SPECTRAL RANKING USING SERIATION

FAJWEL FOGEL, ALEXANDRE D'ASPREMONT, AND MILAN VOJNOVIC

ABSTRACT. We describe a seriation algorithm for ranking a set of items given pairwise comparisons between these items. Intuitively, the algorithm assigns similar rankings to items that compare similarly with all others. It does so by constructing a similarity matrix from pairwise comparisons, using seriation methods to reorder this matrix and construct a ranking. We first show that this spectral seriation algorithm recovers the true ranking when all pairwise comparisons are observed and consistent with a total order. We then show that ranking reconstruction is still exact when some pairwise comparisons are corrupted or missing, and that seriation based spectral ranking is more robust to noise than classical scoring methods. Finally, we bound the ranking error when only a random subset of the comparions are observed. An additional benefit of the seriation formulation is that it allows us to solve semi-supervised ranking problems. Experiments on both synthetic and real datasets demonstrate that seriation based spectral ranking achieves competitive and in some cases superior performance compared to classical ranking methods.

1. INTRODUCTION

We study the problem of ranking a set of n items given pairwise comparisons between these items. The setting we study here goes back at least to [Kendall and Smith, 1940] and seeks to reconstruct a ranking of items from pairwise comparisons reflecting a total ordering. In this case, the directed graph of all pairwise comparisons, where every pair of vertices is connected by exactly one of two possible directed edges, is usually called a *tournament* graph in the theoretical computer science literature or a "round robin" in sports, where every player plays every other player once and each preference marks victory or defeat. The motivation for this formulation often stems from the fact that in many applications, e.g. music, images, and movies, preferences are easier to express in relative terms (e.g. a is better than b) rather than absolute ones (e.g. a should be ranked fourth, and b seventh). In practice, the information about pairwise comparisons is usually *incomplete*, especially in the case of a large set of items, and the data may also be *noisy*, that is some pairwise comparisons could be incorrectly measured and inconsistent with a total order.

Ranking is a classical problem but its formulations vary widely. Website ranking methods such as PageRank [Page et al., 1998] and HITS [Kleinberg, 1999] seek to rank web pages based on the hyperlink structure of the web, where links do not necessarily express consistent preference relationships (e.g. a can link to b and b can link c, and c can link to a). Assumptions about how the pairwise preference information is obtained also vary widely. A subset of preferences is measured adaptively in [Ailon, 2011; Jamieson and Nowak, 2011], while [Negahban et al., 2012], for example, assume that preferences are observed iteratively, and [Freund et al., 2003] extract them at random. In other settings, the full preference matrix is observed, but is perturbed by noise: in e.g. [Bradley and Terry, 1952; Luce, 1959; Herbrich et al., 2006], a parametric model is assumed over the set of permutations, which reformulates ranking as a maximum likelihood problem.

Loss functions, performance metrics and algorithmic approaches vary as well. Kenyon-Mathieu and Schudy [2007], for example, derive a PTAS for the minimum feedback arc set problem on tournaments, i.e. the problem of finding a ranking that minimizes the number of upsets (a pair of players where the player ranked lower on the ranking beats the player ranked higher). In practice, the complexity of this method is

Date: June 25, 2015.

2010 *Mathematics Subject Classification.* 62F07, 06A07, 90C27.

Key words and phrases. Ranking, seriation, spectral methods.

relatively high, and other authors [see e.g. Keener, 1993; Negahban et al., 2012] have been using spectral methods to produce more efficient algorithms (each pairwise comparison is understood as a link pointing to the preferred item). In other cases, such as the classical Analytic Hierarchy Process (AHP) [Saaty, 1980; Barbeau, 1986] preference information is encoded in a "reciprocal" matrix whose Perron-Frobenius eigenvector provides the global ranking. Simple scoring methods such as the point difference rule [Huber, 1963; Wauthier et al., 2013] produce efficient estimates at very low computational cost. Ranking has also been approached as a prediction problem, i.e. learning to rank [Schapire and Singer, 1998], with [Joachims, 2002] for example using support vector machines to learn a score function. Finally, in the Bradley-Terry-Luce framework, where multiple observations on pairwise preferences are observed and assumed to be generated by a generalized linear model, the maximum likelihood problem is usually solved using fixed point algorithms or EM-like majorization-minimization techniques [Hunter, 2004].

Here, we show that the ranking problem is directly related to another classical ordering problem, namely *seriation*. Given a similarity matrix between a set of n items and assuming that the items can be ordered along a chain such that the similarity between items decreases with their distance within this chain (i.e. a total order exists), the seriation problem seeks to reconstruct the underlying linear ordering based on unsorted, possibly noisy, pairwise similarity information. Atkins et al. [1998] produced a spectral algorithm that exactly solves the seriation problem in the noiseless case, by showing that for similarity matrices computed from serial variables, the ordering of the eigenvector corresponding to the second smallest eigenvalue of the Laplacian matrix (a.k.a. the Fiedler vector) matches that of the variables. In practice, this means that performing spectral clustering on the similarity matrix exactly reconstructs the correct ordering provided items are organized in a chain.

We adapt these results to ranking to produce a very efficient *spectral ranking algorithm with provable recovery and robustness guarantees*. Furthermore, the seriation formulation allows us to handle semi-supervised ranking problems. Fogel et al. [2013] show that seriation is equivalent to the 2-SUM problem and study convex relaxations to seriation in a semi-supervised setting, where additional structural constraints are imposed on the solution. Several authors [Blum et al., 2000; Feige and Lee, 2007] have also focused on the directly related Minimum Linear Arrangement (MLA) problem, for which excellent approximation guarantees exist in the noisy case, albeit with very high polynomial complexity.

The main contributions of this paper can be summarized as follows. We link seriation and ranking by showing how to construct a consistent similarity matrix based on consistent pairwise comparisons. We then recover the true ranking by applying the spectral seriation algorithm in [Atkins et al., 1998] to this similarity matrix (we call this method SerialRank in what follows). In the noisy case, we then show that spectral seriation can perfectly recover the true ranking even when some of the pairwise comparisons are either corrupted or missing, provided that the pattern of errors is somewhat unstructured. We show in particular that, in a regime where a high proportion of comparisons are observed, some incorrectly, the spectral solution is more robust to noise than classical scoring based methods. On the other hand, when only few comparisons are observed, we show that for Erdös-Rényi graphs, *i.e.* when pairwise comparisons are observed independently with a given probability, $\Omega(n(\log n)^4)$ comparisons suffice for ℓ_2 consistency of the Fiedler vector w.h.p., while $\Omega(n^{3/2}(\log n)^4)$ comparisons suffice to retrieve a ranking whose relative ℓ_∞ distance to the true ranking goes to 0 with n, which requires ℓ_∞ consistency of the eigenvector. Since for Erdös-Rényi graphs the induced graph of comparisons is connected with high probability only when the total number of pairs sampled scales as $\Omega(n \log n)$ (aka the coupon collector effect), we need at least that many comparisons in order to retrieve a ranking, therefore the ℓ_2 consistency result can be seen as optimal up to a polylogarithmic factor. Finally, we use the seriation results in [Fogel et al., 2013] to produce semi-supervised ranking solutions.

The paper is organized as follows. In Section 2 we recall definitions related to seriation, and link ranking and seriation by showing how to construct well ordered similarity matrices from well ranked items. In Section 3 we apply the spectral algorithm of [Atkins et al., 1998] to reorder these similarity matrices and reconstruct the true ranking in the noiseless case. In Section 4 we then show that this spectral solution remains exact in a noisy regime where a random subset of comparisons is corrupted. In Section 5 we analyze ranking perturbation results when only few comparisons are given following an Erdös-Rényi graph. Finally, in Section 6 we illustrate our results on both synthetic and real datasets, and compare ranking performance with classical MLE, spectral and scoring based approaches.

2. Seriation, Similarities & Ranking

In this section we first introduce the seriation problem, i.e. reordering items based on pairwise similarities. We then show how to write the problem of ranking given pairwise comparisons as a seriation problem.

2.1. **The Seriation Problem.** The seriation problem seeks to reorder n items given a similarity matrix between these items, such that the more similar two items are, the closer they should be. This is equivalent to supposing that items can be placed on a chain where the similarity between two items decreases with the distance between these items in the chain. We formalize this below, following [Atkins et al., 1998].

Definition 2.1. *We say that a matrix $A \in \mathbf{S}_n$ is an R-matrix (or Robinson matrix) if and only if it is symmetric and $A_{i,j} \leq A_{i,j+1}$ and $A_{i+1,j} \leq A_{i,j}$ in the lower triangle, where $1 \leq j < i \leq n$.*

Another way to formulate R-matrix conditions is to impose $A_{ij} \leq A_{kl}$ if $|i-j| \leq |k-l|$ off-diagonal, i.e. the coefficients of A decrease as we move away from the diagonal. We also introduce a definition for strict R-matrices A, whose rows and columns cannot be permuted without breaking the R-matrix monotonicity conditions. We call *reverse identity* permutation the permutation that puts rows and columns $\{1,\ldots,n\}$ of a matrix A in reverse order $\{n, n-1, \ldots, 1\}$.

Definition 2.2. *An R-matrix $A \in \mathbf{S}_n$ is called strict-R if and only if the identity and reverse identity permutations of A are the only permutations reordering A as a R-matrix.*

Any R-matrix with only strict R-constraints is a strict R-matrix. Following [Atkins et al., 1998], we will say that A is *pre-R* if there is a permutation matrix Π such that $\Pi A \Pi^T$ is a R-matrix. Given a pre-R matrix A, the seriation problem consists in finding a permutation Π such that $\Pi A \Pi^T$ is a R-matrix. Note that there might be several solutions to this problem. In particular, if a permutation Π is a solution, then the reverse permutation is also a solution. When only two permutations of A produce R-matrices, A will be called *pre-strict-R*.

2.2. **Constructing Similarity Matrices from Pairwise Comparisons.** Given an ordered input pairwise comparison matrix, we now show how to construct a similarity matrix which is *strict-R* when all comparisons are given and consistent with the identity ranking (*i.e.* items are ranked in increasing order of indices). This means that the similarity between two items decreases with the distance between their ranks. We will then be able to use the spectral seriation algorithm by [Atkins et al., 1998] described in Section 3 to reconstruct the true ranking from a disordered similarity matrix.

We first show how to compute a pairwise similarity from binary comparisons between items by counting the number of matching comparisons. Another formulation allows to handle the generalized linear model. These two examples are only two particular instances of a broader class of ranking algorithms derived here. Any method which produces R-matrices from pairwise preferences yields a valid ranking algorithm.

2.2.1. *Similarities from Pairwise Comparisons.* Suppose we are given a matrix of pairwise comparisons $C \in \{-1,0,1\}^{n\times n}$ such that $C_{i,j} = -C_{j,i}$ for every $i \neq j$ and

$$C_{i,j} = \begin{cases} 1 & \text{if } i \text{ is ranked higher than } j \\ 0 & \text{if } i \text{ and } j \text{ are not compared or in a draw} \\ -1 & \text{if } j \text{ is ranked higher than } i \end{cases} \tag{1}$$

setting $C_{i,i} = 1$ for all $i \in \{1, \ldots, n\}$. We define the pairwise similarity matrix S^{match} as

$$S^{\text{match}}_{i,j} = \sum_{k=1}^{n} \left(\frac{1 + C_{i,k} C_{j,k}}{2} \right). \tag{2}$$

Since $C_{i,k} C_{j,k} = 1$, if $C_{i,k}$ and $C_{j,k}$ have matching signs, and $C_{i,k} C_{j,k} = -1$ if they have opposite signs, $S^{\text{match}}_{i,j}$ counts the number of matching comparisons between i and j with other reference items k. If i or j is not compared with k, then $C_{i,k} C_{j,k} = 0$ and the term $(1 + C_{i,k} C_{j,k})/2$ has an average effect on the similarity of $1/2$. Note that we also have

$$S^{\text{match}} = \frac{1}{2} \left(n\mathbf{1}\mathbf{1}^T + CC^T \right). \tag{3}$$

The intuition behind the similarity S^{match} is easy to understand in a tournament setting: players that beat the same players and are beaten by the same players should have a similar ranking.

