk, it might give us a way to estimate the weight vector $\mathbf{w}$. Moreover, for computational efficiency, we would also want this random walk to have a fast *mixing time*, i.e. it should rapidly converge to its stationary distribution.

The rank centrality (RC) algorithm (Negahban et al., 2017) for the BTL model, and its generalization the Luce spectral ranking (LSR) algorithm (Maystre & Grossglauser, 2015) for the MNL model, are based on a similar idea of constructing a random walk over the comparison graph. These algorithms construct a random walk whose stationary distribution, in expectation, is exactly $\mathbf{w}$. However, this construction forces their Markov chain to have self loops with large mass, slowing down the convergence rate.

In this section we will show that it is possible to design a *significantly* faster mixing random walk that belongs to a different class of random walks over the comparison graph. More precisely, the random walk that we construct is such that it is possible to recover the weight vector $\mathbf{w}$ from its stationary distribution using a fixed linear transformation, while for RC and LSR, the stationary distribution is exactly $\mathbf{w}$. Our theoretical analysis in Section 5 as well as experiments on synthetic and real world datasets in Section 7 will show that this difference can lead to vastly improved results.

Given comparison data $\mathbf{Y}$, let us denote by $G_c([n], E)$ the undirected graph on n vertices, with an edge $(i, j) \in E$ for any i, j that are a part of an m-way comparison. More formally, $(i, j) \in E$ if there exists an index $a \in [d]$ such that $i, j \in S_a$. We will call G_c the *comparison* graph, and throughout this paper, we will assume that $\mathbf{Y}$ is such that G_c is connected. We will denote by d_i the number of unique m-way comparisons of which $i \in [n]$ was a part, i.e. $d_i = \sum_{a \in [d]} \mathbf{1}[i \in S_a]$. Let $\mathbf{D} \in \mathbb{R}^{n \times n}$ be a diagonal matrix, with D_{ii} being equal to d_i, $\forall i \in [n]$. Also, let $d_{\max} := \max_i d_i$ and $d_{\min} := \min_i d_i$.

Suppose for each $a \in [d]$ and $j \in S_a$, one had access to the true probability $p_{j|S_a}$ of j being the most preferred item in S_a. Then one could define a random walk on G_c with transition probability from node $i \in [n]$ to $j \in [n]$ given by

$$P_{ij} := \frac{1}{d_i} \sum_{a \in [d]: i, j \in S_a} p_{j|S_a} = \frac{1}{d_i} \sum_{a \in [d]: i, j \in S_a} \frac{w_j}{\sum_{j' \in S_a} w_{j'}} \,. \tag{1}$$

Let $\mathbf{P} := [P_{ij}]$. One can verify that $\mathbf{P}$ corresponds to a valid transition probability matrix as it is non-negative and row stochastic. Furthermore, $\mathbf{P}$ defines a reversible Markov chain as it satisfies the detailed balance equations

$$w_i \, d_i \, P_{ij} = w_j \, d_j \, P_{ji} \,,$$

for all $i, j \in [n]$. If the graph G_c is connected then $\boldsymbol{\pi} = \mathbf{D}\,\mathbf{w}/\|\mathbf{D}\,\mathbf{w}\|_1$ is the unique stationary distribution of $\mathbf{P}$, and one can recover the true weight vector $\mathbf{w}$ from this stationary distribution using a linear transform $\mathbf{D}^{-1}$.

In practice one does not have access to $\mathbf{P}$, so we propose an *empirical* estimate of $\mathbf{P}$ that can be computed from the given comparison data. Formally, define $\hat{p}_{i|S_a}$ to be the fraction of times that i won a m-way comparison amongst items in the

[1] Throughout this paper we will use the terminology Markov chain and random walk interchangeably.

Algorithm 1 ASR

Input Markov chain $\hat{\mathbf{P}}$ according to Eq. (2)
Initialize $\hat{\boldsymbol{\pi}} = (\frac{1}{n}, \cdots, \frac{1}{n})^\top \in \Delta_n$
while estimates do not converge **do**
  $\hat{\boldsymbol{\pi}} \leftarrow \hat{\mathbf{P}}^\top \hat{\boldsymbol{\pi}}$
end while
Output $\hat{\mathbf{w}} = \frac{\mathbf{D}^{-1}\hat{\boldsymbol{\pi}}}{\|\mathbf{D}^{-1}\hat{\boldsymbol{\pi}}\|_1}$

set S_a, i.e. $\hat{p}_{i|S_a} := \frac{1}{L}\sum_{l=1}^{L} \mathbf{1}[y_a^l = i]$. Let us then define a random walk where the probability of transition from node $i \in [n]$ to node $j \in [n]$ is given by

$$\hat{P}_{ij} := \frac{1}{d_i} \sum_{a \in [d]: i,j \in S_a} \hat{p}_{j|S_a} \,. \tag{2}$$

Let $\hat{\mathbf{P}} := [\hat{P}_{ij}]$. One can again verify that $\hat{\mathbf{P}}$ corresponds to a valid transition probability matrix. We can think of $\hat{\mathbf{P}}$ as a perturbation of $\mathbf{P}$, with the error due to perturbation decreasing with more and more comparisons. There is a rich literature (Cho & Meyer, 2001; Mitrophanov, 2005) on analyzing sensitivity of the stationary distribution of a Markov chain under small perturbations. Hence, given a large number of comparisons, one can expect the stationary distribution of $\hat{\mathbf{P}}$ to be close to that of $\mathbf{P}$. Since we take a linear transform of these stationary distributions, one also needs to show that closeness is preserved under this linear transform. We defer this analysis to Section 5.

The pseudo-code for our algorithm is given in Algorithm 1. The algorithm computes the stationary distribution $\hat{\boldsymbol{\pi}}$ of the Markov chain $\hat{\mathbf{P}}$ using the power method.[2] It then outputs the (normalized) vector $\hat{\mathbf{w}}$ that is obtained after applying the linear transform $\mathbf{D}^{-1}$ to $\hat{\boldsymbol{\pi}}$, i.e. $\hat{\mathbf{w}} = \frac{\mathbf{D}^{-1}\hat{\boldsymbol{\pi}}}{\|\mathbf{D}^{-1}\hat{\boldsymbol{\pi}}\|_1}$. In the next section we will compare the convergence time of our algorithm with previous algorithms (Negahban et al., 2017; Maystre & Grossglauser, 2015).

4. Comparison of Mixing Time with Rank Centrality (RC) and Luce Spectral Ranking (LSR)

The random walk $\mathbf{P}^{\text{RC}}$ constructed by the RC (Negahban et al., 2017) algorithm for the BTL model is given by

$$P_{ij}^{\text{RC}} := \begin{cases} \frac{1}{d_{\max}} \sum_{a \in [d]: i,j \in S_a} p_{j|S_a} & \text{if } i \neq j \\ 1 - \frac{1}{d_{\max}} \sum_{j' \neq i} P_{ij'}^{\text{RC}} & \text{if } i = j \end{cases}, \tag{3}$$

[2]The stationary distribution of the Markov chain may also be computed using other linear algebraic techniques, but these techniques typically have a running time of $O(n^3)$ which is impractical for most modern applications.

and the random walk $\mathbf{P}^{\text{LSR}}$ constructed by LSR (Maystre & Grossglauser, 2015) for the MNL model is given by

$$P_{ij}^{\text{LSR}} := \begin{cases} \epsilon \sum_{a \in [d]: i,j \in S_a} p_{j|S_a} & \text{if } i \neq j \\ 1 - \epsilon \sum_{j' \neq i} P_{ij'}^{\text{LSR}} & \text{if } i = j \end{cases}, \tag{4}$$

where $\epsilon > 0$ is chosen such that the diagonal entries are non-negative. In general ϵ would be $O(\frac{1}{d_{\max}})$. The random walks $\hat{\mathbf{P}}^{\text{RC}}$ and $\hat{\mathbf{P}}^{\text{LSR}}$ constructed from the comparison data are defined analogously using empirical probabilities $\hat{p}_{j|S_a}$ instead of $p_{j|S_a}$.

We first begin by showing that for any given comparison data $\mathbf{Y}$, both RC/LSR and our algorithm will return the same estimate upon convergence.

Proposition 1. *Given items $[n]$ and comparison data $\mathbf{Y} = \{(S_a, \mathbf{y}_a)\}_{a=1}^d$, let $\hat{\boldsymbol{\pi}}$ be the stationary distribution of the Markov chain $\hat{\mathbf{P}}$ constructed by* ASR*, and let $\hat{\mathbf{w}}^{LSR}$ be the stationary distribution of the Markov chain $\hat{\mathbf{P}}^{LSR}$. Then $\hat{\mathbf{w}}^{LSR} = \frac{\mathbf{D}^{-1}\hat{\boldsymbol{\pi}}}{\|\mathbf{D}^{-1}\hat{\boldsymbol{\pi}}\|_1}$. The same result is also true for $\hat{\mathbf{w}}^{RC}$ for the case of pairwise comparisons.*

We give a proof of this result in the supplementary material. Although the above lemma shows that in a convergent state both these algorithms will return the same estimates, it does not say anything about the time it takes to reach this convergent state. This is where the *key difference* lies.

Observe that each row $i \in [n]$ of our matrix $\mathbf{P}$ is divided by d_i, whereas each row of $\mathbf{P}^{\text{RC}}$ is divided by $d_{\max}$ except the diagonal entries. Now if $d_{\max}$ is very large, a row $i \in [n]$ of $\mathbf{P}^{\text{RC}}$ that corresponds to an item i with small d_i would have very small non-diagonal entries. This can make the diagonal entry P_{ii}^{RC} very large, which amounts to having a heavy self loop at node i. This heavy self loop can significantly reduce the time it takes for the random walk to reach its stationary distribution, since a lot of transitions starting from i will return back to i. The same analysis holds true for LSR under multiway comparisons.

To formalize this intuition, we need to analyze the spectral gap of a random walk $\mathbf{X}$, which we denote by $\mu(\mathbf{X})$, which plays an important role in determining its mixing time. The spectral gap of a reversible random walk (or Markov chain) $\mathbf{X}$ is defined as $\mu(\mathbf{X}) := 1 - \lambda_2(\mathbf{X})$, where $\lambda_2(\mathbf{X})$ is the second largest eigenvalue of $\mathbf{X}$ in terms of absolute value. The following lemma (see Levin et al. (2008) for more details) gives both upper and lower bounds on the mixing time (w.r.t. the total variation distance) of a random walk in terms of the spectral gap.

Lemma 1. *(Levin et al., 2008) Let $\mathbf{X}$ be the transition probability matrix of a reversible, irreducible Markov chain with state space $[n]$, $\boldsymbol{\pi}$ be the stationary distribution of $\mathbf{X}$, and $\pi_{\min} := \min_{i \in [n]} \pi_i$, and let*

$$d(r) = \sup_{\mathbf{p} \in \Delta_n} \|\mathbf{p}\mathbf{X}^r - \boldsymbol{\pi}\|_{TV} \,.$$

For any $\gamma > 0$, let $t(\gamma) = \min\{r \in \mathbb{N} : d(r) \leq \gamma\}$; then

$$\log(\frac{1}{2\gamma})\Big(\frac{1}{\mu(\mathbf{X})} - 1\Big) \leq t(\gamma) \leq \log(\frac{1}{\gamma\pi_{\min}})\frac{1}{\mu(\mathbf{X})}\,.$$

The above lemma states that the mixing time of a Markov chain $\mathbf{X}$ is inversely proportional to its spectral gap $\mu(\mathbf{X})$. Now, we will compare the spectral gap of our Markov chain $\mathbf{P}$ with the spectral gap of $\mathbf{P}^{\text{RC}}$ (and $\mathbf{P}^{\text{LSR}}$).

Proposition 2. *Let the probability transition matrix $\mathbf{P}$ for our random walk be as defined in Eq. (1). Let $\mathbf{P}^{RC}$ and $\mathbf{P}^{LSR}$ be as defined in Eq. (3) and Eq. (4), respectively. Then*

$$\frac{d_{\min}}{d_{\max}}\mu(\mathbf{P}) \leq \mu(\mathbf{P}^{RC}) \leq \mu(\mathbf{P})\,, \tag{5}$$

and

$$\epsilon d_{\min}\mu(\mathbf{P}) \leq \mu(\mathbf{P}^{LSR}) \leq \mu(\mathbf{P})\,, \tag{6}$$

where $\epsilon = O(\frac{1}{d_{\max}})$.

A formal proof of this lemma is given in the supplementary material, and uses comparison theorems for reversible Markov chains due to Diaconis & Saloff-Coste (1993). This lemma shows that the spectral gap of $\mathbf{P}$ is always lower bounded by that of $\mathbf{P}^{\text{RC}}$ (and $\mathbf{P}^{\text{LSR}}$), but can be much larger than it. In the latter case one would observe, using Lemma 1, that our algorithm will converge faster than the RC algorithm (and LSR). In fact there are instances where $O(d_{\max}/d_{\min}) = \Omega(n)$ and the leftmost inequalities in both Eq. (5) and Eq. (6) hold with equality. In these instances the convergence of our algorithm will be $\Omega(n)$ times faster. We give examples of two such instances.