The next result shows that when all comparisons are given and consistent with the identity ranking, then the similarity matrix S^{match} is a strict R-matrix. Without loss of generality, we assume that items are ranked in increasing order of their indices. In the general case, we can simply replace the *strict-R* property by the *pre-strict-R* property.

Proposition 2.3. *Given all pairwise comparisons $C_{i,j} \in \{-1, 0, 1\}$ between items ranked according to the identity permutation (with no ties), the similarity matrix S^{match} constructed in (2) is a strict R-matrix and*

$$S^{\text{match}}_{i,j} = n - |i - j| \tag{4}$$

for all $i, j = 1, \ldots, n$.

Proof. Since items are ranked as $\{1, \ldots, n\}$ with no ties and all comparisons given, $C_{i,j} = -1$ if $i < j$ and $C_{i,j} = 1$ otherwise. Therefore we get from definition (2)

$$\begin{aligned} S^{\text{match}}_{i,j} &= \sum_{k=1}^{\min(i,j)-1} \left(\frac{1+1}{2} \right) + \sum_{k=\min(i,j)}^{\max(i,j)-1} \left(\frac{1-1}{2} \right) + \sum_{k=\max(i,j)}^{n} \left(\frac{1+1}{2} \right) \\ &= n - (\max\{i,j\} - \min\{i,j\}) \\ &= n - |i - j| \end{aligned}$$

This means in particular that S^{match} is strictly positive and its coefficients are strictly decreasing when moving away from the diagonal, hence S^{match} is a strict R-matrix. ■

2.2.2. *Similarities in the Generalized Linear Model.* Suppose that paired comparisons are generated according to a generalized linear model (GLM), *i.e.* we assume that the outcomes of paired comparisons are independent and for any pair of distinct items, item i is observed ranked higher than item j with probability

$$P_{i,j} = H(\nu_i - \nu_j) \tag{5}$$

where $\nu \in \mathbb{R}^n$ is a vector of skill parameters and $H : \mathbb{R} \to [0, 1]$ is a function that is increasing on $\mathbb{R}$ and such that $H(-x) = 1 - H(x)$ for all $x \in \mathbb{R}$, and $\lim_{x \to -\infty} H(x) = 0$ and $\lim_{x \to \infty} H(x) = 1$. A well known special instance of the generalized linear model is the Bradley-Terry-Luce model for which $H(x) = 1/(1 + e^{-x})$, for $x \in \mathbb{R}$.

Let $m_{i,j}$ be the number of times items i and j were compared, $C^s_{i,j} \in \{-1, 1\}$ be the outcome of comparison s and Q be the matrix of corresponding sample probabilities, i.e. if $m_{i,j} > 0$ we have

$$Q_{i,j} = \frac{1}{m_{i,j}} \sum_{s=1}^{m_{i,j}} \frac{C^s_{i,j} + 1}{2}$$

and $Q_{i,j} = 1/2$ in case $m_{i,j} = 0$. We define the similarity matrix S^{glm} from the observations Q as

$$S^{\text{glm}}_{i,j} = \sum_{k=1}^{n} \mathbf{1}_{\{m_{i,k} m_{j,k} > 0\}} \left(1 - \frac{|Q_{i,k} - Q_{j,k}|}{2}\right) + \frac{\mathbf{1}_{\{m_{i,k} m_{j,k} = 0\}}}{2}. \tag{6}$$

Since the comparison observations are independent we have that $Q_{i,j}$ converges to $P_{i,j}$ as $m_{i,j}$ goes to infinity and the CLT means that $S^{\text{glm}}_{i,j}$ converges to a Gaussian variable with mean

$$\sum_{k=1}^{n} \left(1 - \frac{|P_{i,k} - P_{j,k}|}{2}\right).$$

The result below shows that this limit similarity matrix is a strict R-matrix when the variables are properly ordered.

Proposition 2.4. *If the items are ordered according to the order in decreasing values of the skill parameters, the similarity matrix S^{glm} is a strict R matrix with high probability as n goes to infinity.*

Proof. Without loss of generality, we suppose the true order is $\{1, \dots, n\}$, with $\nu(1) > \dots > \nu(n)$. For any i, j, k such that $i > j$, using the GLM assumption (i) we get

$$P_{i,k} = H(\nu(i) - \nu(k)) > H(\nu(j) - \nu(k)) = p_{j,k}.$$

Since empirical probabilities Q_{ij} converge to p_{ij}, when the number of observations is large enough, we also get $Q_{i,k} > Q_{j,k}$ for any i, j, k such that $i > j \geq k$ (we focus wlog on the lower triangle), and we can therefore remove the absolute value in the expression of S^{glm}_{ij} for $i > j$. Hence for any $i > j$ we have

$$\begin{aligned} S^{\text{glm}}_{i+1,j} - S^{\text{glm}}_{i,j} &= \frac{1}{2}\left(-\sum_{k=1}^{n} |Q_{i+1,k} - Q_{j,k}| + \sum_{k=1}^{n} |Q_{i,k} - Q_{j,k}|\right) \\ &= \frac{1}{2}\left(\sum_{k=1}^{n} -(Q_{i+1,k} - Q_{j,k}) + (Q_{i,k} - Q_{j,k})\right) \\ &= \frac{1}{2}\left(\sum_{k=1}^{n} Q_{i,k} - Q_{i+1,k}\right) < 0. \end{aligned}$$

Similarly for any $i > j$, $S^{\text{glm}}_{i,j-1} - S^{\text{glm}}_{i,j} < 0$, so S^{glm} is a strict R-matrix. ■

Notice that we recover the original definition of S^{match} in the case of binary probabilities, though it does not fit in the Generalized Linear Model. Note also that these definitions can be directly extended to the setting where multiple comparisons are available for each pair and aggregated in comparisons that take fractional values (*e.g.* a tournament setting where participants play several times against each other).

3. Spectral Algorithms

We first recall how spectral clustering can be used to recover the true ordering in seriation problems. We then apply this method to the ranking problem.

3.1. **Spectral Seriation Algorithm.**

We use the spectral computation method originally introduced in [Atkins et al., 1998] to solve the seriation problem based on the similarity matrices defined in the previous section. We first recall the definition of the Fiedler vector.

Definition 3.1. *The Fiedler value of a symmetric, nonnegative and irreducible matrix A is the smallest non-zero eigenvalue of its Laplacian matrix $L_A = \mathbf{diag}(A\mathbf{1}) - A$. The corresponding eigenvector is called Fiedler vector and is the optimal solution to* $\min\{y^T L_A y : y \in \mathbb{R}^n, y^T \mathbf{1} = 0, \|y\|_2 = 1\}$.

The main result from [Atkins et al., 1998], detailed below, shows how to reorder pre-R matrices in a noise free case.

Proposition 3.2. *[Atkins et al., 1998, Th. 3.3] Let $A \in \mathbf{S}_n$ be an irreducible pre-R-matrix with a simple Fiedler value and a Fiedler vector v with no repeated values. Let $\Pi_1 \in \mathcal{P}$ (respectively, Π_2) be the permutation such that the permuted Fiedler vector $\Pi_1 v$ is strictly increasing (decreasing). Then $\Pi_1 A \Pi_1^T$ and $\Pi_2 A \Pi_2^T$ are R-matrices, and no other permutations of A produce R-matrices.*

The next technical lemmas extend the results in Atkins et al. [1998] to strict R-matrices and will be used to prove proposition 3.6 in next section. The first one shows that without loss of generality, the Fiedler value is simple.

Lemma 3.3. *If A is an irreducible R-matrix, up to a uniform shift of its coefficients, A has a simple Fiedler value and a monotonic Fiedler vector.*

Proof. We use [Atkins et al., 1998, Th. 4.6] which states that if A is an irreducible R-matrix with $A_{n,1} = 0$, then the Fiedler value of A is a simple eigenvalue. Since A is a R-matrix, $A_{n,1}$ is among its minimal elements. Subtracting it from A does not affect the positivity of A and we can apply [Atkins et al., 1998, Th. 4.6]. Monotonicity of the Fiedler vector then follows from [Atkins et al., 1998, Th. 3.2]. ■

The next lemma shows that the Fiedler vector is strictly monotonic if A is a strict R-matrix.

Lemma 3.4. *Let $A \in \mathbf{S}_n$ be a R-matrix. Suppose there are no distinct indices $r < s$ such that for any $k \notin [r; s]$, $A_{r,k} = A_{r+1,k} = \ldots = A_{s,k}$, then, up to a uniform shift, the Fiedler value of A is simple and its Fiedler vector is strictly monotonic.*

Proof. By Lemma 3.3, the Fiedler value of A is simple (up to a uniform shift of A). Let x be the corresponding Fiedler vector of A, x is monotonic by Lemma 3.3. Suppose $[r; s]$ is a nontrivial maximal interval such that $x_r = x_{r+1} = \ldots = x_s$, then by [Atkins et al., 1998, lemma 4.3], for any $k \notin [r; s]$, $A_{r,k} = A_{r+1,k} = \ldots = A_{s,k}$, which contradicts the initial assumption. Therefore x is strictly monotonic. ■

In fact, we only a small portion of the R-constraints to be strict for the previous lemma to hold. We now show that the main assumption on A in Lemma 3.4 is equivalent to A being strict-R.

Lemma 3.5. *An R-matrix $A \in \mathbf{S}_n$ is strictly R if and only if there are no distinct indices $r < s$ such that for any $k \notin [r; s]$, $A_{r,k} = A_{r+1,k} = \ldots = A_{s,k}$.*

Proof. Let $A \in \mathbf{S}_n$ a R-matrix. Let us first suppose there are no distinct indices $r < s$ such that for any $k \notin [r; s]$, $A_{r,k} = A_{r+1,k} = \ldots = A_{s,k}$. By lemma 3.4 the Fiedler value of A is simple and its Fiedler vector is strictly monotonic. Hence by proposition 3.2, only the identity and reverse identity permutations of A produce R-matrices. Now suppose there exist two distinct indices $r < s$ such that for any $k \notin [r; s]$, $A_{r,k} = A_{r+1,k} = \ldots = A_{s,k}$. In addition to the identity and reverse identity permutations, we can locally reverse the order of rows and columns from r to s, since the sub matrix $A_{r:s,r:s}$ is an R-matrix and for any $k \notin [r; s]$, $A_{r,k} = A_{r+1,k} = \ldots = A_{s,k}$. Therefore at least four different permutations of A produce R-matrices, which means that A is not strictly R. ■

3.2. **SerialRank: a Spectral Ranking Algorithm.** In Section 2, we showed that similarities S^{match} and S^{glm} are *pre-strict-R* when all comparisons are available and consistent with an underlying ranking of items. We now use the spectral seriation method in [Atkins et al., 1998] to reorder these matrices and produce a ranking. Spectral clustering requires computing an extremal eigenvector, at a cost of $O(n^2 \log n)$ flops [Kuczynski and Wozniakowski, 1992]. We call this algorithm SerialRank and prove the following result.