Example 1. *Let $n = 3$, $m = 2$, $w_1 = 1/2, w_2 = 1/4$ and $w_3 = 1/4$. In the comparison data 1 is compared to both 2 and 3; but items 2 and 3 are not compared to each other. This implies that $d_1 = 2$, and $d_i = 1$ for $i \neq 1$. One can calculate the matrices $\mathbf{P}$ and $\mathbf{P}^{RC}$, and their respective eigenvalues, and observe that $\mu(\mathbf{P}) = 2\mu(\mathbf{P}^{RC})$.*

Example 2. *Let $m = 2$, $\mathbf{w} = (1/n, \cdots, 1/n)^\top$, and the comparison data be such that item 1 is compared to every other item, and no other items are compared to each other. This implies that $d_1 = n-1$, and $d_i = 1$ for $i \neq 1$. One can calculate the matrix $\mathbf{P}$ and $\mathbf{P}^{RC}$ again, and their respective eigenvalues, and observe that $\mu(\mathbf{P}) = (n-1)\cdot\mu(\mathbf{P}^{RC})$.*

Note that in the above lemma, we only show the relation between the spectral gaps of the matrices $\mathbf{P}$ and $\mathbf{P}^{\text{RC}}$, and not for any particular realization $\hat{\mathbf{P}}$ and $\hat{\mathbf{P}}^{\text{RC}}$. If the Markov chains $\hat{\mathbf{P}}$ and $\hat{\mathbf{P}}^{\text{RC}}$ are reversible, then identical results hold. However, similar results are very hard to prove for non-reversible Markov chains (Dyer et al., 2006). Nevertheless, for large L, one can expect the realized matrices $\hat{\mathbf{P}}$ and $\hat{\mathbf{P}}^{\text{RC}}$ to be close to their expected matrices $\mathbf{P}$ and $\mathbf{P}^{\text{RC}}$, respectively. Hence, using eigenvalue perturbation bounds (Horn & Johnson, 1990), one can show that the spectrum of $\hat{\mathbf{P}}$ and $\hat{\mathbf{P}}^{\text{RC}}$ is close to the spectrum of $\mathbf{P}$ and $\mathbf{P}^{\text{RC}}$, respectively. The same analysis holds true for LSR under multiway comparisons. In Section 7 we perform experiments on synthetic and real world datasets which empirically show that the mixing times of the realized Markov chains behave as predicted.

It has been observed that faster mixing rates of Markov chains gives us the ability to prove sharper perturbation bounds for these Markov chains (Mitrophanov, 2005). In the following section we will use these perturbation bounds to prove sharper sample complexity bounds for our algorithm.

5. Sample Complexity Bounds

In this section we will present sample complexity bounds for the estimates returned by ASR in terms of total variation distance. The following theorem gives an error bound in terms of the total variation distance for estimates $\hat{\mathbf{w}}$ of the MNL weights returned by our algorithm

Theorem 1. *Given items $[n]$ and comparison data $\mathbf{Y} = \{(S_a, \mathbf{y}_a)\}_{a=1}^d$, let each set S_a of cardinality m be compared L times, with outcomes $\mathbf{y}_a = (y_a^1, \cdots, y_a^L)$ produced as per a MNL model with parameters $\mathbf{w} = (w_1, \ldots, w_n)$, such that $\|\mathbf{w}\|_1 = 1$. If the random walk $\hat{\mathbf{P}}$ (Eq. (2)) on the comparison graph $G_c([n], E)$ induced by the comparison data $\mathbf{Y}$ is strongly connected, then the* ASR *algorithm (Algorithm 1) converges to a unique distribution $\hat{\mathbf{w}}$, which with probability $\geq 1 - 3n^{-(C^2-50)/25}$ satisfies the following error bound*[3]

$$\|\mathbf{w} - \hat{\mathbf{w}}\|_{TV} \leq \frac{C\,\kappa\,d_{\text{avg}}}{\mu(\mathbf{P})\,d_{\min}}\sqrt{\frac{\max\{m, \log(n)\}}{L}}\,,$$

where $\kappa = \log\left(\frac{d_{\text{avg}}}{d_{\min} w_{\min}}\right)$, $w_{\min} = \min_{i\in[n]} w_i$, $d_{\text{avg}} = \sum_{i\in[n]} w_i d_i$, $d_{\min} = \min_{i\in[n]} d_i$, $\mu(\mathbf{P})$ is the spectral gap of the random walk $\mathbf{P}$ (Eq. (1)), and C is any constant.

Recall from Section 3 that the Markov chain $\hat{\mathbf{P}}$ can be viewed as a perturbation of $\mathbf{P}$. To show that the stationary distributions of $\hat{\mathbf{P}}$ and $\mathbf{P}$ are close, we use the results of Mitrophanov (2005) on the stability of Markov chains under perturbations. We also show closeness is preserved under the linear transformation $\mathbf{D}^{-1}$, giving the final bound stated in the aforementioned theorem. We present a formal proof in the supplementary material.

In the error bound of Theorem 1, one can further bound the spectral gap $\mu(\mathbf{P})$ of $\mathbf{P}$ in terms of the spectral gap of the *random walk normalized Laplacian* of G_c, which is a

[3] The dependence on κ is due to the dependence on $\frac{1}{\pi_{\min}}$ in the mixing time upper bounds in Lemma 1. There are other bounds for κ in terms of the condition number for Markov chains, for example see (Mitrophanov, 2005), and any improvement on these bounds will lead to an improvement in our sample complexity. In the worst case, κ has a trivial upper bound of $O(\log n)$.

fundamental quantity associated with G_c. The Laplacian represents a random walk on G_c that transitions from a node i to one of its neighbors uniformly at random. Formally, the Laplacian $\mathbf{L} := \mathbf{C}^{-1}\mathbf{A}$, where $\mathbf{C}$ is a diagonal matrix with $C_{ii} = \left|\bigcup_{a\in[d]:i\in S_a} S_a\right|$, i.e. the number of unique items i was compared with, and $\mathbf{A}$ is the adjacency matrix, such that for $i, j \in [n]$, $A_{ij} = 1$ if $(i, j) \in E$, and $A_{ij} = 0$ otherwise. Let $\xi := \mu(\mathbf{L})$ be the spectral gap of $\mathbf{L}$. Then we can lower bound $\mu(\mathbf{P})$ as follows (proof in the supplement)

$$\mu(\mathbf{P}) \geq \frac{\xi}{m\, b^2}\,,$$

where b is the ratio of the maximum to the minimum weight, i.e. $b = \max_{i,j\in[n]} w_i/w_j$. This gives us the following.

Corollary 1. *In the setting of Theorem 1, the* ASR *algorithm converges to a unique distribution* $\hat{\mathbf{w}}$*, which with probability* $\geq 1 - 3n^{-(C^2-50)/25}$ *satisfies the following error bound:*

$$\|\mathbf{w} - \hat{\mathbf{w}}\|_{TV} \leq \frac{C\, m\, b^2\, \kappa\, d_{\text{avg}}}{\xi\, d_{\min}} \sqrt{\frac{\max\{m, \log(n)\}}{L}}\,,$$

where $b = \max_{i,j\in[n]} \frac{w_i}{w_j}$.

In the discussion that follows, we will assume $b = O(1)$, and hence, $\mu(\mathbf{P}) = \Omega(\xi/m)$. The quantity d_{avg} has an interesting interpretation: it is the weighted average of the number of sets in which each item was shown. It has a trivial upper bound of $d_{\max}$, however, a careful analysis will reveal a better bound of $O(|E|/n)$ where E is the set of edges in the comparison graph G_c. Using this observation we can give the following corollary of the above theorem.

Corollary 2. *If the conditions of Theorem 1 are satisfied, and if the number of edges in the comparison graph* G_c *are* $O(n\,\text{poly}(\log n))$*, i.e.* $|E| = O(n\,\text{poly}(\log n))$*, then in order to ensure a total variation error of* $o(1)$*, the required number of comparisons per set is upper bounded as*

$$L = O\big(\mu(\mathbf{P})^{-2}\,\text{poly}(\log n)\big) = O\big(\xi^{-2}\, m^3\,\text{poly}(\log n)\big)\,.$$

Hence, the sample complexity, i.e. total number of m*-way comparisons needed to estimate* $\mathbf{w}$ *with error* $o(1)$*, is given by* $|E| \times L = O\big(\xi^{-2}\, m^3\, n\,\text{poly}(\log n)\big)$.

We again provide a proof of this corollary in the appendix. Note that the case when the total number of edges in the comparison graph is $O(n\,\text{poly}(\log n))$ captures the most interesting case in ranking and sorting. Also, in most practical settings the size m of comparison sets will be $O(\log n)$. In this case, the above corollary implies a sample complexity bound of $O\big(\xi^{-2}\, n\,\text{poly}(\log n)\big)$, which is sometimes referred to as *quasi-linear* complexity. The following simple example illustrates this sample complexity bound.

Example 3. *Consider a star comparison graph, discussed in Example 2, where there is one item* $i \in [n]$ *that is compared to all other* $n - 1$ *items, and no other items are compared to each other. Let* $\mathbf{w} = (\frac{1}{n}, \cdots, \frac{1}{n})^\top$*. One can calculate the spectral gap* $\mu(\mathbf{P})$ *to be* 0.5 *exactly. In this case, the sample complexity bound given by our result is* $O(n\,\text{poly}(\log n))$.

Discussion/Comparison. For the special case of pairwise comparisons under the BTL model ($m = 2$), Negahban et al. (2017) give a sample complexity bound of $O\big(\frac{d_{\max}}{d_{\min}}\xi^{-2}\, n\,\text{poly}(\log n)\big)$ for recovering the estimates $\hat{\mathbf{w}}$ with low (normalized) L_2 error. Using Proposition 1 one can see that this bound also applies to the estimates returned by our algorithm, and our bound in terms of L_1 applies to rank centrality as well. However, the bounds due to Negahban et al. (2017) have a dependence on the ratio $\frac{d_{\max}}{d_{\min}}$ due to the large spectral gap of their Markov chain as compared to ξ, the spectral gap of the Laplacian. In Section 7 we show that for many real world datasets $\frac{d_{\max}}{d_{\min}}$ can be much larger than $\log n$, and hence, their bounds are no longer quasi-linear. A large class of graphs that occur in many real world scenarios and exhibit this behavior are the power-law graphs. Another real world scenario in which $\frac{d_{\max}}{d_{\min}} = \Omega(n)$ arises is choice modeling (Agrawal et al., 2016), where one explicitly models the *'no choice option'* where the user has an option of not selecting any item from the set of items presented to her. In this case the 'no choice option' will be present in each comparison set, and the comparison graph will behave like a star graph discussed in Example 2. In fact for such graphs, the results of (Negahban et al., 2017) give a trivial bound of $\text{poly}(n)$ in terms of the L_2 error.

For the general case of multiway comparisons we are not aware of any other sample complexity bounds. It is also important to note that the dependence on the number of comparison sets comes only through the spectral gap ξ of the natural random walk on the comparison graph. For example, if the graph is a cycle ($d = n$), then the spectral gap is $O(1/n^2)$, whereas if the graph is a clique ($d = O(n^2)$) the spectral gap is $O(1)$.

6. Message Passing Interpretation of ASR

In this section, we show our spectral ranking algorithm can be interpreted as a message passing/belief propagation algorithm. This connection can be used to design a decentralized distributed version of our algorithm.

Let us introduce the *factor graph*, which is an important data structure used in message passing algorithms. The factor graph is a bipartite graph $G_f([n] \cup [d], E_f)$ which has two type of nodes– *item nodes* which correspond to the n items, and *set nodes* which correspond to the d sets. More formally, there is an item node i for each item $i \in [n]$, and there is a set node a for each set S_a, $\forall a \in [d]$. There is an edge $(i, a) \in E_f$ between node i and a if and only if $i \in S_a$. There is a weight $\hat{p}_{i|S_a}$ on the edge (i, a) which corresponds

to the fraction of times i won in the set S_a.

Algorithm 2 Message Passing

Input Graph $G_f = ([n] \cup [d], E_f)$, edge $(i,a) \in E$ has weight $\hat{p}_{i|S_a}$
Initialize Set $m^{(0)}_{a\to i} \leftarrow m/n, \forall a \in [d], \forall i \in S_a$
for $t = 1, 2, \cdots$ until convergence **do**
  for all $i \in [n]$ **do** $m^{(t)}_{i\to a} = \frac{1}{d_i} \sum_{a':i\in S_{a'}} \hat{p}_{i|S_{a'}} \cdot m^{(t-1)}_{a'\to i}$
  for all $a \in [d]$ **do** $m^{(t)}_{a\to i} = \sum_{i'\in S_a} m^{(t)}_{i'\to a}$
end for
Set $\hat{w}_i \leftarrow m^{(t-1)}_{i\to a}, \forall i \in [n]$
Output $\hat{\mathbf{w}}/\|\hat{\mathbf{w}}\|_1$

We shall now describe the algorithm. In each iteration of this algorithm, the item nodes send a message to their neighboring set nodes, and the set nodes respond to these messages. A message from an item node i to a set node a represents an estimate of the weight w_i of item i, and a message from a set node a to an item i represents an estimate of the sum of weights of items contained in set S_a.