Proposition 3.6. *Given all pairwise comparisons for a set of totally ordered items and assuming there are no ties between items, algorithm SerialRank, i.e. sorting the Fiedler vector of the matrix S^{match} defined in (3), recovers the true ranking of items.*

Algorithm 1 (SerialRank)

```
Input: A set of pairwise comparisons $C_{i,j} \in \{-1, 0, 1\}$ or $[-1, 1]$.
Compute a similarity matrix $S$ as in §2.2
Compute the Laplacian matrix
    $L_S = \mathbf{diag}(S\mathbf{1}) - S$    (SerialRank)
Compute the Fiedler vector of $S$.
Output: A ranking induced by sorting the Fiedler vector of $S$ (choose either increasing or decreasing order
    to minimize the number of upsets).
```

Proof. From Proposition 2.3 we get that, under our assumptions, S^{match} is a pre-strict R-matrix. Now combining the definition of strict-R matrices in Lemma 3.5 with Lemma 3.4, we deduce that Fiedler value of S^{match} is simple and its Fiedler vector has no repeated values. Hence by Theorem 3.2, only the two permutations that sort the Fiedler vector in increasing and decreasing order produce strict R-matrices and are therefore candidate rankings (by Proposition 2.3 S^{match} is a strictly R-matrix when ordered according to the true ranking). Finally we can choose between the two candidate rankings (increasing and decreasing) by picking the one with the least upsets. ■

Similar results apply for S^{glm} given enough comparisons in the Generalized Linear Model. This last result guarantees recovery of the true ranking of items in the noiseless case. In the next section, we will study the impact of corrupted or missing comparisons on the inferred ranking of items.

3.3. **Hierarchical Ranking.** In a large dataset, the goal may be to rank only a subset of top items. In this case, we can first perform spectral ranking, then refine the ranking of the top set of items using either the SerialRank algorithm on the top comparison submatrix, or another seriation algorithm such as the convex relaxation in [Fogel et al., 2013]. This last method also allows us to solve semi-supervised ranking problems, given additional information on the structure of the solution.

4. Exact Recovery with Corrupted and Missing Comparisons

In this section we study the robustness of SerialRank using S^{match} with respect to noisy and missing pairwise comparisons. We will see that noisy comparisons cause ranking ambiguities for the point score method and that such ambiguities are be lifted by the spectral ranking algorithm. We show in particular that the SerialRank algorithm recovers the exact ranking when the pattern of errors is random and errors are not

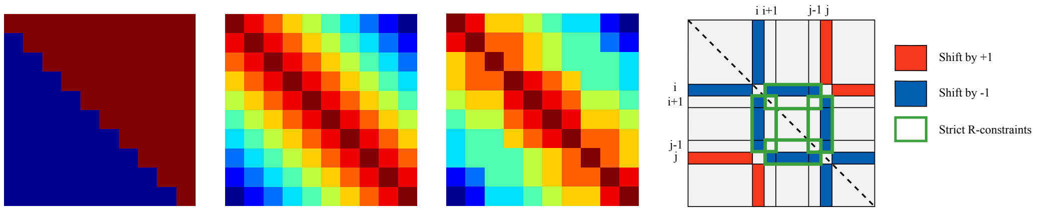

Figure 1. The matrix of pairwise comparisons C *(far left)* when the rows are ordered according to the true ranking. The corresponding similarity matrix S^{match} is a strict R-matrix *(center left)*. The same S^{match} similarity matrix with comparison (3,8) corrupted *(center right)*. With one corrupted comparison, S^{match} keeps enough strict R-constraints to recover the right permutation. In the noiseless case, the difference between all coefficients is at least one and after introducing an error, the coefficients inside the green rectangles still enforce strict R-constraints *(far right)*.

too numerous. We first study the impact of one corrupted comparison on SerialRank, then extend the result to multiple corrupted comparisons. A similar analysis is provided for missing comparisons as Corollary 7.3. in the Appendix. Finally, Proposition 4.4 provides an estimate of the number of randomly corrupted entries that can be tolerated for perfect recovery of the true ranking. We begin by recalling the definition of the *point score* of an item.

Definition 4.1. *The* point score w_i *of an item* i*, also known as point-difference, or* row-sum *is defined as* $w_i = \sum_{k=1}^n C_{k,i}$*, which corresponds to the number of wins minus the number of losses in a tournament setting.*

In the following we will denote by w the score vector.

Proposition 4.2. *Given all pairwise comparisons* $C_{s,t} \in \{-1, 1\}$ *between items ranked according to their indices, suppose the sign of one comparison* $C_{i,j}$ *is switched, with* $i < j$*. If* $j - i > 2$ *then* S^{match} *defined in* (3) *remains* strict-R*, whereas the score vector* w *has ties between items* i *and* $i+1$ *and items* j *and* $j-1$*.*

Proof. We give some intuition for the result in Figure 1. We write the true score and comparison matrix w and C, while the observations are written $\hat{w}$ and $\hat{C}$ respectively. This means in particular that $\hat{C}_{i,j} = -C_{i,j} = 1$. To simplify notations we denote by S the similarity matrix S^{match} (respectively $\hat{S}$ when the similarity is computed from observations). We first study the impact of a corrupted comparison $C_{i,j}$ for $i < j$ on the score vector $\hat{w}$. We have

$$\hat{w}_i = \sum_{k=1}^n \hat{C}_{k,i} = \sum_{k=1}^n C_{k,i} + \hat{C}_{j,i} - C_{j,i} = w_i + 2 = w_{i+1},$$

similarly $\hat{w}_j = w_{j-1}$, whereas for $k \neq i, j$, $\hat{w}_k = w_k$. Hence, the incorrect comparison induces two ties in the score vector w. Now we show that the similarity matrix defined in (3) breaks these ties, by showing that it is a strict R-matrix. Writing $\hat{S}$ in terms of S, we get

$$[\hat{C}\hat{C}^T]_{i,t} = \sum_{k \neq j} \left(\hat{C}_{i,k}\hat{C}_{t,k} \right) + \hat{C}_{i,j}\hat{C}_{t,j} = \sum_{k \neq j} (C_{i,k}C_{t,k}) + \hat{C}_{i,j}C_{t,j} = \begin{cases} [CC^T]_{i,t} - 2 & \text{if } t < j \\ \left[CC^T\right]_{i,t} + 2 & \text{if } t > j. \end{cases}$$

We thus get

$$\hat{S}_{i,t} = \begin{cases} S_{i,t} - 1 & \text{if } t < j \\ S_{i,t} + 1 & \text{if } t > j, \end{cases}$$

(remember there is a factor $1/2$ in the definition of S). Similarly we get for any $t \neq i$

$$\hat{S}_{j,t} = \begin{cases} S_{j,t} + 1 & \text{if } t < i \\ S_{j,t} - 1 & \text{if } t > i. \end{cases}$$

Finally, for the single corrupted index pair (i, j), we get

$$\hat{S}_{i,j} = \frac{1}{2} \left(n + \sum_{k \neq i,j} \left(\hat{C}_{i,k}\hat{C}_{j,k} \right) + \hat{C}_{i,i}\hat{C}_{j,i} + \hat{C}_{i,j}\hat{C}_{j,j} \right) = S_{i,j} - 1 + 1 = S_{i,j}.$$

For all other coefficients (s, t) such that $s, t \neq i, j$, we have $\hat{S}_{s,t} = S_{s,t}$. Meaning all rows or columns outside of i, j are left unchanged. We first observe that these last equations, together with our assumption that $j - i > 2$ and the fact that the elements of the exact S in (4) differ by at least one, mean that

$$\hat{S}_{s,t} \geq \hat{S}_{s+1,t} \quad \text{and} \quad \hat{S}_{s,t+1} \geq \hat{S}_{s,t}, \quad \text{for any } s < t$$

so $\hat{S}$ remains an R-matrix. Note that this result remains true even when $j - i = 2$, but we need some strict inequalities to show uniqueness of the retrieved order. Indeed, because $j - i > 2$ all these R constraints are strict except between elements of rows i and $i+1$, and rows $j-1$ and j (idem for columns). These ties can be broken using the fact that

$$\hat{S}_{i,j-1} = S_{i,j-1} - 1 < S_{i+1,j-1} - 1 = \hat{S}_{i+1,j-1} - 1 < \hat{S}_{i+1,j-1}$$

which means that $\hat{S}$ is still a strict R-matrix (see Figure 1) since $j-1 > i+1$ by assumption. ■

We now extend this result to multiple errors.

Proposition 4.3. *Given all pairwise comparisons $C_{s,t} \in \{-1, 1\}$ between items ranked according to their indices, suppose the signs of m comparisons indexed $(i_1, j_1), \dots, (i_m, j_m)$ are switched. If the following condition (7) holds true,*

$$|s-t| > 2, \textit{ for all } s, t \in \{i_1, \dots, i_m, j_1, \dots, j_m\} \textit{ with } s \neq t, \tag{7}$$

then S^{match} defined in (3) remains strict-R*, whereas the score vector w gets $2m$ ties.*

Proof. We write the true score and comparison matrix w and C, while the observations are written $\hat{w}$ and $\hat{C}$ respectively, and without loss of generality we suppose $i_l < j_l$. This means in particular that $\hat{C}_{i_l,j_l} = -C_{i_l,j_l} = 1$ for all l in $\{1, \dots, m\}$. To simplify notations we denote by S the similarity matrix S^{match} (respectively $\hat{S}$ when the similarity is computed from observations).

As in the proof of proposition 4.2, corrupted comparisons indexed (i_l, j_l) induce shifts of ±1 on columns and rows i_l and j_l of the similarity matrix S^{match}, while $S^{\text{match}}_{i_l,j_l}$ values remain the same. Since there are several corrupted comparisons, we also need to check the values of $\hat{S}$ at the intersections of rows and columns with indices of corrupted comparisons. Formally, for any $(i,j) \in \{(i_1, j_1), \dots (i_m, j_m)\}$ and $t \notin \{i_1, \dots, i_m, j_1, \dots, j_m\}$

$$\hat{S}_{i,t} = \begin{cases} S_{i,t} + 1 & \text{if } t < j \\ S_{i,t} - 1 & \text{if } t > j, \end{cases}$$

Similarly for any $t \notin \{i_1, \dots, i_m, j_1, \dots, j_m\}$

$$\hat{S}_{j,t} = \begin{cases} S_{j,t} - 1 & \text{if } t < i \\ S_{j,t} + 1 & \text{if } t > i. \end{cases}$$

Let (s, s') and $(t, t') \in \{(i_1, j_1), \dots (i_m, j_m)\}$, we have

$$\begin{aligned} \hat{S}_{s,t} &= \tfrac{1}{2}\left(n + \textstyle\sum_{k \neq s',t'} \left(\hat{C}_{s,k}\hat{C}_{t,k}\right) + \hat{C}_{s,s'}\hat{C}_{t,s'} + \hat{C}_{s,t'}\hat{C}_{t,t'}\right) \\ &= \tfrac{1}{2}\left(n + \textstyle\sum_{k \neq s',t'} \left(C_{s,k}C_{t,k}\right) - C_{s,s'}C_{t,s'} - C_{s,t'}C_{t,t'}\right) \end{aligned}$$

Without loss of generality we suppose $s < t$, and since $s < s'$ and $t < t'$, we get

$$\hat{S}_{s,t} = \begin{cases} S_{s,t} & \text{if } t > s' \\ S_{s,t} + 2 & \text{if } t < s'. \end{cases}$$

Similar results apply for other intersections of rows and columns with indices of corrupted comparisons (*i.e.* shifts of 0, +2, or −2). For all other coefficients (s,t) such that $s, t \notin \{i_1, \dots, i_m, j_1, \dots, j_m\}$, we have $\hat{S}_{s,t} = S_{s,t}$. We first observe that these last equations, together with our assumption that $j_l - i_l > 2$, mean that

$$\hat{S}_{s,t} \geq \hat{S}_{s+1,t} \quad \text{and} \quad \hat{S}_{s,t+1} \geq \hat{S}_{s,t}, \qquad \text{for any } s < t$$

so $\hat{S}$ remains an R-matrix. Moreover, since $j_l - i_l > 2$ all these R constraints are strict except between elements of rows i_l and $i_l + 1$, and rows $j_l - 1$ and j_l (similar for columns). These ties can be broken using the fact that for $k = j_l - 1$

$$\hat{S}_{i_l,k} = S_{i_l,k} - 1 < S_{i_l+1,k} - 1 = \hat{S}_{i_l+1,k} - 1 < \hat{S}_{i_l+1,k}$$

which means that $\hat{S}$ is still a strict R-matrix since $k = j_l - 1 > i_l + 1$. Moreover, using the same argument as in the proof of proposition 4.2, corrupted comparisons induces $2m$ ties in the score vector w. ■

For the case of one corrupted comparison, note that the separation condition on the pair of items (i,j) is necessary. When the comparison $C_{i,j}$ between two adjacent items is corrupted, no ranking method can break the resulting tie. For the case of arbitrary number of corrupted comparisons, condition (7) is a sufficient condition only. We study exact ranking recovery conditions with missing comparisons in the Appendix, using similar arguments. We now estimate the number of randomly corrupted entries that can be tolerated while maintaining exact recovery of the true ranking.

Proposition 4.4. *Given a comparison matrix for a set of n items with m corrupted comparisons selected uniformly at random from the set of all possible item pairs. Algorithm SerialRank guarantees that the probability of recovery $p(n,m)$ satisfies $p(n,m) \geq 1-\delta$, provided that $m = O(\sqrt{\delta n})$. In particular, this implies that $p(n,m) = 1 - o(1)$ provided that $m = o(\sqrt{n})$.*

Proof. Let $\mathcal{P}$ be the set of all distinct pairs of items from the set $\{1,2,\ldots,n\}$. Let $\mathcal{X}$ be the set of all admissible sets of pairs of items, i.e. containing each $X \subseteq \mathcal{P}$ such that X satisfies condition (7). We consider the case of $m \geq 1$ distinct pairs of items sampled from the set $\mathcal{P}$ uniformly at random without replacement. Let X_i denote the set of sampled pairs given that i pairs are sampled. We seek to bound $p(n,m) = \mathbf{Prob}(X_m \in \mathcal{X})$. Given a set of pairs $X \in \mathcal{X}$, let $T(X)$ be the set of non-admissible pairs, i.e. containing $(i,j) \in \mathcal{P} \setminus X$ such that $X \cup (i,j) \notin \mathcal{X}$. We have

$$\mathbf{Prob}(X_m \in \mathcal{X}) = \sum_{x \in \mathcal{X} : |x| = m-1} \left(1 - \frac{|T(x)|}{|\mathcal{P}| - (m-1)}\right) \mathbf{Prob}(X_{m-1} = x). \tag{8}$$

Note that every selected pair from $\mathcal{P}$ contributes at most $6n-10$ non-admissible pairs, hence, for every $x \in \mathcal{X}$ we have

$$|T(x)| \leq 2(3n-5)|x|.$$

Combined with (8) and the fact $|\mathcal{P}| = \binom{n}{2}$, we have

$$\mathbf{Prob}(X_m \in \mathcal{X}) \geq \left(1 - \frac{2(3n-5)}{\binom{n}{2} - (m-1)}(m-1)\right) \mathbf{Prob}(X_{m-1} \in \mathcal{X}).$$

From this it follows

$$\begin{aligned} p(n,m) &\geq \prod_{i=1}^{m-1} \left(1 - \frac{2(3n-5)}{\binom{n}{2} - (i-1)} i\right) \\ &\geq \prod_{i=1}^{m-1} \left(1 - \frac{i}{a(n,m)}\right) \end{aligned}$$

where

$$a(n,m) = \frac{\binom{n}{2} - (m-1)}{2(3n-5)}.$$

Notice that for $m = o(n)$ we have

$$\prod_{i=1}^{m-1} \left(1 - \frac{i}{a(n,m)}\right) \sim \exp\left(-6\frac{m^2}{n}\right) \text{ for large } n.$$

Hence, given $\delta > 0$, $p(n,m) \geq 1-\delta$ provided that $m = O(\sqrt{n\delta})$. If $\delta = o(1)$, the condition is $m = o(\sqrt{n})$. ■

5. Spectral Perturbation Analysis

In this section we analyze how SerialRank performs when only a small fraction of pairwise comparisons are given. We show that for Erdös-Rényi graphs, *i.e.* when pairwise comparisons are observed independently with a given probability, $\Omega(n(\log n)^4)$ comparisons suffice for ℓ_2 consistency of the Fiedler vector. On the other hand we need $\Omega(n^{3/2}(\log n)^4)$ comparisons to retrieve a ranking whose relative ℓ_∞ distance to the true ranking goes to zero when n grows, which requires ℓ_∞ consistency of the eigenvector. Since Erdös-Rényi graphs are connected with high probability only when the total number of pairs sampled scales as $\Omega(n \log n)$, we need at least that many comparisons in order to retrieve a ranking, therefore the ℓ_2 consistency result can be seen as optimal up to a polylogarithmic factor.

Throughout this section, we only focus on the similarity S^{match} in (3) and write it S to simplify notations. We refer to the eigenvalues of the Laplacian as λ_i, with $\lambda_1 = 0 \leq \lambda_2 \leq \ldots \leq \lambda_n$. For any quantity x, we denote by $\tilde{x}$ its perturbed analogue. We define $R = \tilde{S} - S$ and write f the Fiedler vector of the Laplacian matrix L_S. We denote by $D_S = \mathbf{diag}(D\mathbf{1})$ the degree of matrix S. Finally, we will use $c > 0$ for absolute constants, whose values are allowed to vary from one equation to another.

We assume that our information on preferences is both incomplete and corrupted. Specifically, pairwise comparisons are independently sampled with probability q and these sampled comparisons are consistent with the underlying total ranking with probability p. Let us define $\tilde{C} = B \circ C$ the matrix of observed comparisons, where C is the true comparison matrix defined in (1), B is a symmetric matrix with entries

$$B_{i,j} = \begin{cases} 0 & \text{with probability } 1-q \\ 1 & \text{with probability } qp \\ -1 & \text{with probability } q(1-p). \end{cases}$$

In order to obtain an unbiased estimator of the similarity matrix defined in (3), we normalize $\tilde{C}$ by its mean value $q(2p-1)$, *i.e.* redefine $\tilde{S}$ as

$$\tilde{S} = \frac{1}{q^2(2p-1)^2}\tilde{C}\tilde{C}^T + n\mathbf{1}\mathbf{1}^T.$$

For ease of read we have dropped the factor $1/2$ in (3) w.l.o.g. since positive multiplicative factors of the Laplacian do not affect its eigenvectors.

We now state our main results. The first one bounds ℓ_2 perturbations of the Fiedler vector with both missing and corrupted comparisons.

Proposition 5.1. *For every $\mu \in (0,1)$ and n large enough, if $q > \frac{(\log n)^4}{\mu^2(2p-1)^4 n}$, then*

$$\|\tilde{f} - f\|_2 \leq c\frac{\mu}{\sqrt{\log n}}$$

with probability at least $1 - 2/n$, where $c > 0$ is an absolute constant.

The second result bounds local perturbations of the ranking with π referring to the "true" ranking and $\tilde{\pi}$ to the ranking retrieved by SerialRank.

Proposition 5.2. *For every $\mu \in (0,1)$ and n large enough, if $q > \frac{(\log n)^4}{\mu^2(2p-1)^4\sqrt{n}}$, then*

$$\|\tilde{\pi} - \pi\|_\infty \leq c\mu n$$

with probability at least $1 - 2/n$, where $c > 0$ is an absolute constant.

We conjecture proposition 5.2 still holds if the condition $q > (\log n)^4/\mu^2(2p-1)^4\sqrt{n}$ is replaced by the weaker assumption $q > (\log n)^4/\mu^2(2p-1)^4 n$.

5.1. **Proof sketch.** The proof of these results relies on classical perturbation arguments and is structured as follows.

- **Step 1:** Bound $\|\tilde{D}_S - D_S\|_2$, $\|\tilde{S} - S\|_2$, and hence $\|\tilde{L}_S - L_S\|_2$ with high probability using concentration inequalities on quadratic forms of Bernoulli variables.
- **Step 2.** Bound the eigengap between the Fiedler value and other eigenvalues.
- **Step 3.** Bound $\|\tilde{f} - f\|_2$ using Davis-Kahan theorem and the bound on $\|\tilde{L}_S - L_S\|_2$.
- **Step 4.** Translate this result into a bound on ranking consistency $\|\tilde{\pi} - \pi\|_\infty$.

We now turn to the proof itself.

5.2. **Step 1: Bounding $\|\tilde{L}_S - L_S\|_2$.** Here, we seek to bound $\|\tilde{D}_S - D_S\|_2$, $\|\tilde{S} - S\|_2$ and finally $\|\tilde{L}_S - L_S\|_2$, with high probability using concentration inequalities.

5.2.1. *Bounding the norm of the degree.* We first bound perturbations of the degree with both missing and corrupted comparisons.

Lemma 5.3. *For every $\mu \in (0,1)$ and $n \geq 100$, if $q \geq \frac{(\log n)^4}{\mu^2(2p-1)^4}$ then*

$$\|\tilde{D}_S - D_S\|_2 \leq \frac{3\mu n^2}{\sqrt{\log n}}$$

with probability at least $1 - 1/n$.

Proof. Let $R = \tilde{S} - S$. We have

$$R_{ij} = \sum_{k=1}^{n} C_{ik}C_{jk}\left(\frac{B_{ik}B_{jk}}{q^2(2p-1)^2} - 1\right).$$

Letting $d = \mathbf{diag}\, D_R$, we obtain

$$d_i = \sum_{j=1}^{n} R_{ij} = \sum_{j=1}^{n}\sum_{k=1}^{n} C_{ik}C_{jk}\left(\frac{B_{ik}B_{jk}}{q^2(2p-1)^2} - 1\right).$$

Notice that we can arbitrarily fix the diagonal values of R, $\tilde{C}$, and C to zeros. Hence we could take $j \neq i \neq k$ in the definition of d_i. This means in particular that the B_{ik} are independent of the B_{jk} in the summation.