In each iteration, the item nodes update their estimates based on the messages they receive in the previous iteration, and send these estimates to their neighboring set nodes. The set nodes then update their estimate by summing up the messages they receive from their neighboring item nodes, and then send these estimates to their neighboring item nodes. This process continues until the messages converge.

Formally, let $m^{(t-1)}_{i\to a}$ be the message from item node i to set node a in iteration $t-1$, and $m^{(t-1)}_{a\to i}$ be the corresponding message from the set node a to item node i. Then the messages in the next iteration are updated as follows:

$$m^{(t)}_{i\to a} = \frac{1}{d_i} \sum_{a'\in[d]:i\in S_{a'}} \hat{p}_{i|S_{a'}} \cdot m^{(t-1)}_{a'\to i} \,,$$

$$m^{(t)}_{a\to i} = \sum_{i'\in S_a} m^{(t)}_{i'\to a} \,.$$

Now, suppose that the empirical edge weights $\hat{p}_{i|S_a}$ are equal to the true weights $p_{i|S_a} = \frac{w_i}{\sum_{j\in S_a} w_j}, \forall i \in [n], a \in [d]$. Also, suppose on some iteration $t \geq 1$, the item messages $m^{(t)}_{i\to a}$ become equal to the item weights $w_i, \forall i \in [n]$. Then it is easy to observe that the next iteration of messages $m^{(t+1)}_{i\to a}$ are also equal to w_i. Therefore, the true weights $\mathbf{w}$, in some sense, are a fixed point of the above set of equations. The following lemma shows that the ASR algorithm is equivalent to this message passing algorithm.

Lemma 2. *For any realization of comparison data* $\mathbf{Y}$*, there is a one-to-one correspondence d each iteration of the message passing algorithm (2) and the corresponding power iteration of the ASR algorithm (1), and both algorithms return the same estimates* $\hat{\mathbf{w}}$ *for any* $\mathbf{Y}$.

We give a proof of the above lemma in the supplementary material. The above lemma gives an interesting connection between spectral ranking under the MNL model and message passing/belief propagation. Such connections have been observed for other problem such as the problem of aggregating crowdsourced binary tasks (Khetan & Oh, 2016). A consequence of this connection is that it facilitates a fully decentralized distributed implementation of the ASR algorithm. This can be very useful for modern applications, where machines can communicate local parameter updates to each other, without explicitly communicating the data.

7. Experiments

In this section we perform experiments on both synthetic and real data to compare our algorithm to the existing LSR (Maystre & Grossglauser, 2015) and RC (Negahban et al., 2017) algorithms for recovering the weight vector $\mathbf{w}$ under the MNL and BTL model, respectively. The implementation[4] of our algorithm is based on applying the power method on $\hat{\mathbf{P}}$ (Eq. (2)). The power method was chosen due to its simplicity, efficiency, and scalability to large problem sizes. Similarly, the implementations of LSR and RC are based on applying the power method on $\hat{\mathbf{P}}^{\text{LSR}}$ (Eq. (4)), and $\hat{\mathbf{P}}^{\text{RC}}$ (Eq. (3)), respectively. In the definition of $\hat{\mathbf{P}}^{\text{LSR}}$, the parameter ϵ was chosen to be the maximum possible value that ensures $\hat{\mathbf{P}}^{\text{LSR}}$ is a Markov chain.

7.1. Synthetic Data

We conducted experiments on synthetic data generated according to the MNL model, with weight vectors $\mathbf{w}$ generated randomly (details below). We compared our algorithm with the LSR algorithm for comparison sets of size $m = 5$, and with the RC algorithm for sets of size $m = 2$. We used two different graph topologies for generating the comparison graph G_c, or equivalently the comparison sets:

1. **Random Topology:** This graph topology corresponds to random graphs where $n \log_2(n)$ comparison sets are chosen uniformly at random from all the $\binom{n}{m}$ unique sets of cardinality m. This topology is very close to the Erdős-Rényi topology which has been well-studied in the literature. In fact the degree distributions of nodes in this random topology are very close to the degree distributions in the Erdős-Rényi topology (Mezard & Montanari, 2009). The only reason we study the former is computational, as iterating over all $\binom{n}{m}$ hyper-edges is computationally challenging.

2. **Star Topology:** In this graph topology, there is a single item that belongs to all sets; the remaining $(m-1)$ items in each set are contained only in that set. We study this topology because it corresponds to the choice sets used in Example 2, where there was a factor of $\Omega(n)$ gap in the

[4] code available: https://github.com/agarpit/asr

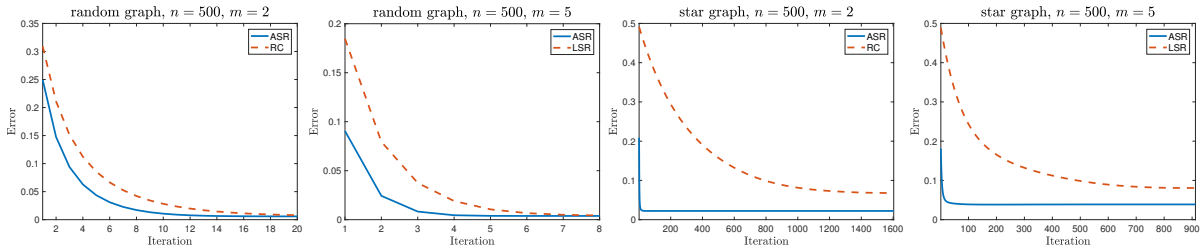

Figure 1. Results on synthetic data: L_1 error vs. number of iterations for our algorithm, ASR, compared with the RC algorithm (for $m = 2$) and the LSR algorithm (for $m = 5$), on data generated from the MNL/BTL model with the random and star graph topologies.

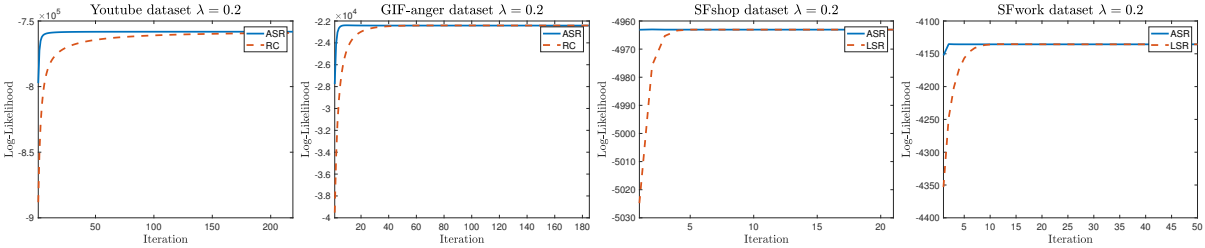

Figure 2. Results on real data: Log-likelihood vs. number of iterations for our algorithm, ASR, compared with the RC algorithm (for pairwise comparison data) and the LSR algorithm (for multi-way comparison data), all with regularization parameter set to 0.2.

spectral gap between our algorithm and the other algorithms.

In our experiments we selected $n = 500$[5], and the weight w_i of each item $i \in [n]$ was drawn uniformly at random from the range $(0, 1)$; the weights were then normalized so they sum to 1. A comparison graph G_c was generated according to each of the graph topologies above. The parameter L was set to $300 \log_2 n$. The winner for each comparison set was drawn according to the MNL model with weights $\mathbf{w}$. The convergence criterion for all algorithms was the same: we run the algorithm until the L_1 distance between the new estimates and the old estimates is ≤ 0.0001. Each experiment was repeated 100 times and the average values over all trials are reported. For $n = 500$, $m \in \{2, 5\}$, and both graph topologies described above, we compared the convergence as a function of the number of iterations[6] for each algorithm. We plotted the L_1 error of the estimates produced by these algorithms after each iteration. The plots are given in Figure 1. These plots verify the mixing time analysis of Section 4, and show that our algorithm converges much faster than RC and LSR, and orders of magnitude faster in the case of the star graph.

7.2. Real World Datasets

We conducted experiments on the YouTube dataset (Shetty, 2012), GIF-anger dataset (Rich et al.), and the SFwork and SFshop (Koppelman & Bhat, 2006) datasets. Table 1 gives some statistics about these datasets. We also plot the degree distributions of these datasets in the supplementary material.

[5]Results for other values of n are given in the supplement.

[6]We also plotted the convergence as a function of the running time; the results were similar as the running time of each iteration is similar for all these algorithm.

Table 1. Statistics for real world datasets

Dataset	n	m	d	$d_{\max}/d_{\min}$
Youtube	21207	2	394007	600
GIF-anger	6119	2	64830	106
SFwork	6	3-6	12	4.3
SFshop	8	4-8	10	1.9

For these datasets, a ground-truth $\mathbf{w}$ is either unknown or undefined; and hence, we compare our algorithm and the RC/LSR algorithm with respect to the log-likelihood of the estimates as a function of number of iterations. Due to the number of comparisons per set (or pair) being very small, in order to ensure irreducibility of random walks, we use a regularized version of all algorithms (see supplementary material, and also Section 3.3 in Negahban et al. (2017), for more details). Here, we give results when the regularization parameter λ is set to 0.2, and defer the results for other parameter values to the supplementary material. The results are given in Figure 2. We observe that our algorithm converges rapidly to the peak log-likelihood value while RC and LSR are always slower in converging to this value.

8. Conclusion and Future Work

We presented a spectral algorithm for the problem of rank aggregation from pairwise and multiway comparisons. Our algorithm is considerably faster than previous algorithms; in addition, our analysis yields improved sample complexity results for estimation under the BTL and MNL model. We also give a message passing/belief propagation interpretation for our algorithm. It would be interesting to see if one can use our algorithm to give better guarantees for recovery of top-k items under MNL.

Acknowledgements

We would like to thank Rohan Ghuge and Ashish Khetan for helpful discussions. We would also like to thank the anonymous reviewers for helpful comments. This material is based upon work supported by the US National Science Foundation under Grant No. 1717290. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Agarwal, A., Agarwal, S., Assadi, S., and Khanna, S. Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *COLT*, 2017.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. A near-optimal exploration-exploitation approach for assortment selection. In *EC*, 2016.

Arrow, K. J. Social choice and individual values, 1963.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Chen, X., Li, Y., and Mao, J. A nearly instance optimal algorithm for top-k ranking under the multinomial logit model. In *SODA*, 2017.

Chen, Y. and Suh, C. Spectral MLE: Top-k rank aggregation from pairwise comparisons. In *ICML*, 2015.

Cho, G. and Meyer, C. Comparison of perturbation bounds for the stationary distribution of a Markov chain. *Linear Algebra and its Applications*, 335(1-3):137–150, 2001.

Condorcet, M. d. Essay on the application of analysis to the probability of majority decisions. *Paris: Imprimerie Royale*, 1785.

de Borda, J. C. Mémoire sur les élections au scrutin. 1781.

Devroye, L. The equivalence of weak, strong and complete convergence in l1 for kernel density estimates. *The Annals of Statistics*, pp. 896–904, 1983.

Diaconis, P. and Saloff-Coste, L. Comparison theorems for reversible Markov chains. *The Annals of Applied Probability*, pp. 696–730, 1993.

Diaconis, P. and Stroock, D. Geometric bounds for eigenvalues of Markov chains. *The Annals of Applied Probability*, pp. 36–61, 1991.

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. Rank aggregation methods for the web. In *WWW*, 2001.

Dyer, M., Goldberg, L. A., Jerrum, M., Martin, R., et al. Markov chain comparison. *Probability Surveys*, 3:89–111, 2006.

Guiver, J. and Snelson, E. Bayesian inference for Plackett-Luce ranking models. In *ICML*, 2009.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge university press, 1990.

Hunter, D. R. MM algorithms for generalized Bradley-Terry models. *Annals of Statistics*, pp. 384–406, 2004.

Jang, M., Kim, S., Suh, C., and Oh, S. Top-k Ranking from Pairwise Comparisons: When Spectral Ranking is Optimal. *arXiv preprint arXiv:1603.04153*, 2016.

Jang, M., Kim, S., Suh, C., and Oh, S. Optimal sample complexity of m-wise data for top-k ranking. In *NIPS*, 2017.

Khetan, A. and Oh, S. Achieving budget-optimality with adaptive schemes in crowdsourcing. In *NIPS*, 2016.

Koppelman, F. S. and Bhat, C. A self instructing course in mode choice modeling: multinomial and nested logit models. *US Department of Transportation, Federal Transit Administration*, 2006.

Levin, D. A., Peres, Y., and Wilmer, E. L. *Markov Chains and Mixing Times*. American Mathematical Society, Providence, RI, USA, 2008.

Luce, D. R. Individual choice behavior. 1959.

Maystre, L. and Grossglauser, M. Fast and accurate inference of plackett-luce models. In *NIPS*, 2015.

McFadden, D. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pp. 105–142, 1974.

Mezard, M. and Montanari, A. *Information, Physics, and Computation*. Oxford University Press, Inc., New York, NY, USA, 2009.