We first seek a concentration inequality for each d_i. Notice that

$$\begin{aligned} d_i &= \sum_{j=1}^{n}\sum_{k=1}^{n} C_{ik}C_{jk}\left(\frac{B_{ik}B_{jk}}{q^2(2p-1)^2} - 1\right) \\ &= \sum_{k=1}^{n}\left(\frac{C_{ik}B_{ik}}{q(2p-1)}\sum_{j=1}^{n} C_{jk}\left(\frac{B_{jk}}{q(2p-1)} - 1\right)\right) + \sum_{k=1}^{n}\sum_{j=1}^{n} C_{ik}C_{jk}\left(\frac{B_{ik}}{q(2p-1)} - 1\right). \end{aligned}$$

The first term is quadratic while the second is linear, both terms have mean zero since the B_{ik} are independent of the B_{jk}. We begin by bounding the quadratic term. Let $X_{jk} = C_{jk}\left(\frac{1}{q(2p-1)}B_{jk} - 1\right)$. We have

$$\mathbf{E}(X_{jk}) = C_{jk}\left(\tfrac{qp-q(1-p)}{q(2p-1)} - 1\right) = 0,$$

$$\mathbf{var}(X_{jk}) = \frac{\mathbf{var}(B_{jk})}{q^2(2p-1)^2} = \frac{1}{q^2(2p-1)^2}(q - q^2(2p-1)^2) = \frac{1}{q(2p-1)^2} - 1 \leq \frac{1}{q(2p-1)^2},$$

and

$$|X_{jk}| = \left|\frac{B_{jk}}{q(2p-1)} - 1\right| \leq 1 + \frac{1}{q(2p-1)} \leq \frac{2}{q(2p-1)} \leq \frac{2}{q(2p-1)^2}.$$

By applying Bernstein's inequality we then get for any $t > 0$

$$\mathbf{Prob}\left(\left|\sum_{j=1}^{n} X_{jk}\right| > t\right) \leq 2\exp\left(\frac{-q(2p-1)^2t^2}{2(n+2t/3)}\right) \leq 2\exp\left(\frac{-q(2p-1)^2t^2}{2(n+t)}\right). \tag{9}$$

Now notice that

$$\mathbf{Prob}\left(\left|\sum_{k=1}^{n}\left(C_{ik}\frac{B_{ik}}{q(2p-1)}\sum_{j=1}^{n} X_{jk}\right)\right| > t\right) \leq \mathbf{Prob}\left(\sum_{k=1}^{n}\left(\frac{|B_{ik}|}{q(2p-1)}\right)\max_{l}|\sum_{j=1}^{n} X_{jl}| > t\right).$$

By applying a union bound to the first Bernstein inequality (9) we get for any $t > 0$

$$\mathbf{Prob}\left(\max_{l}\left|\sum_{j=1}^{n} X_{jl}\right| > \sqrt{t}\right) \leq 2n\exp\left(\frac{-tq(2p-1)^2}{2(n+\sqrt{t})}\right).$$

Moreover, since $\mathbf{E}\,|B_{ik}| = q$ we also get from Bernstein's inequality that for any $t > 0$

$$\mathbf{Prob}\left(\sum_{k=1}^{n}\frac{|B_{ik}|}{q(2p-1)} > \frac{n}{2p-1} + \sqrt{t}\right) \leq \exp\left(\frac{-tq(2p-1)^2}{2(n+\sqrt{t})}\right).$$

We deduce from these last three inequalities that for any $t > 0$

$$\mathbf{Prob}\left(\left|\sum_{k=1}^{n}\left(C_{ik}\frac{B_{ik}}{q(2p-1)}\sum_{j=1}^{n} X_{jk}\right)\right| > \sqrt{t}\left(\sqrt{t} + \frac{n}{2p-1}\right)\right) \leq (2n+1)\exp\left(\frac{-tq(2p-1)^2}{2(n+\sqrt{t})}\right).$$

Taking $t = \mu^2(2p-1)^2n^2/\log n$ and $q \geq \frac{(\log n)^4}{\mu^2(2p-1)^4}$, with $\mu \leq 1$, we have $\sqrt{t} \leq n$ and we deduce that

$$\mathbf{Prob}\left(\left|\sum_{k=1}^{n}\left(C_{ik}\frac{B_{ik}}{q(2p-1)}\sum_{j=1}^{n} X_{jk}\right)\right| > \frac{2\mu n^2}{\sqrt{\log n}}\right) \leq (2n+1)\exp\left(-\frac{(\log n)^3}{4}\right). \tag{10}$$

We now bound the linear term in B.

$$\begin{aligned}
\mathbf{Prob}\left(\left|\sum_{j=1}^{n}\sum_{k=1}^{n} C_{ik}C_{jk}\left(\frac{B_{ik}}{q(2p-1)} - 1\right)\right| > t\right) \quad &\leq \quad \mathbf{Prob}\left(\sum_{k=1}^{n}|C_{ik}|\max_{l}|\sum_{j=1}^{n} X_{jl}| > t\right) \\
&\leq \quad \mathbf{Prob}\left(\max_{k}|\sum_{j=1}^{n} X_{jk}| > t/n\right),
\end{aligned}$$

hence

$$\mathbf{Prob}\left(\left|\sum_{j=1}^{n}\sum_{k=1}^{n} C_{ik}C_{jk}\left(\frac{B_{ik}}{q(2p-1)} - 1\right)\right| > t\right) \leq 2n\exp\left(\frac{-t^2q(2p-1)^2}{2n^2(n+t/n)}\right).$$

Taking $t = \mu n^2/(\log n)^{1/2}$ and $q \geq \frac{(\log n)^4}{\mu^2(2p-1)^4}$, with $\mu \leq 1$, we have $t \leq n^2$ and we deduce that

$$\mathbf{Prob}\left(\left|\sum_{j=1}^{n}\sum_{k=1}^{n} C_{ik}C_{jk}\left(\frac{B_{ik}}{q(2p-1)} - 1\right)\right| > \frac{\mu n^2}{\sqrt{\log n}}\right) \leq 2n\exp\left(-\frac{(\log n)^3}{4}\right). \tag{11}$$

Finally, combining equations (10) and (11), we obtain for $q \geq \frac{(\log n)^4}{\mu^2(2p-1)^4}$, with $\mu \leq 1$

$$\mathbf{Prob}\left(|d_i| > \frac{3\mu n^2}{\sqrt{\log n}}\right) \leq (2n+1)\exp\left(-\frac{(\log n)^3}{4}\right).$$

Now, using a union bound, this shows that for $q \geq \frac{(\log n)^4}{\mu^2(2p-1)^4}$,

$$\textbf{Prob}\left(\max |d_i| > \frac{3\mu n^2}{\sqrt{\log n}}\right) \leq n(2n+1)\exp\left(-\frac{(\log n)^3}{4}\right),$$

which is less than $1/n$ for $n \geq 100$. ■

5.2.2. *Bounding perturbations of the comparison matrix C.* Here, we adapt results in [Achlioptas and McSherry, 2007] to bound perturbations of the comparison matrix. We will then use bounds on the perturbations of C to bound $\|\tilde{S} - S\|$.

Lemma 5.4. *For $n \geq 104$ and $q \geq \frac{(\log n)^4}{n}$,*

$$\|C - \tilde{C}\|_2 \leq \frac{c}{2p-1}\sqrt{\frac{n}{q}}, \tag{12}$$

with probability at least $1 - 2/n$, where c is an absolute constant.

Proof. The main argument of the proof is to use the independence of the C_{ij} for $i < j$ in order to bound $\|\tilde{C} - C\|_2$ by a constant times $\sigma\sqrt{n}$, where σ is the standard deviation of C_{ij}. To isolate independent entries in the perturbation matrix, we first need to break the anti-symmetry of $\tilde{C} - C$ by decomposing $X = \tilde{C} - C$ into its upper triangular part and its lower triangular part, *i.e.* $\tilde{C} - C = X_{\text{up}} + X_{\text{low}}$, with $X_{\text{up}} = -X_{\text{low}}^T$ (diagonal entries of $\tilde{C} - C$ can be arbitrarily set to 0). Entries of X_{up} are all independent, with variance less than the variance of $\tilde{C}_{ij}$. Indeed, lower entries of X_{up} are equal to 0 and hence have variance 0. Notice that

$$\|\tilde{C} - C\|_2 = \|X_{\text{up}} + X_{\text{low}}\|_2 \leq \|X_{\text{up}}\|_2 + \|X_{\text{low}}\|_2 \leq 2\|X_{\text{up}}\|_2,$$

so bounding $\|X_{\text{up}}\|_2$ will give us a bound on $\|X\|_2$. In the rest of the proof we write X_{up} instead of X to simplify notations. We can now apply [Achlioptas and McSherry, 2007, Th. 3.1] to X. Since

$$X_{ij} = \tilde{C}_{ij} - C_{ij} = C_{ij}\left(\frac{B_{ij}}{q(2p-1)} - 1\right),$$

we have (*cf.* proof of lemma 5.3) $\mathbf{E}(X_{ij}) = 0$, $\textbf{var}(X_{ij}) \leq \frac{1}{q(2p-1)^2}$, and $|X_{ij}| \leq \frac{2}{q(2p-1)}$. Hence for a given $\epsilon > 0$ such that

$$\frac{4}{q(2p-1)} \leq \left(\frac{\log(1+\epsilon)}{2\log(2n)}\right)^2 \frac{\sqrt{2n}}{\sqrt{q}(2p-1)}, \tag{13}$$

for any $\theta > 0$ and $n \geq 76$,

$$\textbf{Prob}\left(\|X\|_2 \geq 2(1+\epsilon+\theta)\frac{1}{\sqrt{q}(2p-1)}\sqrt{2n}\right) < 2\exp\left(-16\frac{\theta^2}{\epsilon^4}(\log n)^4\right). \tag{14}$$

For $q \geq \frac{(\log 2n)^4}{n}$ and taking $\epsilon \geq \exp(\sqrt{(}16/\sqrt{(}2))) - 1$ (so $\log(1+\epsilon)^2 \geq 16/\sqrt{2}$) means inequality (13) holds. Taking (14) with $\epsilon = 30$ and $\theta = 30$ we get

$$\textbf{Prob}\left(\|X\|_2 \geq \frac{112}{2p-1}\sqrt{\frac{n}{q}}\right) < 2\exp\left(-10^{-2}(\log n)^4\right). \tag{15}$$

Hence for $n \geq 104$, we have $(\log n)^3 > 100$ and

$$\textbf{Prob}\left(\|X\|_2 \geq \frac{112}{2p-1}\sqrt{\frac{n}{q}}\right) < 2/n.$$

Noting that $\log 2n \leq 1.15 \log n$ for $n \geq 104$, we obtain the desired result by choosing $c = 2 \times 112 \times \sqrt{1.15} \leq 241$. ■

5.2.3. *Bounding the perturbation of the similarity matrix* $\|S\|$. We now seek to bound $\|\tilde{S} - S\|$ with high probability.

Lemma 5.5. *For every* $\mu \in (0,1)$, $n \geq 104$, *if* $q > \frac{(\log n)^4}{\mu^2(2p-1)^2 n}$, *then*

$$\|\tilde{S} - S\|_2 \leq c\frac{\mu n^2}{\sqrt{\log n}},$$

with probability at least $1 - 2/n$, *where c is an absolute constant.*

Proof. Let $X = \tilde{C} - C$. We have

$$\tilde{C}\tilde{C}^T = (C + X)(C + X)^T = CC^T + XX^T + XC^T + CX^T,$$

hence

$$\tilde{S} - S = XX^T + XC^T + CX^T,$$

and

$$\|\tilde{S} - S\|_2 \leq \|XX^T\|_2 + \|XC^T\|_2 + \|CX^T\|_2 \leq \|X\|_2^2 + 2\|X\|_2\|C\|_2.$$

From lemma 5.4 we deduce that for $n \geq 104$ and $q \geq \frac{(\log n)^4}{n}$, with probability at least $1 - 2/n$

$$\|\tilde{S} - S\|_2 \leq \frac{c^2 n}{q(2p-1)^2} + \frac{2c}{2p-1}\sqrt{\frac{n}{q}}\|C\|_2. \tag{16}$$

Notice that $\|C\|_2^2 \leq \mathbf{Tr}(CC^T) = n^2$, hence $\|C\|_2 \leq n$ and

$$\|\tilde{S} - S\|_2 \leq \frac{c^2 n}{q(2p-1)^2} + \frac{2cn}{2p-1}\sqrt{\frac{n}{q}}. \tag{17}$$

By taking $q > \frac{(\log n)^4}{\mu^2(2p-1)^2 n}$, we get for $n \geq 104$ with probability at least $1 - 2/n$

$$\|\tilde{S} - S\|_2 \leq \frac{c^2\mu^2 n^2}{(\log n)^4} + \frac{2c\mu n^2}{(\log n)^2}.$$

Hence setting a new constant c with $c = \max(c^2(\log 104)^{-7/2}, 2c(\log 104)^{-3/2}) \leq 270$,

$$\|\tilde{S} - S\|_2 \leq c\frac{\mu n^2}{\sqrt{\log n}}$$

with probability at least $1 - 2/n$, which is the desired result. ■

5.3. **Step 2: Controlling the eigengap.** In the spectral clustering literature, several constructions for the Laplacian operators are suggested, namely the unnormalized Laplacian (used in SerialRank), the symmetric normalized Laplacian, and the non-symmetric normalized Laplacian. [Von Luxburg et al., 2008] show stronger consistency results for spectral clustering by using the non-symmetric normalized Laplacian.