Mitrophanov, A. Y. Sensitivity and convergence of uniformly ergodic Markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005.

Negahban, S., Oh, S., and Shah, D. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65 (1):266–287, 2017.

Plackett, R. L. The analysis of permutations. *Applied Statistics*, pp. 193–202, 1975.

Rajkumar, A. and Agarwal, S. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *ICML*, 2014.

Rich, T., Hu, K., and Tome, B. GIFGIF dataset. Data Available: http://www.gif.gf.

Shetty, S. Quantifying comedy on YouTube: why the number of o's in your LOL matter. Data Available: https://archive.ics.uci.edu/ml/datasets/YouTube+Comedy+Slam+Preference+Data, 2012.

Soufiani, H. A., Chen, W. Z., Parkes, D. C., and Xia, L. Generalized method-of-moments for rank aggregation. In *NIPS*, 2013.

Supplemental Material

A. Generalization of the ASR algorithm with Regularization

In this section, we shall present a generalized version of the ASR algorithm that relaxes the assumption that each set S_a is of the same fixed cardinality m, and each set S_a is compared the same number of times L. The intuition behind this generalization is that each comparison carries an equal amount of information, and thus, we should give a higher preference to the empirical estimates $\hat{p}_{i|S_a}$ corresponding to sets with more comparisons. Furthermore, comparisons on smaller sets are more reliable than comparisons on larger sets. In general, sets with larger cardinality should have proportionately more comparisons. Lastly, in practice, we often encounter comparison data for which the random walk $\hat{\mathbf{P}}$ on the comparison graph G_c is not strongly connected. We can resolve this issue through regularization. With these in mind, we update our algorithm as discussed below:

Given general comparison data $\mathbf{Y}' = \{(S_a, \mathbf{y}_a)_{a=1}^d\}$, where $S_a \subseteq [n]$ is of cardinality $|S_a|$, and $\mathbf{y}_a = (y_a^1, \ldots, y_a^{L_a})$, we define d_i' for each $i \in [n]$ as

$$d_i' := \sum_{a \in [d]: i \in S_a} \left(\frac{L_a}{|S_a|} + \lambda \right)$$

where λ is a regularization parameter. Intuitively, one can think of the regularization as adding $\lambda |S_a|$ pseudo-comparisons to each set S_a, with each item in the set winning an equal λ times. Furthermore, we define $n_{i|S_a}$ to be the number of times item $i \in S_a$ won in a $|S_a|$-way comparison amongst items in S_a, i.e. for all $a \in [d]$, for all $i \in S_a$,

$$n_{i|S_a} := \sum_{l=1}^{L_a} \mathbf{1}[y_a^l = i] \tag{7}$$

Using the above notation, we set up a Markov chain $\hat{\mathbf{P}}' \in \mathbb{R}_+^{n \times n}$ such that entry (i, j) is

$$\hat{P}'_{ij} := \frac{1}{d_i'} \sum_{a \in [d]: i, j \in S_a} \left(\frac{n_{j|S_a} + \lambda}{|S_a|} \right) \tag{8}$$

One can verify that this non-negative matrix is indeed row stochastic, hence corresponds to the transition matrix of a Markov chain. One can also verify that this construction reduces to a regularized version of $\hat{\mathbf{P}}$ (Eq. (2)) when all sets are of an equal size and are compared an equal number of times, and is identical to $\hat{\mathbf{P}}$ when $\lambda = 0$. Lastly, we define the matrix $\mathbf{D}'$ as a diagonal matrix, with diagonal entry $D'_{ii} := d_i'$, $\forall i \in [n]$. Similar to ASR, we compute the stationary distribution of $\hat{\mathbf{P}}'$, and output a (normalized) $\mathbf{D}'^{-1}$ transform of this stationary distribution.

Algorithm 3 Generalized-ASR

Input Markov chain $\hat{\mathbf{P}}'$ (according to Eq. (8))
Initialize $\hat{\boldsymbol{\pi}} = (\frac{1}{n}, \cdots, \frac{1}{n})^\top \in \Delta_n$
while estimates do not converge **do**
  $\hat{\boldsymbol{\pi}}' \leftarrow \hat{\mathbf{P}}'^\top \hat{\boldsymbol{\pi}}'$
end while
Output $\hat{\mathbf{w}}' = \frac{\mathbf{D}'^{-1}\hat{\boldsymbol{\pi}}'}{\|\mathbf{D}'^{-1}\hat{\boldsymbol{\pi}}'\|_1}$

B. Proof of Proposition 1

Proposition 1. *Given items $[n]$ and comparison data $\mathbf{Y} = \{(S_a, \mathbf{y}_a)\}_{a=1}^d$, let $\hat{\boldsymbol{\pi}}$ be the stationary distribution of the Markov chain $\hat{\mathbf{P}}$ constructed by* ASR*, and let $\hat{\mathbf{w}}^{LSR}$ be the stationary distribution of the Markov chain $\hat{\mathbf{P}}^{LSR}$. Then $\hat{\mathbf{w}}^{LSR} = \frac{\mathbf{D}^{-1}\hat{\boldsymbol{\pi}}}{\|\mathbf{D}^{-1}\hat{\boldsymbol{\pi}}\|_1}$. The same result is also true for $\hat{\mathbf{w}}^{RC}$ for the case of pairwise comparisons.*

Proof. Consider the estimates $\hat{\mathbf{w}} = \mathbf{D}^{-1}\hat{\boldsymbol{\pi}}/\|\mathbf{D}^{-1}\hat{\boldsymbol{\pi}}\|_1$ returned by the ASR algorithm upon convergence. In order to prove this lemma it is sufficient to prove that $\mathbf{D}\hat{\mathbf{w}}^{\text{LSR}}$ is an invariant measure (an eigenvector associated with eigenvalue 1) of the Markov chain $\hat{\mathbf{P}}$ corresponding to the ASR algorithm.

Since $\hat{\mathbf{w}}^{\text{LSR}}$ is the stationary distribution (also an eigenvector corresponding to eigenvalue 1) of $\hat{\mathbf{P}}^{\text{LSR}}$, we have

$$\hat{\mathbf{w}}^{\text{LSR}} = (\hat{\mathbf{P}}^{\text{LSR}})^\top \hat{\mathbf{w}}^{\text{LSR}}.$$

Following the definition (Eq. (4)) of $\hat{\mathbf{P}}^{\text{LSR}}$, we have the following relation for all $1 \leq i \leq n$

$$\begin{aligned}
\hat{w}_i^{\text{LSR}} &= \hat{w}_i^{\text{LSR}} \left(1 - \epsilon \sum_{j \neq i} \sum_{a: i, j \in S_a} p_{j|S_a} \right) \\
&\quad + \epsilon \sum_{j \neq i} \sum_{a: i, j \in S_a} p_{j|S_a} \hat{w}_j^{\text{LSR}} \\
\implies & \sum_{j \neq i} \sum_{a: i, j \in S_a} p_{j|S_a} \hat{w}_i^{\text{LSR}} = \sum_{j \neq i} \sum_{a: i, j \in S_a} p_{j|S_a} \hat{w}_j^{\text{LSR}}.
\end{aligned}$$

We shall use this relation to prove that $\hat{\mathbf{P}}^\top \mathbf{D}\hat{\mathbf{w}}^{\text{LSR}} = \mathbf{D}\hat{\mathbf{w}}^{\text{LSR}}$, where $\hat{\mathbf{P}}$ is the transition matrix corresponding to the Markov chain constructed by ASR. Consider the i^{th}

coordinate $[\hat{\mathbf{P}}^\top \mathbf{D}\hat{\mathbf{w}}^{\text{LSR}}]_i$ of the vector $\hat{\mathbf{P}}^\top \mathbf{D}\hat{\mathbf{w}}^{\text{LSR}}$

$$\begin{aligned}
[\hat{\mathbf{P}}^\top \mathbf{D}\hat{\mathbf{w}}^{\text{LSR}}]_i &= \frac{1}{d_i}\sum_{a:i\in S_a} p_{i|S_a} d_i \hat{w}_i^{\text{LSR}} \\
&\quad + \sum_{j\neq i}\frac{1}{d_j}\sum_{b:i,j\in S_b} p_{j|S_b} d_j \hat{w}_j^{\text{LSR}} \\
&= \sum_{a:i\in S_a} p_{i|S_a}\hat{w}_i^{\text{LSR}} + \sum_{j\neq i}\sum_{b:i,j\in S_b} p_{j|S_b}\hat{w}_i^{\text{LSR}} \\
&= \sum_{a:i\in S_a}\Big(\sum_{j\in S_a} p_{j|S_a}\Big)\hat{w}_i^{\text{LSR}} \\
&= \sum_{a:i\in S_a}\hat{w}_i^{\text{LSR}} \\
&= d_i\hat{w}_i^{\text{LSR}} = [\mathbf{D}\hat{\mathbf{w}}^{\text{LSR}}]_i\,,
\end{aligned}$$

where the second equality follows from the relation we proved earlier. Furthermore, this identity holds for all $1 \leq i \leq n$, from which we can conclude $\hat{\mathbf{P}}^\top \mathbf{D}\hat{\mathbf{w}}^{\text{LSR}} = \mathbf{D}\hat{\mathbf{w}}^{\text{LSR}}$. Furthermore, if the respective Markov chains induced by the comparison data are ergodic, then the corresponding stationary distributions must be unique, which is sufficient to prove both LSR and ASR return the same estimates upon convergence.

Since Luce spectral ranking is a generalization of the rank centrality algorithm, the transition matrix $\hat{\mathbf{P}}^{\text{LSR}}$ is identical to the transition matrix $\hat{\mathbf{P}}^{\text{RC}}$ in the pairwise comparison setting after setting $\epsilon = \frac{1}{d_{\max}}$, and thus, we can also conclude $\hat{\mathbf{P}}^\top \mathbf{D}\hat{\mathbf{w}}^{\text{RC}} = \mathbf{D}\hat{\mathbf{w}}^{\text{RC}}$. Thus, the statement of the lemma follows. □

C. Proof of Proposition 2

Proposition 2. *Let the probability transition matrix $\mathbf{P}$ for our random walk be as defined in Eq. (1). Let $\mathbf{P}^{RC}$ and $\mathbf{P}^{LSR}$ be as defined in Eq. (3) and Eq. (4), respectively. Then*

$$\frac{d_{\min}}{d_{\max}}\mu(\mathbf{P}) \leq \mu(\mathbf{P}^{RC}) \leq \mu(\mathbf{P})\,,$$

and

$$\epsilon d_{\min}\mu(\mathbf{P}) \leq \mu(\mathbf{P}^{LSR}) \leq \mu(\mathbf{P})\,,$$

where $\epsilon = O(\frac{1}{d_{\max}})$.

In order to prove this lemma, we will use the following result due to (Diaconis & Saloff-Coste, 1993) which compares the spectral gaps of two reversible random walks.

Lemma 3. *(Diaconis & Saloff-Coste, 1993) Let $\mathbf{Q}$ and $\mathbf{P}$ be reversible Markov chains on a finite set $[n]$ representing random walks on a graph $G = ([n], E)$, i.e. $P_{ij} = Q_{ij} = 0$ for all $(i,j) \notin E$. Let $\boldsymbol{\nu}$ and $\boldsymbol{\pi}$ be the stationary distributions of $\mathbf{Q}$ and $\mathbf{P}$, respectively. Then the spectral gaps of $\mathbf{Q}$ and $\mathbf{P}$ are related as*

$$\frac{\mu(\mathbf{P})}{\mu(\mathbf{Q})} \geq \frac{\alpha}{\beta}$$

where $\alpha := \min_{(i,j)\in E}\{\pi_i P_{ij}/\nu_i Q_{ij}\}$ and $\beta := \max_{i\in[n]}\{\pi_i/\nu_i\}$.

We are now ready to prove Proposition 2.

Proof. (of Proposition 2) To prove this lemma, we shall leverage the above comparison lemma due to (Diaconis & Saloff-Coste, 1993), that compares the spectral gaps of two arbitrary reversible Markov Chains. Let $\mathbf{P}$ (Eq. (2)) be the reversible Markov chain corresponding to ASR with stationary distribution $\boldsymbol{\pi} = \mathbf{D}\mathbf{w}/\|\mathbf{D}\mathbf{w}\|_1$, and let $\mathbf{P}^{\text{LSR}}$ (Eq. (4)) be the reversible Markov chain corresponding to LSR (RC in the pairwise case) with stationary distribution $\boldsymbol{\pi}^{\text{LSR}}$. Then by Lemma 3,

$$\frac{\mu(\mathbf{P}^{\text{LSR}})}{\mu(\mathbf{P})} \geq \frac{\alpha}{\beta}$$

where

$$\begin{aligned}
\alpha &:= \min_{(i,j):\exists a \text{ s.t. } i,j\in S_a}\left(\frac{\pi_i^{\text{LSR}} P_{ij}^{\text{LSR}}}{\pi_i P_{ij}}\right), \\
\beta &:= \max_{i\in[n]}\left(\frac{\pi_i^{\text{LSR}}}{\pi_i}\right).
\end{aligned}$$

From the definition of $\mathbf{P}$, and $\mathbf{P}^{\text{LSR}}$, we have

$$\begin{aligned}
\mathbf{P}_{ij} &= \frac{1}{d_i}\sum_{a\in[d]:i,j\in S_a}\frac{w_j}{\sum_{k\in S_a} w_k}, \\
\mathbf{P}_{ij}^{\text{LSR}} &= \epsilon\sum_{a\in[d]:i,j\in S_a}\frac{w_j}{\sum_{k\in S_a} w_k}
\end{aligned}$$

From the above equations and Proposition 1, it is easy to see that

$$\begin{aligned}
\alpha &= \epsilon\|\mathbf{D}\mathbf{w}\|_1, \quad \text{and} \\
\beta &= \frac{\|\mathbf{D}\mathbf{w}\|_1}{d_{\min}} \\
\implies \mu(\mathbf{P}^{\text{LSR}}) &\geq \epsilon d_{\min}(\mu(\mathbf{P}))
\end{aligned}$$

Following an identical line of reasoning, we have

$$\frac{\mu(\mathbf{P})}{\mu(\mathbf{P}^{\text{LSR}})} \geq \frac{\alpha'}{\beta'}$$

where

$$\begin{aligned}
\alpha' &= \min_{(i,j):\exists a \text{ s.t. } i,j\in S_a}\left(\frac{\pi_i P_{ij}}{\pi_i^{\text{LSR}} P_{ij}^{\text{LSR}}}\right), \\
\beta' &= \max_{i\in[n]}\left(\frac{\pi_i}{\pi_i^{\text{LSR}}}\right)
\end{aligned}$$

From the definition of $\mathbf{P}$, and $\mathbf{P}^{\text{LSR}}$, we have

$$\alpha' = \frac{1}{\|\mathbf{D}\mathbf{w}\|_1 \epsilon}, \quad \text{and}$$
$$\beta' = \frac{d_{\max}}{\|\mathbf{D}\mathbf{w}\|_1}$$
$$\implies \mu(\mathbf{P}) \geq \frac{1}{\epsilon d_{\max}} (\mu(\mathbf{P}^{\text{LSR}})) .$$

Since $\epsilon \leq 1/d_{\max}$, we get the following comparison between the spectral gaps of the Markov chains corresponding to the two approaches

$$\epsilon d_{\min} \mu(\mathbf{P}) \leq \mu(\mathbf{P}^{\text{LSR}}) \leq \mu(\mathbf{P}) .$$

The same analysis works for the Markov chain $\mathbf{P}^{\text{RC}}$ constructed by rank centrality for the pairwise comparison case with $\epsilon = 1/d_{\max}$, from which we can conclude

$$\frac{d_{\min}}{d_{\max}} \mu(\mathbf{P}) \leq \mu(\mathbf{P}^{\text{RC}}) \leq \mu(\mathbf{P}) .$$

□

D. Proof of Theorem 1

Theorem 1. *Given items $[n]$ and comparison data $\mathbf{Y} = \{(S_a, \mathbf{y}_a)\}_{a=1}^d$, let each set S_a of cardinality m be compared L times, with outcomes $\mathbf{y}_a = (y_a^1, \cdots, y_a^L)$ produced as per a MNL model with parameters $\mathbf{w} = (w_1, \ldots, w_n)$, such that $\|\mathbf{w}\|_1 = 1$. If the random walk $\hat{\mathbf{P}}$ (Eq. (2)) on the comparison graph $G_c([n], E)$ induced by the comparison data $\mathbf{Y}$ is strongly connected, then the* ASR *algorithm (Algorithm 1) converges to a unique distribution $\hat{\mathbf{w}}$, which with probability $\geq 1 - 3n^{-(C^2-50)/25}$ satisfies the following error bound*

$$\|\mathbf{w} - \hat{\mathbf{w}}\|_{TV} \leq \frac{C\, \kappa\, d_{\text{avg}}}{\mu(\mathbf{P})\, d_{\min}} \sqrt{\frac{\max\{m, \log(n)\}}{L}} ,$$

where $\kappa = \log\left(\frac{d_{\text{avg}}}{d_{\min} w_{\min}}\right)$, $w_{\min} = \min_{i\in[n]} w_i$, $d_{\text{avg}} = \sum_{i\in[n]} w_i d_i$, $d_{\min} = \min_{i\in[n]} d_i$, $\mu(\mathbf{P})$ is the spectral gap of the random walk $\mathbf{P}$ (Eq. (1)), and C is any constant.

Let us first state the concentration inequality for multinomial distributions due to (Devroye, 1983), which will be useful in proving this theorem.

Lemma 4 (Multinomial distribution inequality)**.** *(Devroye, 1983) Let $Y_1, \ldots, Y_n$ be a sequence of n independent random variables drawn from the multinomial distribution with parameters $(p_1, \ldots, p_k)$. Let X_i be the number of times i occurs in the n draws, i.e. $X_i = \sum_{j=1}^n \mathbf{1}[Y_j = i]$. For all $\epsilon \in (0, 1)$, and all k satisfying $k/n \leq \epsilon^2/20$, we have*

$$P(\sum_{i=1}^k |X_i - np_i| \geq n\epsilon) \leq 3\exp(-n\epsilon^2/25).$$

To prove Theorem 1, we shall first prove a bound on the total variation distance between the stationary states $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$ of the transition matrices $\mathbf{P}$ and $\hat{\mathbf{P}}$ respectively. We shall then prove a bound on the distance between the true weights $\mathbf{w}$ and estimates $\hat{\mathbf{w}}$ in terms of the distance between $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$.

An important result in the stability theory of Markov chains shows a connection between the stability of a chain and its speed of convergence to equilibrium (Mitrophanov, 2005). In fact, we can bound the sensitivity of a Markov chain under perturbation as a function of the convergence rate of the chain, with the accuracy of the sensitivity bound depending on the sharpness of the bound on the convergence rate. The following theorem is a specialization of Theorem 3.1 of (Mitrophanov, 2005), which gives perturbation bounds for Markov chains with general state spaces.

Theorem 2. *(Mitrophanov, 2005) Consider two discrete-time Markov chains $\mathbf{P}$ and $\hat{\mathbf{P}}$, with finite state space $\Omega = \{1, \ldots, n\}, n \geq 1$, and stationary distributions $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$, respectively. If there exist positive constants $1 < R < \infty$ and $\rho < 1$ such that*

$$\max_{x\in\Omega} \|\mathbf{P}^t(x, \cdot) - \boldsymbol{\pi}\|_{TV} \leq R\rho^t, \qquad \forall t \in \mathbb{N}$$

then for $\mathbf{E} := \mathbf{P} - \hat{\mathbf{P}}$, we have

$$\|\boldsymbol{\pi} - \hat{\boldsymbol{\pi}}\|_{TV} \leq \left(\hat{t} + \frac{1}{1-\rho}\right) \cdot \|\mathbf{E}\|_\infty .$$

where $\hat{t} = \log(R)/\log(1/\rho)$, and $\|\cdot\|_\infty$ is the matrix norm induced by the L_∞ vector norm.

It is well known that all ergodic Markov chains satisfy the conditions imposed by Theorem 2. In order to obtain sharp bounds on the convergence rate, we shall leverage the fact that the (unperturbed) Markov chain corresponding to the ideal transition probability matrix $\mathbf{P}$ is time-reversible.

Theorem 3. *(Diaconis & Stroock, 1991) Let $\mathbf{P}$ be an irreducible, reversible Markov chain with finite state space $\Omega = \{1, \ldots, n\}, n \geq 1$, and stationary distribution $\boldsymbol{\pi}$. Let $\lambda_2 := \lambda_2(\mathbf{P})$ be the second largest eigenvalue of $\mathbf{P}$ in terms of absolute value. Then for all $x \in \Omega$, $t \in \mathbb{N}$,*

$$\|\mathbf{P}^t(x, \cdot) - \boldsymbol{\pi}\|_{TV} \leq \sqrt{\frac{1 - \pi(x)}{4\pi(x)}} \lambda_2^t$$

Comparing these bounds with the conditions imposed by

Theorem 2, we can observe that

$$\begin{aligned}\rho &= \lambda_2,\\ R &= \max_{i\in[n]}\sqrt{\frac{1-\pi(i)}{4\pi(i)}}\\ &= \max_{i\in[n]}\sqrt{\frac{\|\mathbf{D}\mathbf{w}\|_1 - w_i d_i}{4w_i d_i}}\\ &\le \sqrt{\frac{d_{\text{avg}}}{4d_{\min}w_{\min}}},\end{aligned}$$

where $w_{\min} = \min_{i\in[n]} w_i$. Substituting these values into the perturbation bounds of Theorem 2, we get

$$\begin{aligned}\hat{t} + \frac{1}{1-\rho} &= \frac{\log(d_{\text{avg}}/(4d_{\min}w_{\min}))}{2\log(1/\lambda_2(\mathbf{P}))} + \frac{1}{1-\lambda_2(\mathbf{P})}\\ &\le \frac{\log(d_{\text{avg}}/(4d_{\min}w_{\min}))}{2(1-\lambda_2(\mathbf{P}))} + \frac{1}{1-\lambda_2(\mathbf{P})}\\ &< \frac{\kappa}{2\mu(\mathbf{P})}, \qquad \text{where } \kappa = \log(\frac{2d_{\text{avg}}}{d_{\min}w_{\min}})\end{aligned}$$

Now, the next step is to show that the perturbation error $\mathbf{E} := \mathbf{P} - \hat{\mathbf{P}}$ is bounded in terms of the matrix L_∞ norm.

Lemma 5. *For $\mathbf{E} := \mathbf{P} - \hat{\mathbf{P}}$, we have with probability $\ge 1 - 3n^{-(C^2-50)/25}$,*

$$\|\mathbf{E}\|_\infty \le C\sqrt{\frac{\max\{m, \log n\}}{L}}$$

where C is any constant.

Proof. By definition, $\|\mathbf{E}\|_\infty = \max_i \sum_{j=1}^n |\hat{P}_{ij} - P_{ij}|$. Fix any row $i \in [n]$. The probability that the absolute row sum exceeds a fixed positive quantity t is given by

$$\begin{aligned}&P(\sum_{j=1}^n |\hat{P}_{ij} - P_{ij}| \ge t)\\ &= P(\sum_{j=1}^n |\frac{1}{d_i}\sum_{a:i,j\in S_a}(\hat{p}_{j|S_a} - p_{j|S_a})| \ge t)\\ &= P(\sum_{j=1}^n |\frac{1}{d_i}\sum_{a:i,j\in S_a}\frac{1}{L}\sum_{l=1}^L(\mathbf{1}(y_a^l = j) - p_{j|S_a})| \ge t)\\ &\le P(\sum_{j=1}^n \sum_{a:i,j\in S_a} |\sum_{l=1}^L(\mathbf{1}(y_a^l = j) - p_{j|S_a})| \ge Ld_i t)\\ &= P(\sum_{a:i\in S_a}\sum_{j\in S_a} |\sum_{l=1}^L(\mathbf{1}(y_a^l = j) - p_{j|S_a})| \ge Ld_i t)\\ &\le d_i P(\sum_{j\in S_a} |\sum_{l=1}^L(\mathbf{1}(y_l^a = j) - p_{j|S_a})| \ge \frac{Ld_i t}{d_i})\end{aligned}$$

with the final pair of inequalities following from rearranging the terms in the summations and applying union bound. We leverage the multinomial distribution concentration inequality (Lemma 4) of Devroye (1983) to obtain the following bound for any set S_a for any m satisfying a technical condition $m/L \le t^2/20$.

$$P(\sum_{j\in S_a} |\sum_{l=1}^L(\mathbf{1}(y_l^a = j) - p_{j|S_a})| \ge Lt) \le 3\exp(\frac{-Lt^2}{25})$$

Thus, using union bound, the probability that any absolute row sum exceeds t is at most $3nd_{\max}\exp(-Lt^2/25)$. By selection of $t = 5C'\sqrt{\max\{m, \log n\}/L}$, we get

$$\begin{aligned}&P\left(\|\mathbf{E}\|_\infty \ge 5C'\sqrt{\frac{\max\{m,\log n\}}{L}}\right)\\ &\le 3n^2\exp\left(\frac{-25C'^2 L\max\{m,\log n\}}{25L}\right)\\ &\le 3n^{-(C'^2-2)}\end{aligned}$$

substituting $C = 5C'$ proves our claim. Lastly, one can verify that the aforementioned choice of t satisfies the technical condition imposed by Lemma 4 for any n, m and L. □

Combining the results of Theorem 2, Theorem 3, and Theorem 5 gives us a high confidence total variation error bound on the stationary states $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$ of the ideal and perturbed Markov chains $\mathbf{P}$ and $\hat{\mathbf{P}}$ respectively. Thus, with confidence $\ge 1 - 3n^{-(C^2-50)/25}$, we have

$$\|\boldsymbol{\pi} - \hat{\boldsymbol{\pi}}\|_{\text{TV}} \le \frac{C\kappa}{\mu(\mathbf{P})}\sqrt{\frac{\max\{m,\log n\}}{L}}, \tag{9}$$

where $\kappa = \log(2d_{\text{avg}}/(d_{\min}w_{\min}))$.