Here, we obtain a closed-form expression for the Fiedler value and the Fiedler vector of the normalized Laplacian, while we only get an asymptotic expression for the unnormalized Laplacian (*cf.* section 7). This motivated us to provide an analysis of SerialRank robustness based on the normalized Laplacian, though in practice the use of the unnormalized Laplacian is valid.

In numerical experiments on synthetic datasets, the normalized Laplacian gives slightly worse results than the unnormalized Laplacian in regimes with small perturbations, and slightly better results in regimes with high perturbations (*cf.* Figure 5).

The following proposition gives a closed-form expression for the Fiedler value and the Fiedler vector of the normalized Laplacian, which yields bounds on the eigengap between the first, second and third smallest eigenvalues of the Laplacian.

Proposition 5.6. *Let $L^{\text{norm}} = \mathbf{I} - D^{-1}S$ be the non-symmetric normalized Laplacian of S. L^{norm} has a linear Fiedler vector, and its Fiedler value is equal to $2/3$.*

Proof. Let $x_i = i - \frac{n+1}{2}$ (x is uniform with mean zero). We want to show that $L^{\text{norm}}x = \lambda x$ or equivalently $Sx = (1-\lambda)Dx$. We develop both sides of the last equation, and use the following facts

$$S_{i,j} = n - |j-i|, \quad \sum_{k=1}^{n} k = \frac{n(n+1)}{2}, \quad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}.$$

We first get an expression for the degree of $d = S\mathbf{1} = \sum_{i=1}^{n} S_{i,k}$, with

$$\begin{aligned} d_i &= \sum_{k=1}^{i-1} S_{i,k} + \sum_{k=i}^{n} S_{i,k} \\ &= \sum_{k=1}^{i-1} (n-i+k) + \sum_{k=i}^{n} (n-k+i) \\ &= \frac{n(n-1)}{2} + i(n-i+1). \end{aligned}$$

Similarly we have

$$\begin{aligned} \sum_{k=1}^{n} kS_{i,k} &= \sum_{k=1}^{i-1} k(n-i+k) + \sum_{k=i}^{n} k(n-k+i) \\ &= \frac{n^2(n+1)}{2} + \frac{i(i-1)(2i-1)}{3} - \frac{n(n+1)(2n+1)}{6} - i(i-1) + i\frac{n(n+1)}{2}. \end{aligned}$$

Finally, setting $\lambda = 2/3$, notice that

$$\begin{aligned} [Sx]_i &= \sum_{k=1}^{n} \left(kS_{i,k} - k\frac{n+1}{2} \right) \\ &= \sum_{k=1}^{n} kS_{i,k} - \frac{n(n+1)^2}{2} \\ &= \frac{1}{3} \left(\frac{n(n-1)}{2} + i(n-i+1) \right) \left(i - \frac{n+1}{2} \right) \\ &= (1-\lambda) d_i x_i, \end{aligned}$$

which shows that $Sx = (1-\lambda)Dx$. ■

The next corollary will be useful in following proofs.

Corollary 5.7. *The Fiedler vector f of the unperturbed Laplacian satisfies $\|f\|_\infty \leq 2/\sqrt{n}$.*

Proof. We use the fact that f is collinear to the vector x defined by $x_i = i - \frac{n+1}{2}$ and verifies $\|f\|_2 = 1$. Indeed, for n odd $f_i = \frac{i-(n+1)/2}{a_n}$, with

$$a_n^2 = 2 \sum_{k=0}^{(n-1)/2} k^2 = \frac{1}{6} \frac{n-1}{2} \left(\frac{n-1}{2} + 1 \right) ((n-1)+1) = \frac{n^3-n}{12}.$$

Hence (for $n \geq 5$)

$$\|f\|_\infty = f_n = \frac{n-1}{2a_n} \leq \sqrt{\frac{3}{n-1}} \leq \frac{2}{\sqrt{n}}.$$

A similar reasoning applies for n even. ■

Lemma 5.8. *The minimum eigengap between the Fiedler value and other eigenvalues is bounded below by a constant for n sufficiently large.*

Proof. The first eigenvalue of the Laplacian is always 0, so we have for any n, $\lambda_2 - \lambda_1 = \lambda_2 = 2/3$. Moreover, using results from [Von Luxburg et al., 2008], we know that eigenvalues of the normalized Laplacian that are different from one converge to an asymptotic spectrum, and that the limit eigenvalues are "isolated". Hence there exists $n_0 > 0$ and $c > 0$ such that for any $n \geq n_0$ we have $\lambda_3 - \lambda_2 > c$. ■

Numerical experiments show that λ_3 converges to $0.9333\ldots$ very fast when n grows towards infinity.

5.4. **Step 3: Bounding the perturbation of the Fiedler vector $\|\tilde{f} - f\|_2$.** We first recall Weyl's inequality and a simplified version of Davis-Kahan theorem which can be found in [Stewart and Sun, 1990; Stewart, 2001].

Theorem 5.9. **(Weyl's inequality)** *Consider a symmetric positive definite matrix A and its perturbed version $\tilde{A}$. For any eigenvalue λ of A and its perturbed analogue $\tilde{\lambda}$, $\|\tilde{\lambda} - \lambda\|_\infty \leq \|\tilde{A} - A\|_2$.*

Theorem 5.10. **(Davis-Kahan)** *Consider a symmetric matrix L_S and its perturbed version $\tilde{L}_S$. If $|\tilde{\lambda}_3 - \lambda_2| > |\lambda_3 - \lambda_2|/2$ and $|\tilde{\lambda}_1 - \lambda_2| > |\lambda_1 - \lambda_2|/2$, then*

$$\|f - \tilde{f}\|_2 \leq \sqrt{2}\frac{\|\tilde{L}_S - L_S\|_2}{\min(\lambda_2 - \lambda_1, \lambda_3 - \lambda_2)}.$$

We can now compile results from previous sections to get a first perturbation bound and show ℓ_2 consistency of the Fiedler vector when comparisons are both missing and corrupted.

Proposition 5.11. *For every $\mu \in (0, 1)$ and n large enough, if $q > \frac{(\log n)^4}{\mu^2 (2p-1)^4 n}$, then*

$$\|\tilde{f} - f\|_2 \leq c\frac{\mu}{\sqrt{\log n}},$$

with probability at least $1 - 2/n$.

Proof. In order to use Davis-Kahan theorem, we need to relate perturbations of the normalized Laplacian matrix to perturbations of the similarity matrix. To simplify notations, we wite $L = \mathbf{I} - D^{-1}S$ and $\tilde{L} = \mathbf{I} - \tilde{D}^{-1}\tilde{S}$.

$$\begin{aligned}
\|\tilde{L} - L\|_2 &= \|\tilde{D}^{-1}\tilde{S} - D^{-1}S\|_2 \\
&= \|\tilde{D}^{-1}\tilde{S} - \tilde{D}^{-1}S + \tilde{D}^{-1}S - D^{-1}S\|_2 \\
&\leq \|\tilde{D}^{-1}\|_2 \|\tilde{S} - S\|_2 + \|S\|_2 \|\tilde{D}^{-1} - D^{-1}\|_2.
\end{aligned}$$

We first bound the second term. Notice that

$$\|\tilde{D}^{-1} - D^{-1}\|_2 = \max_i |\tilde{d}_i^{-1} - d_i^{-1}|,$$

where d_i (respectively $\tilde{d}_i$) is the degree of the i^{th} row of S (respectively $\tilde{S}$). Hence

$$\|\tilde{D}^{-1} - D^{-1}\|_2 = \max_i \frac{|\tilde{d}_i - d_i|}{|\tilde{d}_i||d_i|}.$$

From lemma 5.3, we have

$$\|D_R\|_2 = \max |\tilde{d}_i - d_i| \leq \frac{3\mu n^2}{\sqrt{\log n}},$$

with probability at least $1 - \frac{2}{n}$, hence

$$\|\tilde{D}^{-1} - D^{-1}\|_2 \leq \frac{\frac{3\mu n^2}{\sqrt{\log n}}}{d_i(d_i - \frac{3\mu n^2}{\sqrt{\log n}})}, \quad i = 1, \ldots, n, \text{ w.h.p.}$$

Since $d_i = \frac{n(n-1)}{2} + i(n-i+1)$ (*cf.* proof of proposition 5.6), for $\mu < 1$ there exists a constant c such that $d_i \left(d_i - \frac{3\mu n^2}{\sqrt{\log n}} \right) > cn^2$. We deduce that there exists an absolute constant c such that

$$\|\tilde{D}^{-1} - D^{-1}\|_2 \leq \frac{c\mu}{n^2\sqrt{\log n}} \text{ w.h.p.} \tag{18}$$

Moreover, we have

$$\|S\|_2 = \|CC^T + n\mathbf{1}\mathbf{1}^T\|_2 \leq \|C\|_2^2 + n\|\mathbf{1}\mathbf{1}^T\|_2 \leq 2n^2.$$

Hence,

$$\|S\|_2\|\tilde{D}^{-1} - D^{-1}\|_2 \leq \frac{c\mu}{\sqrt{\log n}} \text{ w.h.p,}$$

where $c = 2c$. Using lemma 5.5, we can similarly bound $\|\tilde{D}^{-1}\|_2\|\tilde{S} - S\|_2$ and we obtain

$$\|\tilde{L} - L\|_2 \leq \frac{c\mu}{\sqrt{\log n}} \text{ w.h.p,} \tag{19}$$

where c is an absolute constant. Finally, for small μ, Weyl's inequality, equation (19) together with Lemma 5.8 ensure that for n large enough with high probability $|\tilde{\lambda}_3 - \lambda_2| > |\lambda_3 - \lambda_2|/2$ and $|\tilde{\lambda}_1 - \lambda_2| > |\lambda_1 - \lambda_2|/2$. Hence we can apply the Davis-Kahan theorem on the following symmetric matrix

$$M = \begin{pmatrix} 0 & L^T \\ L & 0 \end{pmatrix},$$

which satisfies $\|M\|_2 = \|L\|_2$ and $M \begin{pmatrix} f \\ f \end{pmatrix} = \lambda_2 \begin{pmatrix} f \\ f \end{pmatrix}$. Compiling all constants into c we get the desired result. ■

5.5. **Bounding ranking perturbations** $\|\tilde{\pi} - \pi\|_\infty$**.** SerialRank's ranking is derived by sorting the Fiedler vector. While the consistency result in Proposition 5.11 shows the ℓ_2 estimation error going to zero as n goes to infinity, this is not sufficient to show that the ordering of the vector is preserved. To show ranking consistency, as in [Wauthier et al., 2013], we need to bound $\|\tilde{\pi} - \pi\|_\infty$ instead. We make the following conjecture on the behavior of the ℓ_∞ error and prove a weaker result, with $\sqrt{n}$ oversampling compared to the optimal rate.

Conjecture 5.12. *For every $\mu \in (0,1)$ and n large enough, if $q > \frac{(\log n)^4}{\mu^2(2p-1)^4 n}$, then*

$$\|\tilde{\pi} - \pi\|_\infty \leq c\mu n,$$

with probability at least $1 - 2/n$, where c is an absolute constant.

This result appears to be true in numerical experiments simulations, but we are only able to show a suboptimal sampling rate, losing a factor $\sqrt{n}$ in Proposition 5.2. The proof is in two parts: we first bound the l_∞ norm of the perturbation of the Fiedler vector, then translate this perturbation of the Fiedler vector into a perturbation of the ranking.

5.5.1. *Bounding the l_∞ norm of the Fiedler vector perturbation.* We start by a technical lemma bounding $\|(\tilde{S} - S)f\|_\infty$.

Lemma 5.13. *Let $r > 0$, for every $\mu \in (0,1)$ and n large enough, if $q > \frac{(\log n)^4}{\mu^2(2p-1)^4 n}$, then*

$$\|(\tilde{S} - S)f\|_\infty \leq \frac{3\mu n^{3/2}}{\sqrt{\log n}}$$

with probability at least $1 - 2/n$.

Proof. The proof is very much similar to the proof of lemma 5.3. Let $R = \tilde{S} - S$. We have

$$R_{ij} = \sum_{k=1}^{n} C_{ik} C_{jk} \left(\frac{B_{ik} B_{jk}}{q^2 (2p-1)^2} - 1 \right).$$

Therefore, let $\delta = Rf$

$$\delta_i = \sum_{j=1}^{n} R_{ij} f_j = \sum_{j=1}^{n} \sum_{k=1}^{n} C_{ik} C_{jk} \left(\frac{B_{ik} B_{jk}}{q^2 (2p-1)^2} - 1 \right) f_j.$$

Notice that we can arbitrarily fix the diagonal values of R, $\tilde{C}$, and C to zeros. Hence we could take $j \neq i \neq k$ in the definition of δ_i. This means in particular that the B_{ik} are independent of the B_{jk} in the summation.

We first obtain a concentration inequality for each δ_i. We will then use a union bound to bound $\|\delta\|_\infty = \max |\delta_i|$. Notice that

$$\begin{aligned} \delta_i &= \sum_{j=1}^{n} \sum_{k=1}^{n} C_{ik} C_{jk} \left(\frac{B_{ik} B_{jk}}{q^2 (2p-1)^2} - 1 \right) f_j \\ &= \sum_{k=1}^{n} \left(\frac{C_{ik} B_{ik}}{q(2p-1)} \sum_{j=1}^{n} C_{jk} \left(\frac{B_{jk}}{q(2p-1)} - 1 \right) f_j \right) + \sum_{k=1}^{n} \sum_{j=1}^{n} C_{ik} C_{jk} \left(\frac{B_{ik}}{q(2p-1)} - 1 \right) f_j. \end{aligned}$$

The first term is quadratic while the second is linear, both terms have mean zero since the B_{ik} are independent of the B_{jk}. We begin by bounding the quadratic term. Let $X_{jk} = C_{jk} (\frac{1}{q(2p-1)} B_{jk} - 1) f_j$. We have

$$\mathbf{E}(X_{jk}) = f_j C_{jk} (\tfrac{qp - q(1-p)}{q(2p-1)} - 1) = 0,$$

$$\mathbf{var}(X_{jk}) = \frac{f_j^2 \, \mathbf{var}(B_{jk})}{q^2 (2p-1)^2} = \frac{f_j^2}{q^2 (2p-1)^2} (q - q^2 (2p-1)^2) \leq \frac{f_j^2}{q (2p-1)^2},$$

$$|X_{jk}| = |f_j| |\frac{B_{jk}}{q(2p-1)} - 1| \leq \frac{2|f_j|}{q(2p-1)} \leq \frac{2\|f\|_\infty}{q(2p-1)^2}.$$

From corollary 5.7 $\|f\|_\infty \leq 2/\sqrt{n}$. Moreover $\sum_{j=0}^{n} f_j^2 = 1$ since f is an eigenvector. Hence, by applying Bernstein inequality we get for any $t > 0$

$$\mathbf{Prob} \left(|\sum_{j=1}^{n} X_{jk}| > t \right) \leq 2 \exp \left(\frac{-q(2p-1)^2 t^2}{2(1 + 2t/(3\sqrt{n}))} \right) \leq 2 \exp \left(\frac{-q(2p-1)^2 t^2 n}{2(n + \sqrt{n} t)} \right). \tag{20}$$

The rest of the proof is identical to the proof of lemma 5.3, replacing t by $\sqrt{n} t$. ■

We now prove the main result of this section, bounding $\|\tilde{f} - f\|_\infty$ with high probability when roughly $O(n^{3/2})$ comparisons are sampled.

Lemma 5.14. *For every $\mu \in (0,1)$ and n large enough, if $q > \frac{(\log n)^4}{\mu^2 (2p-1)^4 \sqrt{n}}$, then*

$$\|\tilde{f} - f\|_\infty \leq c \frac{\mu}{\sqrt{n \log n}}$$

with probability at least $1 - 2/n$, where c is an absolute constant.

Proof. Notice that by definition $\tilde{L}\tilde{f} = \tilde{\lambda}\tilde{f}$ and $Lf = \lambda f$. Hence for $\tilde{\lambda} > 0$

$$\begin{aligned}\tilde{f} - f &= \frac{\tilde{L}\tilde{f}}{\tilde{\lambda}} - f \\ &= \frac{\tilde{L}\tilde{f} - Lf}{\tilde{\lambda}} + \frac{(\lambda - \tilde{\lambda})f}{\tilde{\lambda}}.\end{aligned}$$

Moreover

$$\begin{aligned}\tilde{L}\tilde{f} - Lf &= (\mathbf{I} - \tilde{D}^{-1}\tilde{S})\tilde{f} - (\mathbf{I} - D^{-1}S)f \\ &= (\tilde{f} - f) + D^{-1}Sf - \tilde{D}^{-1}\tilde{S}\tilde{f} \\ &= (\tilde{f} - f) + D^{-1}Sf - \tilde{D}^{-1}\tilde{S}f + \tilde{D}^{-1}\tilde{S}f - \tilde{D}^{-1}\tilde{S}\tilde{f} \\ &= (\tilde{f} - f) + (D^{-1}S - \tilde{D}^{-1}\tilde{S})f + \tilde{D}^{-1}\tilde{S}(f - \tilde{f})\end{aligned}$$

Hence

$$(\mathbf{I}(\tilde{\lambda} - 1) + \tilde{D}^{-1}\tilde{S})(\tilde{f} - f) = (D^{-1}S - \tilde{D}^{-1}\tilde{S} + (\lambda - \tilde{\lambda})\mathbf{I})f. \tag{21}$$

Writing S_i the i^{th} row of S and d_i the degree of row i, using the triangle inequality, we deduce that

$$|\tilde{f}_i - f_i| \leq \frac{1}{|\tilde{\lambda} - 1|}\left(|(d_i^{-1}S_i - \tilde{d}_i^{-1}\tilde{S}_i)f| + |\lambda - \tilde{\lambda}||f_i| + |\tilde{d}_i^{-1}\tilde{S}_i(\tilde{f} - f)|\right). \tag{22}$$

We will now bound each term separately. Define

$$\begin{aligned}&\mathrm{Denom} = |\tilde{\lambda} - 1|, \\ &\mathrm{Num1} = |(\mathrm{d_i^{-1}S_i - \tilde{d}_i^{-1}\tilde{S}_i)f}|, \\ &\mathrm{Num2} = |\lambda - \tilde{\lambda}||\mathrm{f_i}|, \\ &\mathrm{Num3} = |\mathrm{\tilde{d}_i^{-1}\tilde{S}_i(\tilde{f} - f)}|.\end{aligned}$$

Bounding Denom. First notice that using Weyl's inequality and equation (19) (*cf.* proof of proposition 5.11), we have with probability at least $1 - 2/n$ $|\tilde{\lambda} - \lambda| \leq \|L_R\|_2 \leq \frac{c\mu}{\sqrt{\log n}}$. Therefore there exists an absolute constant c such that with probability at least $1 - 2/n$

$$|\tilde{\lambda} - 1| > c.$$

We now proceed with the numerator terms.
Bounding Num2. Using Weyl's inequality, corollary 5.7 and equation (19) (*cf.* proof of proposition 5.11), we deduce that w.h.p.

$$|(\lambda - \tilde{\lambda})f_i| \leq \frac{c\mu}{\sqrt{n \log n}},$$

where c is an absolute constant.
Bounding Num1. We now bound $|d_i^{-1}S_i - \tilde{d}_i^{-1}\tilde{S}_i|$. We have

$$\begin{aligned}|(\tilde{d}_i^{-1}\tilde{S}_i - d_i^{-1}S_i)f| &= |(\tilde{d}_i^{-1}\tilde{S}_i - \tilde{d}_i^{-1}S_i + \tilde{d}_i^{-1}S_i - d_i^{-1}S_i)f| \\ &\leq |\tilde{d}_i^{-1}||(\tilde{S}_i - S_i)f| + |(\tilde{d}_i^{-1} - d_i^{-1})S_if|.\end{aligned}$$

Using equation (18) from the proof of proposition 5.11, we have w.h.p. $|\tilde{d}_i^{-1} - d_i^{-1}| \leq \frac{c\mu}{n^2\sqrt{\log n}}$. Moreover

$$|\tilde{d}_i^{-1}| \leq |\tilde{d}_i^{-1} - d_i^{-1}| + |d_i^{-1}| \leq \frac{c_1\mu}{n^2\sqrt{\log n}} + \frac{c_2}{n^2} \leq \frac{c}{n^2}$$

w.h.p., where c is an absolute constant. Therefore

$$|(\tilde{d}_i^{-1}\tilde{S}_i - d_i^{-1}S_i)f| \quad \leq \quad \frac{c\mu}{n^2\sqrt{\log n}}|S_if| + \frac{c}{n^2}|(\tilde{S}_i - S_i)f| \text{ w.h.p.} \tag{23}$$

Using the definition of S and corollary 5.7, we get

$$|S_i f| \leq \sum_{j=0}^{n} S_{ij} \max_i |f_i| \leq c\frac{n^2}{\sqrt{n}} \leq cn^{3/2}, \tag{24}$$

where c is an absolute constant. Using lemma 5.13, we get

$$|(\tilde{S}_i - S_i)f| \leq \frac{3\mu n^{3/2}}{\sqrt{\log n}} \text{ w.h.p.} \tag{25}$$

Combining (23), (24) and (25) we deduce that there exists a constant c such that

$$|(\tilde{d_i}^{-1}\tilde{S}_i - d_i^{-1}S_i)f| \leq \frac{c\mu}{\sqrt{n\log n}} \text{ w.h.p.}$$

Bounding Num3. Finally we bound the remaining term $|\tilde{d_i}^{-1}\tilde{S}_i(\tilde{f} - f)|$. By Cauchy-Schwartz inequality we have,

$$|\tilde{d_i}^{-1}\tilde{S}_i(\tilde{f} - f)| \leq |\tilde{d_i}^{-1}| \|\tilde{S}_i\|_2 \|\tilde{f} - f\|_2.$$

Notice that

$$\|\tilde{S}_i\|_2 \leq \|S_i\|_2 + \|\tilde{S}_i - S_i\|_2 \leq \|S_i\|_2 + \|\tilde{S} - S\|_2.$$

Since $\|S_i\|_2^2 \leq \|S_1\|_2^2 \leq \frac{n(n+1)(2n+1)}{6}$ and $q > \frac{(\log n)^4}{\mu^2(2p-1)^2\sqrt{n}}$ we deduce from lemma 5.5 that w.h.p. $\|\tilde{S}_i\|_2 \leq \frac{c\mu n^{7/4}}{\sqrt{\log n}}$, where c is an absolute constant, for n large enough. Moreover, as shown above, $|\tilde{d_i}^{-1}| \leq \frac{c}{n^2}$ and we also get from proposition 5.11 that $\|\tilde{f} - f\|_2 \leq c\frac{\mu}{n^{1/4}\sqrt{\log n}}$ w.h.p. Hence we have

$$|\tilde{d_i}^{-1}\tilde{S}_i(\tilde{f} - f)| \leq \frac{c\mu^2 n^{7/4}}{n^2 n^{1/4}(\log n)} \leq \frac{c\mu}{\sqrt{n\log n}} \text{ w.h.p.},$$

where c is an absolute constant. Combining bounds on the denominator and numerator terms yields the desired result. ■

5.5.2. *Bounding the l_∞ norm of the ranking perturbation.* First note that the l_∞-norm of the ranking perturbation is equal to the number of pairwise disagreements between the true ranking and the retrieved one, *i.e.* for any i

$$|\tilde{\pi}_i - \pi_i| = \sum_{j<i} \mathbf{1}_{\tilde{f}_j > \tilde{f}_i} + \sum_{j>i} \mathbf{1}_{\tilde{f}_j < \tilde{f}_i}.$$

Now we will argue that when i and j are far apart, with high probability

$$\tilde{f}_j - \tilde{f}_i = (\tilde{f}_j - f_j) + (f_j - f_i) + (f_i - \tilde{f}_i)$$

will have the same sign as $j - i$. Indeed $|\tilde{f}_j - f_j|$ and $|\tilde{f}_i - f_i|$ can be bounded with high probability by a quantity less than $|f_j - f_i|/2$ for i and j sufficiently "far apart". Hence, $|\tilde{\pi}_i - \pi_i|$ is bounded by the number of pairs that are not sufficiently "far apart". We quantify the term "far apart" in the following proposition.