The last step in our scheme is to prove that the linear transformation $\mathbf{D}^{-1}\hat{\boldsymbol{\pi}}$ preserves this error bound up to a reasonable factor.

Lemma 6. *Under the conditions of Theorem 1, let $\boldsymbol{\pi} = \mathbf{D}\mathbf{w}/\|\mathbf{D}\mathbf{w}\|_1$ and $\hat{\boldsymbol{\pi}} = \mathbf{D}\hat{\mathbf{w}}/\|\mathbf{D}\hat{\mathbf{w}}\|_1$ be the unique stationary distributions of the Markov chains $\mathbf{P}$ (Eq. (1)) and $\hat{\mathbf{P}}$ (Eq. (2)) respectively. Then we have*

$$\|\mathbf{w} - \hat{\mathbf{w}}\|_{TV} \le \frac{d_{\text{avg}}}{d_{\min}}\|\boldsymbol{\pi} - \hat{\boldsymbol{\pi}}\|_{TV}.$$

Proof. We shall divide our proof into two cases.
Case 1: $\|\mathbf{D}\hat{\mathbf{w}}\|_1 \ge \|\mathbf{D}\mathbf{w}\|_1$.
Let us define the set $A = \{i : w_i \ge \hat{w}_i\}$, and the set $A' = \{j : \pi_j \ge \hat{\pi}_j\}$. When $\|\mathbf{D}\hat{\mathbf{w}}\|_1 \ge \|\mathbf{D}\mathbf{w}\|_1$, it is easy to see that $A \subseteq A'$.

Consider the total variation distance $\|\mathbf{w} - \hat{\mathbf{w}}\|_{TV}$ between the true preferences $\mathbf{w}$ and our estimates $\hat{\mathbf{w}}$. By definition,

$$\begin{aligned}
\|\mathbf{w} - \hat{\mathbf{w}}\|_{TV} &= \sum_{i \in A} (w_i - \hat{w}_i) \\
&= \sum_{i \in A} w_i \left(1 - \frac{\hat{w}_i}{w_i}\right) = \sum_{i \in A} w_i \left(1 - \frac{\hat{w}_i d_i}{w_i d_i}\right) \\
&\leq \sum_{i \in A} w_i \left(1 - \frac{\hat{w}_i d_i \|\mathbf{D}\mathbf{w}\|_1}{w_i d_i \|\mathbf{D}\hat{\mathbf{w}}\|_1}\right) \\
&= \sum_{i \in A} w_i \left(1 - \frac{\hat{\pi}_i}{\pi_i}\right) \\
&= \sum_{i \in A} w_i \left(\frac{(\pi_i - \hat{\pi}_i)\|\mathbf{D}\mathbf{w}\|_1}{w_i d_i}\right) \\
&\leq \sum_{j \in A'} w_j \left(\frac{(\pi_j - \hat{\pi}_j)\|\mathbf{D}\mathbf{w}\|_1}{w_j d_j}\right) \\
&= \sum_{j \in A'} \left(\frac{(\pi_j - \hat{\pi}_j)\|\mathbf{D}\mathbf{w}\|_1}{d_j}\right) \\
&\leq \frac{\|\mathbf{D}\mathbf{w}\|_1}{d_{\min}} \sum_{j \in A'} (\pi_j - \hat{\pi}_j) = \frac{d_{\text{avg}}}{d_{\min}} \|\pi - \hat{\pi}\|_{TV}
\end{aligned}$$

Case 2, where $\|\mathbf{D}\hat{\mathbf{w}}\|_1 < \|\mathbf{D}\mathbf{w}\|_1$ follows symmetrically, giving us the inequality

$$\begin{aligned}
\|\mathbf{w} - \hat{\mathbf{w}}\|_{TV} &\leq \frac{\|\mathbf{D}\hat{\mathbf{w}}\|_1}{d_{\min}} \|\pi - \hat{\pi}\|_{TV} \\
&\leq \frac{\|\mathbf{D}\mathbf{w}\|_1}{d_{\min}} \|\pi - \hat{\pi}\|_{TV} = \frac{d_{\text{avg}}}{d_{\min}} \|\pi - \hat{\pi}\|_{TV}
\end{aligned}$$

where the last inequality follows from the assumption of Case 2, proving our claim. □

Combining the above lemma with Eq. (9) gives us the statement of the theorem.

E. Proof of Corollary 1

Corollary 1. *In the setting of Theorem 1, the* ASR *algorithm converges to a unique distribution* $\hat{\mathbf{w}}$*, which with probability* $\geq 1 - 3n^{-(C^2-50)/25}$ *satisfies the following error bound:*

$$\|\mathbf{w} - \hat{\mathbf{w}}\|_{TV} \leq \frac{C\, m\, b^2\, \kappa\, d_{\text{avg}}}{\xi\, d_{\min}} \sqrt{\frac{\max\{m, \log(n)\}}{L}},$$

where $b = \max_{i,j \in [n]} \frac{w_i}{w_j}$.

Corollary 1 follows from the following lemma which compares the spectral gap of the matrix $\mathbf{P}$ with the spectral gap of the graph Laplacian.

Lemma 7. *Let* $\mathbf{L} := \mathbf{C}^{-1}\mathbf{A}$ *be the Laplacian of the undirected graph* $G_c([n], E)$*. Then the spectral gap* $\mu(\mathbf{P}) = 1 - \lambda_2(\mathbf{P})$ *of the reversible Markov chain* $\mathbf{P}$ *(Eq. (2)) corresponding to the* ASR *algorithm is related to the spectral gap* $\xi = 1 - \lambda_2(\mathbf{L})$ *of the Laplacian as*

$$\mu(\mathbf{P}) \geq \frac{\xi}{mb^2}$$

Proof. To prove this inequality, we shall leverage the comparison Lemma 3 of (Diaconis & Saloff-Coste, 1993), with $\mathbf{Q}, \boldsymbol{\nu} = \mathbf{L}, \boldsymbol{\nu}$. From the definition of the Laplacian, it is clear that for all i, $\nu_i \mathbf{L}_{ij} = 1/2|E|$. Furthermore, $\nu_i = c_i/2|E| \geq d_i/2|E|$, where c_i is the number of unique items i was compared with, which is trivially at least the number of unique multiway comparisons of which i was a part. Thus,

$$\begin{aligned}
\beta &:= \max_{i \in [n]} \frac{\pi_i}{\nu_i} = \max_{i \in [n]} \frac{w_i d_i / \|\mathbf{D}\mathbf{w}\|_1}{c_i/2|E|} \\
&\leq \frac{2|E| w_{\max}}{\|\mathbf{D}\mathbf{w}\|_1} \\
\alpha &:= \min_{(i,j) \in E} \frac{\pi_i P_{ij}}{\nu_i L_{ij}} \\
&= \min_{(i,j) \in E} \frac{\frac{w_i d_i}{\|\mathbf{D}\mathbf{w}\|_1} \frac{1}{d_i} \sum_{a:(i,j) \in S_a} \frac{w_j}{\sum_{k \in S_a} w_k}}{1/2|E|} \\
&\geq \frac{2|E| w_{\min}^2}{m w_{\max} \|\mathbf{D}\mathbf{w}\|_1}
\end{aligned}$$

Thus, $\alpha/\beta \geq 1/mb^2$, which proves our claim. □

F. Proof of Corollary 2

Corollary 2. *If the conditions of Theorem 1 are satisfied, and if the number of edges in the comparison graph* G_c *are* $O(n \operatorname{poly}(\log n))$*, i.e.* $|E| = O(n \operatorname{poly}(\log n))$*, then in order to ensure a total variation error of* $o(1)$*, the required number of comparisons per set is upper bounded as*

$$L = O\big(\mu(\mathbf{P})^{-2} \operatorname{poly}(\log n)\big) = O\big(\xi^{-2}\, m^3 \operatorname{poly}(\log n)\big).$$

Hence, the sample complexity, i.e. total number of m*-way comparisons needed to estimate* $\mathbf{w}$ *with error* $o(1)$*, is given by* $|E| \times L = O\big(\xi^{-2}\, m^3\, n \operatorname{poly}(\log n)\big)$.

In order to prove the above corollary we first give the following claim.

Claim 1. *Given items* $[n]$*, and comparison graph* $G_c = ([n], E)$ *induced by comparison data* $\mathbf{Y} = \{S_a, \mathbf{y}_a\}_{a=1}^d$*, let the vector of true MNL parameters be* $\mathbf{w} = (w_1, \ldots, w_n)$*. Furthermore, let* d_i *represent the number of unique comparisons of which item* $i \in [n]$ *was a part. Then we have*

$$d_{\text{avg}} = \sum_{i \in [n]} w_i d_i \leq \frac{2 w_{\max} |E|}{w_{\min} n},$$

where $w_{\max} = \max_{i\in[n]} w_i$, *and* $w_{\min} = \min_{j\in[n]} w_j$.

Proof. Clearly,

$$w_{\min} \sum_{i\in[n]} w_i d_i \le \frac{1}{n} \sum_{i\in[n]} w_i d_i \le \frac{w_{\max}}{n} \sum_{i\in[n]} d_i,$$

The statement of the lemma follows by realizing that $\sum_{i\in[n]} d_i \le \sum_{i\in[n]} c_i \le 2|E|$. □

Proof. (of Corollary 2) Substituting the above bound on d_{avg} in the sample complexity bounds of Corollary 1, we get the following guarantee on the total variation error between the estimates $\hat{\mathbf{w}}$ and the true weight vector $\mathbf{w}$

$$\|\mathbf{w} - \hat{\mathbf{w}}\|_{\text{TV}} \le \frac{C\, m\, b^3\, \kappa\, |E|}{n\, \xi\, d_{\min}} \sqrt{\frac{\max\{m, \log(n)\}}{L}},$$

where $b = \frac{w_{\max}}{w_{\min}}$. Furthermore, this guarantee holds with probability $\ge 1 - 3n^{-(C^2-50)/25}$. From this, we can conclude that if

$$L \ge \max\{m, \log(n)\} \left(\frac{10\, m\, b^3\, \kappa\, |E|}{n\, \xi\, d_{\min}}\right)^2,$$

then it is sufficient to guarantee that $\|\mathbf{w} - \hat{\mathbf{w}}\|_{\text{TV}} = o(1)$ with probability $\ge 1 - 3n^{-2}$. Trivially bounding $\kappa = O(\log n)$, and from the assumptions $b = O(1)$ and $|E| = O(n\,\text{poly}(\log n))$, we can conclude

$$L = O(\xi^{-2} m^3 \text{poly}(\log n))$$

where the additional m factor comes from trivially bounding $\max\{m, \log n\} \le m \log n$. This gives us a sample complexity bound of

$$|E| \times L = O(\xi^{-2}\, m^3\, n\, \text{poly}(\log n))$$

for our algorithm, which proves the corollary. □

G. Proof of Lemma 2

Lemma 2. *For any realization of comparison data* $\mathbf{Y}$*, there is a one-to-one correspondence d each iteration of the message passing algorithm (2) and the corresponding power iteration of the ASR algorithm (1), and both algorithms return the same estimates* $\hat{\mathbf{w}}$ *for any* $\mathbf{Y}$.

Proof. In the message passing algorithm, the item to set messages $m^{(r)}_{i\to a}$ in round r correspond to the estimates of the item weights. One can verify that the estimate $\hat{w}^{(r)}$ of item i in round r evolves according to the following equation.

$$\hat{w}_i^{(r+1)} = \frac{1}{d_i} \sum_{a: i\in S_a} p_{i|S_a} \cdot \sum_{j\in S_a} \hat{w}_j^{(r)}.$$

We can represent this system of equations compactly using the following matrices. Let $\hat{\mathbf{V}} \in \mathbb{R}^{d\times n}$ be a matrix such that

$$\hat{V}_{ai} := \begin{cases} \frac{p_{i|S_a}}{d_i} & \text{if } (i,a) \in E \\ 0 & \text{otherwise} \end{cases}, \tag{10}$$

and $\mathbf{B} \in \mathbb{R}^{n\times d}$ be a matrix such that

$$B_{ia} := \begin{cases} 1 & \text{if } (i,a) \in E \\ 0 & \text{otherwise} \end{cases}, \tag{11}$$

Thus, we can represent the weight update from round (r) to round $(r+1)$ as

$$\begin{aligned} \hat{\mathbf{w}}^{(r+1)} &= (\mathbf{B}\hat{\mathbf{V}})^\top \hat{\mathbf{w}}^{(r)} = \hat{\mathbf{M}}^\top \hat{\mathbf{w}}^{(r)} \\ &= (\hat{\mathbf{M}}^\top)^r \hat{\mathbf{w}}^{(0)}, \end{aligned}$$

where $\hat{\mathbf{M}} := \mathbf{B}\hat{\mathbf{V}}$, with entry (i,j) of $\hat{\mathbf{M}}$ being

$$\hat{M}_{ij} := \frac{1}{d_j} \sum_{a: i,j\in S_a} p_{j|S_a}. \tag{12}$$

The above equation implies that the message passing algorithm is essentially a power iteration on the matrix $\hat{\mathbf{M}}$. Now, it is easy to see that $\hat{\mathbf{M}} = \mathbf{D}\hat{\mathbf{P}}\mathbf{D}^{-1}$ where $\hat{\mathbf{P}}$ is the transition matrix constructed by ASR ($Eq.$ (2)). Therefore, there is a one-to-one correspondence between the power iterations on $\hat{\mathbf{M}}$ and $\hat{\mathbf{P}}$. More formally, if we initialize with $\hat{\mathbf{w}}^{(0)}$ in the power iteration on $\hat{\mathbf{M}}$, and initialize with $\hat{\boldsymbol{\pi}}^{(0)} = \mathbf{D}\hat{\mathbf{w}}^{(0)}$ in the power iteration on $\mathbf{P}$, then the iterates at the r-th step will be related as $\hat{\boldsymbol{\pi}}^{(r)} = \mathbf{D}\hat{\mathbf{w}}^{(r)}$. Furthermore, if $\hat{\boldsymbol{\pi}}$ is the stationary distribution of $\hat{\mathbf{P}}$, then $\hat{\mathbf{w}} = \mathbf{D}^{-1}\hat{\boldsymbol{\pi}}$ is the corresponding dominant left eigenvector of $\hat{\mathbf{M}}$, i.e. $\mathbf{D}^{-1}\hat{\boldsymbol{\pi}} = \hat{\mathbf{M}}^\top \mathbf{D}^{-1}\hat{\boldsymbol{\pi}}$. Also, $\hat{\mathbf{w}}$ is exactly the estimate (after normalization) returned by both the ASR and the message passing algorithm upon convergence. Thus, we can conclude that the message passing algorithm is identical to ASR for any realization of comparison data generated according to the MNL model. □

H. Additional Experimental Results

In this section we will describe additional experimental results comparing our algorithm and the RC/LSR algorithms on various synthetic and real world datasets. Since we require additional regularization when the random walk induced by comparison data is reducible, we will first describe the regularized version of the RC and LSR algorithms (regularized version of our algorithm is given in Appendix A).

H.1. RC and LSR algorithms with regularization

In this section, for the sake of completeness, we state the regularized version of the RC (Negahban et al., 2017) and

Table 2. Statistics for real world datasets

Dataset	n	m	d	total choices
Youtube	21207	2	394007	1138562
GIF-amusement	6118	2	75649	77609
GIF-anger	6119	2	64830	66505
GIF-contentment	6118	2	70230	72175
GIF-excitement	6119	2	80493	82564
GIF-happiness	6119	2	104801	107816
GIF-pleasure	6119	2	86499	88959
GIF-relief	6112	2	38770	39853
GIF-sadness	6118	2	63577	65263
GIF-satisfaction	6118	2	78401	80474
GIF-shame	6116	2	46249	47550
GIF-surprise	6118	2	63850	65591
SFWork	6	3-6	12	5029
SFShop	8	4-8	10	3157

LSR (Maystre & Grossglauser, 2015) algorithms.[7] These algorithms are based on computing the stationary distribution of a Markov chain. In the case of pairwise comparisons, for a regularization parameter $\lambda > 0$, the Markov chain $\hat{\mathbf{P}}'^{\text{RC}} := [\hat{P}'^{\text{RC}}_{ij}]$, where, $\forall i, j \in [n]$,

$$\hat{P}'^{\text{RC}}_{ij} := \begin{cases} \frac{1}{d_{\max}} \left(\frac{n_{j|\{i,j\}} + \lambda}{n_{j|\{i,j\}} + n_{i|\{i,j\}} + 2\lambda} \right), & \text{if } i \neq j \\ 1 - \frac{1}{d_{\max}} \sum_{j' \neq i} \hat{P}'^{\text{RC}}_{ij}, & \text{if } i = j \end{cases}$$

and $n_{j|\{i,j\}}$ is defined according to Eq. (7). In the case of multi-way comparisons, the Markov chain $\hat{\mathbf{P}}'^{\text{LSR}} := [\hat{P}'^{\text{LSR}}_{ij}]$, where, $\forall i, j \in [n]$,

$$\hat{P}'^{\text{LSR}}_{ij} := \begin{cases} \epsilon \sum_{a \in [d]: i,j \in S_a} \left(\frac{n_{j|S_a} + \lambda}{|S_a|} \right), & \text{if } i \neq j \\ 1 - \epsilon \sum_{j' \neq i} \hat{P}'^{\text{LSR}}_{ij}, & \text{if } i = j \end{cases}$$

where ϵ is a quantity small enough to make the diagonal entries of $\hat{\mathbf{P}}'^{\text{LSR}}$ non negative, and $n_{j|S_a}$ is again defined according to Eq. (7).

H.2. Synthetic Datasets

In this section, we give additional experimental results for various other values of parameters m and n. The plots are given in the figures below. The general trends observed from these experiments are exactly as predicted by our theoretical analysis. In particular, we note that even in the case of a star graph topology, the convergence rate of ASR remains essentially the same with increasing n, while the performance of RC and LSR degrades smoothly. This really conveys the low dependence on the ratio $d_{\max}/d_{\min}$.

[7] See Section 3.3 in Negahban et al. (2017) for more details.

H.3. Real Datasets

In this section, we provide additional experimental results for more datasets, and additional values of the regularization parameter λ. We conducted experiments on the YouTube dataset (Shetty, 2012), various GIF datasets (Rich et al.), and the SFwork and SFshop (Koppelman & Bhat, 2006) datasets. Below we briefly describe each of these datasets (additional statistics are given in Table 2).

1. **YouTube Comedy Slam Preference Data**. This dataset is due to a video discovery experiment on YouTube in which users were shown a pair of videos and were asked to vote for the video they found funnier out of the two.[8]

2. **GIFGIF datasets**. These datasets are due to a experiment that tries to understand the emotional content present in animated GIFs. In this experiment users are shown a pair of GIFs and asked to vote for the GIF that most accurately represents a particular emotion. These votes are collected for several different emotions.[9]

3. **SF datasets**. These datasets are from a survey of transportation preferences around the San Francisco Bay Area in which citizens were asked to vote on their preferred commute option amongst different options.[10]

As expected, the peak log likelihood decreases with increasing λ, as this regularization parameter essentially dampens the information imparted by the comparison data. We also plot degree distributions of these real world datasets in order to explore the behavior of the ratio $d_{\max}/d_{\min}$ in practice. In particular, we observe that this quantity does not really behave like a constant, and is very large in most cases. This is particularly evident in the Youtube dataset, where the degree distribution closely follows the power law relationship with n.

[8] See https://archive.ics.uci.edu/ml/datasets/YouTube+Comedy+Slam+Preference+Data for more details.

[9] See http://gif.gf for more details.

[10] These datasets are available at https://github.com/sragain/pcmc-nips.

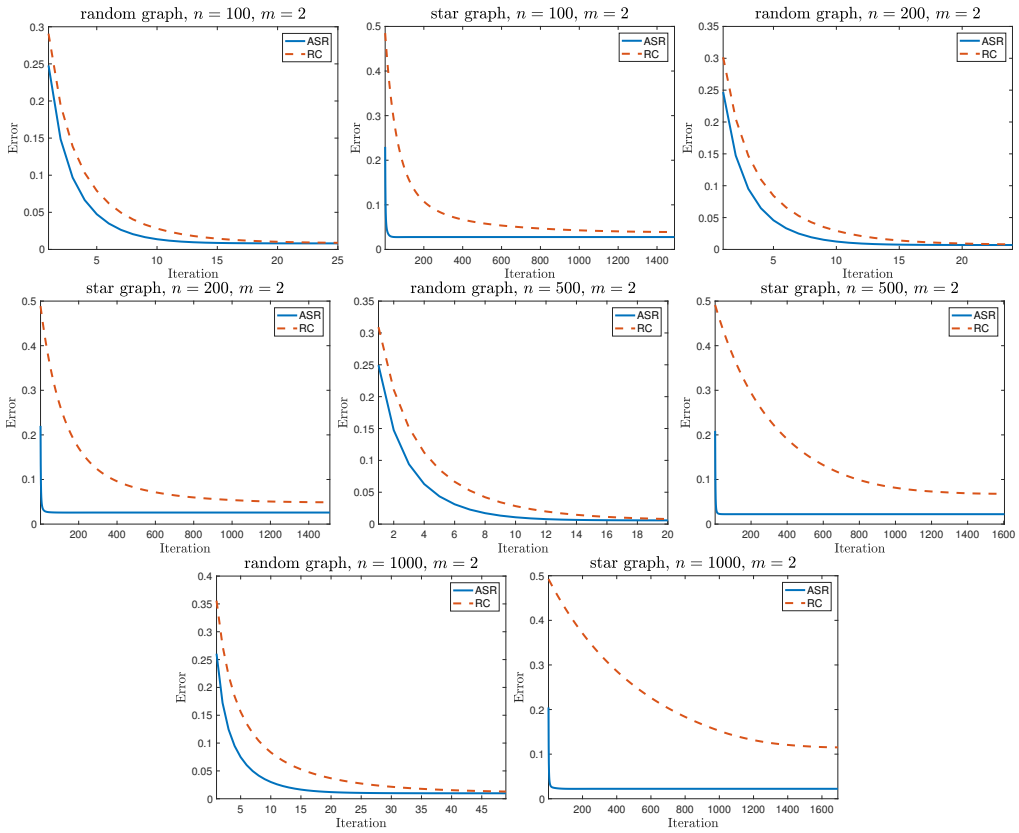

Figure 3. Results on synthetic data: L_1 error vs. number of iterations for our algorithm, ASR, compared with the RC algorithm (for $m = 2$) on data generated from the MNL/BTL model with the random and star graph topologies.

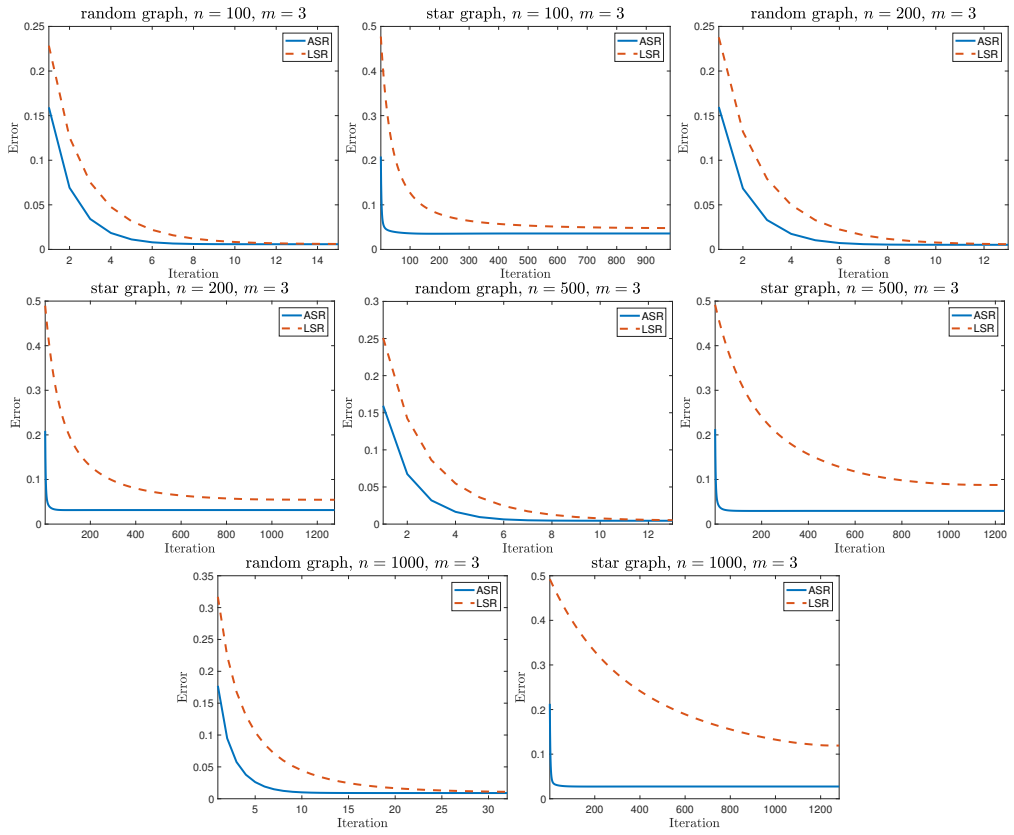

Figure 4. Results on synthetic data: L_1 error vs. number of iterations for our algorithm, ASR, compared with the LSR algorithm (for $m = 3$) on data generated from the MNL/BTL model with the random and star graph topologies.

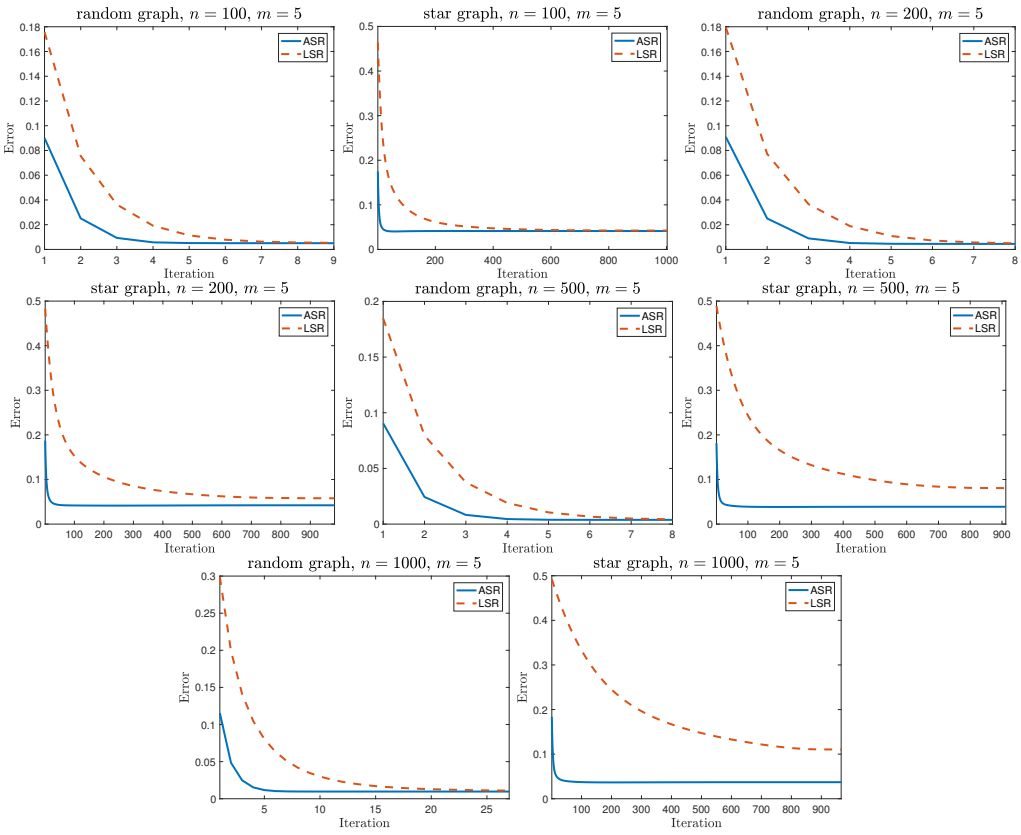

Figure 5. Results on synthetic data: L_1 error vs. number of iterations for our algorithm, ASR, compared with the LSR algorithm (for $m = 5$) on data generated from the MNL/BTL model with the random and star graph topologies.

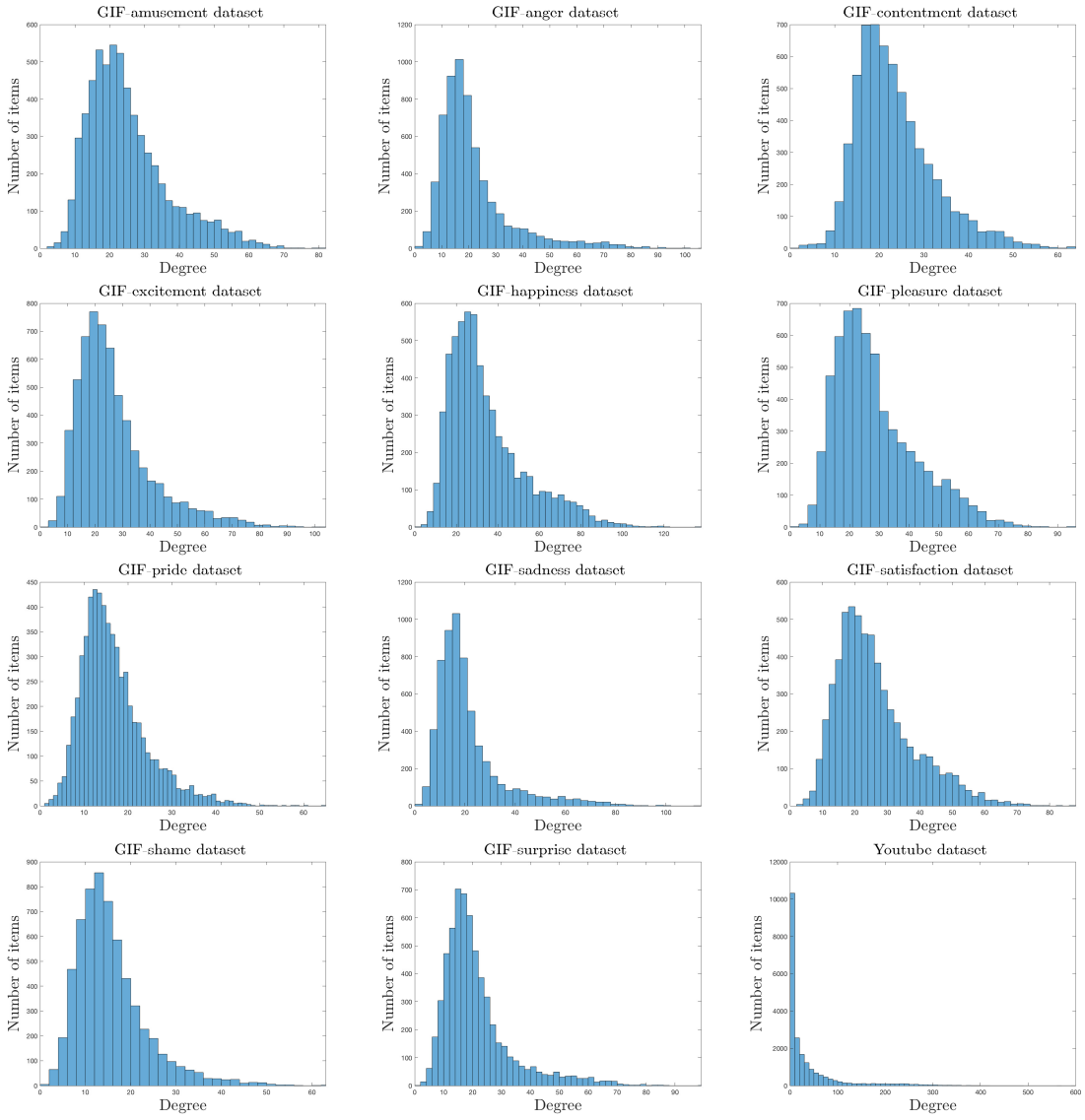

Figure 6. Degree distributions of various real world datasets.

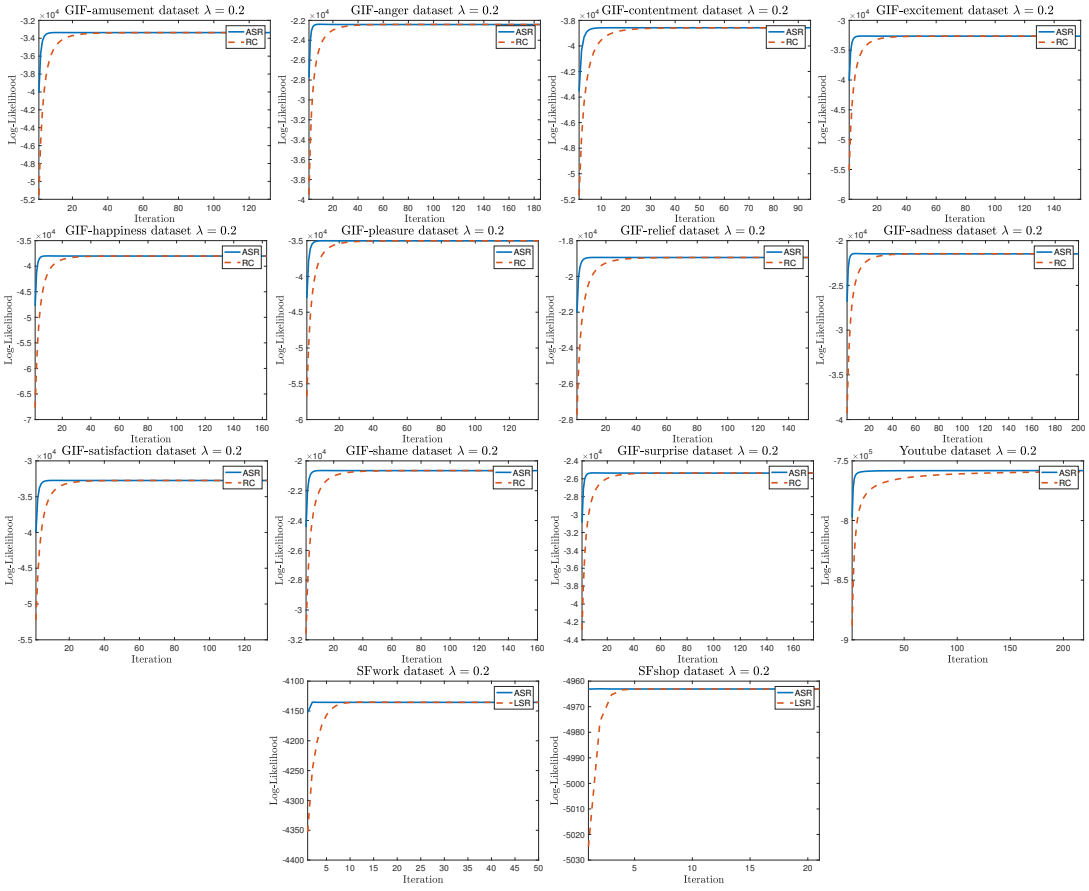

Figure 7. Results on real data: Log-likelihood vs. number of iterations for our algorithm, ASR, compared with the RC algorithm (for pairwise comparison data) and the LSR algorithm (for multi-way comparison data), all with regularization parameter set to 0.2.

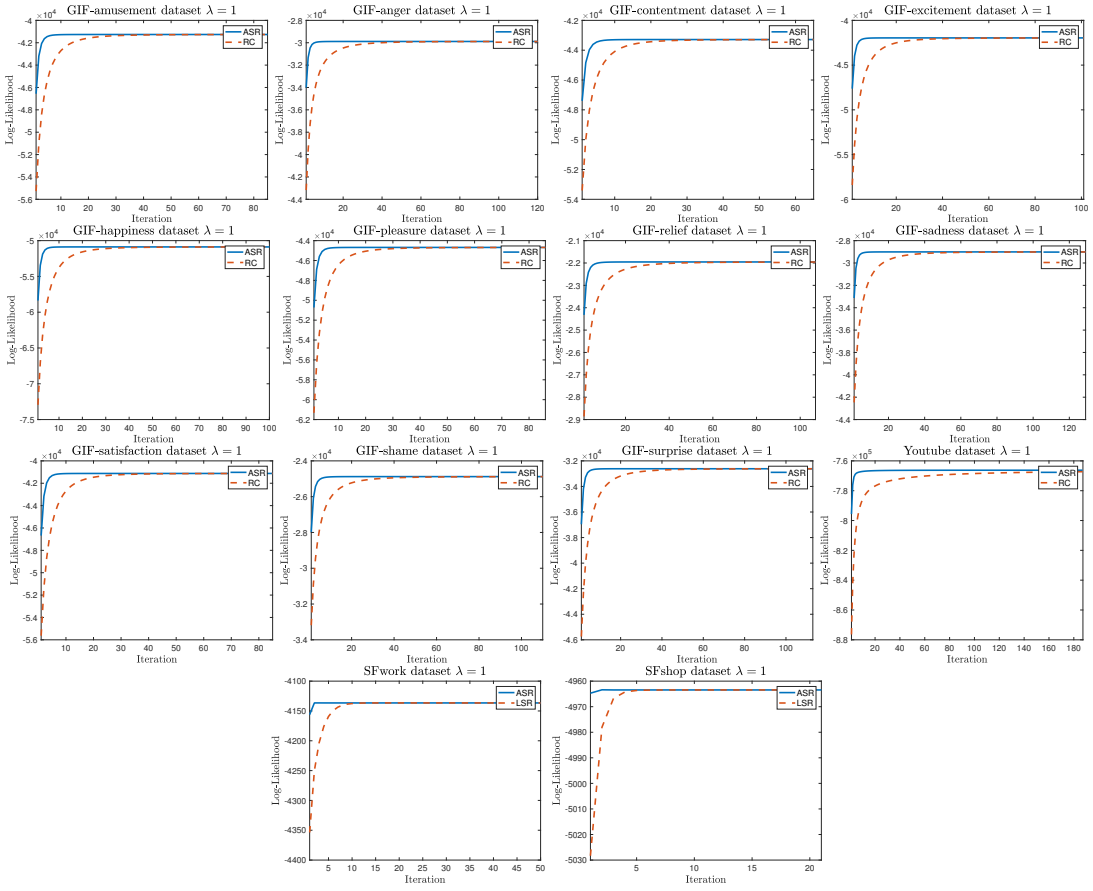

Figure 8. Results on real data: Log-likelihood vs. number of iterations for our algorithm, ASR, compared with the RC algorithm (for pairwise comparison data) and the LSR algorithm (for multi-way comparison data), all with regularization parameter set to 1.