Proposition 5.15. *For every $\mu \in (0, 1)$ and n large enough, if $q > \frac{(\log n)^4}{\mu^2(2p-1)^2\sqrt{n}}$, then*

$$\|\tilde{\pi} - \pi\|_\infty \leq c\mu n,$$

with probability at least $1 - 2/n$, where c is an absolute constant.

Proof. We suppose w.l.o.g. in the following that the true ranking is the identity, hence the unperturbed Fiedler vector f is strictly increasing. We first notice that since

$$\tilde{f}_j - \tilde{f}_i = (\tilde{f}_j - f_j) + (f_j - f_i) + (f_i - \tilde{f}_i),$$

for any $j > i$

$$\|\tilde{f} - f\|_\infty \leq \frac{|f_j - f_i|}{2} \implies \tilde{f}_j \geq \tilde{f}_i.$$

Therefore, for a given i,

$$\sum_{j>i} \mathbf{1}_{\tilde{f}_j < \tilde{f}_i} \leq \sum_{j>i} \mathbf{1}_{\|\tilde{f}-f\|_\infty > \frac{|f_j - f_i|}{2}}.$$

By lemma 5.14 for $q > \frac{(\log n)^4}{\mu^2(2p-1)^2\sqrt{n}}$ with probability at least $1 - 2/n$

$$\|\tilde{f} - f\|_\infty \leq c\frac{\mu}{\sqrt{n \log n}}.$$

Hence w.h.p.

$$\sum_{j>i} \mathbf{1}_{\tilde{f}_j < \tilde{f}_i} \leq \sum_{j>i} \mathbf{1}_{\|\tilde{f}-f\|_\infty > \frac{|f_j - f_i|}{2}} \leq \sum_{j>i} \mathbf{1}_{c\frac{\mu}{\sqrt{n \log n}} > \frac{|f_j - f_i|}{2}}.$$

Now we use the fact that $f_i = \frac{i-(n+1)/2}{a_n}$, with

$$a_n^2 = 2\sum_{k=0}^{(n-1)/2} k^2 = \frac{1}{6}\frac{n-1}{2}\left(\frac{n-1}{2}+1\right)((n-1)+1) = \frac{n^3 - n}{12}.$$

(for n odd). Therefore (up to a negligible change in c)

$$\frac{c\mu}{\sqrt{n \log n}} > \frac{|f_j - f_i|}{2} \iff \frac{c\mu}{\sqrt{n \log n}} > \frac{|j-i|\sqrt{3}}{n^{3/2}} \iff \frac{c\mu n}{\sqrt{3 \log n}} > |j - i|.$$

Redefining $c = \frac{c}{\sqrt{3}}$, we deduce that w.h.p.

$$\sum_{j>i} \mathbf{1}_{\tilde{f}_j < \tilde{f}_i} \leq \sum_{j>i} \mathbf{1}_{\frac{c\mu n}{\sqrt{\log n}} > |j-i|} = \sum_{i=1}^{\left\lfloor \frac{c\mu n}{\sqrt{\log n}} \right\rfloor} (n - i) \leq \frac{c\mu n}{\sqrt{\log n}}.$$

Similarly w.h.p.

$$\sum_{j<i} \mathbf{1}_{\tilde{f}_j > \tilde{f}_i} \leq \frac{c\mu n}{\sqrt{\log n}}.$$

Finally, we get the desired result (w.h.p)

$$|\tilde{\pi}_i - \pi_i| = \sum_{j<i} \mathbf{1}_{\tilde{f}_j > \tilde{f}_i} + \sum_{j>i} \mathbf{1}_{\tilde{f}_j < \tilde{f}_i} \leq \frac{c\mu n}{\sqrt{\log n}},$$

where c is an absolute constant. ■

6. Numerical Experiments

We now describe numerical experiments using both synthetic and real datasets to compare the performance of SerialRank with several classical ranking methods.

6.1. **Synthetic Datasets.** The first synthetic dataset consists of a binary matrix of pairwise comparisons derived from a given ranking of n items with uniform, randomly distributed corrupted or missing entries. A second synthetic dataset consists of a full matrix of pairwise comparisons derived from a given ranking of n items, with added "local" noise on the similarity between nearby items. Specifically, given a positive integer m, we let $C_{i,j} = 1$ if $i < j - m$, $C_{i,j} \sim \text{Unif}[-1, 1]$ if $|i - j| \leq m$, and $C_{i,j} = -1$ if $i > j + m$. In Figure 2, we measure the Kendall τ correlation coefficient between the true ranking and the retrieved ranking, when varying either the percentage of corrupted comparisons or the percentage of missing comparisons. Kendall's τ counts the number of agreeing pairs minus the number of disagreeing pairs between two rankings, scaled by the total number of pairs, so that it takes values between -1 and 1. Experiments were performed with $n = 100$ and reported Kendall τ values were averaged over 50 experiments, with standard deviation less than 0.02 for points of interest (*i.e.* with Kendall $\tau > 0.8$).

6.2. **Real Datasets.** The first real dataset consists of pairwise comparisons derived from outcomes in the TopCoder algorithm competitions. We collected data from 103 competitions among 2742 coders over a period of about one year. Pairwise comparisons are extracted from the ranking of each competition and then averaged for each pair. TopCoder maintains ratings for each participant, updated in an online scheme after each competition, which were also included in the benchmarks. To measure performance in Figure 3, we compute the percentage of upsets (i.e. comparisons disagreeing with the computed ranking), which is closely related to the Kendall τ (by an affine transformation if comparisons were coming from a consistent ranking). We refine this metric by considering only the participants appearing in the top k, for various values of k, i.e. computing

$$l_k = \frac{1}{|\mathcal{C}_k|} \sum_{i,j \in \mathcal{C}_k} \mathbf{1}_{r(i)>r(j)} \mathbf{1}_{C_{i,j}<0}, \tag{26}$$

where $\mathcal{C}$ are the pairs (i, j) that are compared and such that i, j are both ranked in the top k, and $r(i)$ is the rank of i. Up to scaling, this is the loss considered in [Kenyon-Mathieu and Schudy, 2007].

TABLE 1. Ranking of teams in the England premier league season 2013-2014.

Official	**Row-sum**	**RC**	**BTL**	**SerialRank**	**Semi-Supervised**
Man City (86)	Man City	Liverpool	Man City	Man City	Man City
Liverpool (84)	Liverpool	Arsenal	Liverpool	Chelsea	Chelsea
Chelsea (82)	Chelsea	Man City	Chelsea	Liverpool	Liverpool
Arsenal (79)	Arsenal	Chelsea	Arsenal	Arsenal	Everton
Everton (72)	Everton	Everton	Everton	Everton	Arsenal
Tottenham (69)	Tottenham	Tottenham	Tottenham	Tottenham	Tottenham
Man United (64)	Man United	Man United	Man United	Southampton	Man United
Southampton (56)	Southampton	Southampton	Southampton	Man United	Southampton
Stoke (50)	Stoke	Stoke	Stoke	Stoke	Newcastle
Newcastle (49)	Newcastle	Newcastle	Newcastle	Swansea	Stoke
Crystal Palace (45)	Crystal Palace	Swansea	Crystal Palace	Newcastle	West Brom
Swansea (42)	Swansea	Crystal Palace	Swansea	West Brom	Swansea
West Ham (40)	West Brom	West Ham	West Brom	Hull	Crystal Palace
Aston Villa (38)	West Ham	Hull	West Ham	West Ham	Hull
Sunderland (38)	Aston Villa	Aston Villa	Aston Villa	Cardiff	West Ham
Hull (37)	Sunderland	West Brom	Sunderland	Crystal Palace	Fulham
West Brom (36)	Hull	Sunderland	Hull	Fulham	Norwich
Norwich (33)	Norwich	Fulham	Norwich	Norwich	Sunderland
Fulham (32)	Fulham	Norwich	Fulham	Sunderland	Aston Villa
Cardiff (30)	Cardiff	Cardiff	Cardiff	Aston Villa	Cardiff

6.3. **Semi-Supervised Ranking.** We illustrate here how, in a semi-supervised setting, one can interactively enforce some constraints on the retrieved ranking, using e.g. the semi-supervised seriation algorithm in [Fogel et al., 2013]. We compute rankings of England Football Premier League teams for season 2013-2014

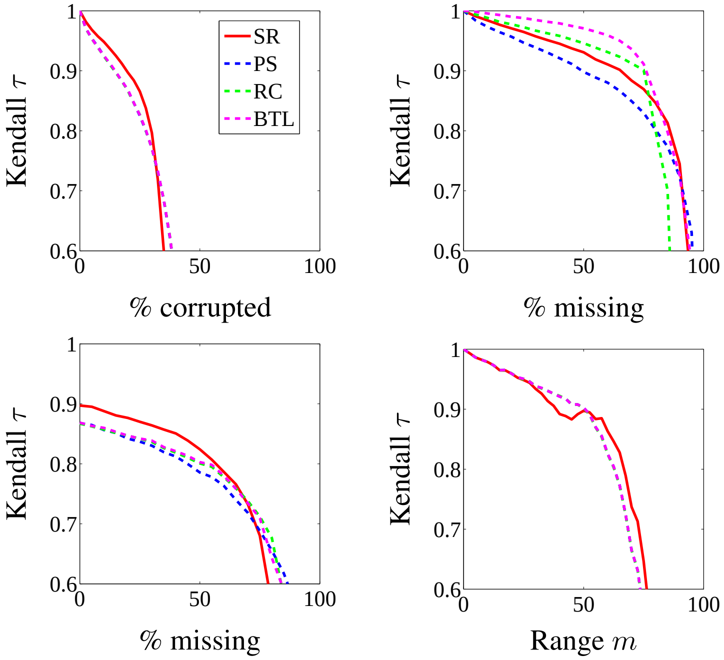

FIGURE 2. Kendall τ (higher is better) for SerialRank (SR, full red line), row-sum (PS, [Wauthier et al., 2013] dashed blue line), rank centrality (RC [Negahban et al., 2012] dashed green line), and maximum likelihood (BTL [Bradley and Terry, 1952], dashed magenta line). In the first synthetic dataset, we vary the proportion of corrupted comparisons *(top left)*, the proportion of observed comparisons *(top right)* and the proportion of observed comparisons, with 20% of comparisons being corrupted *(bottom left)*. We also vary the parameter m in the second synthetic dataset *(bottom right)*.

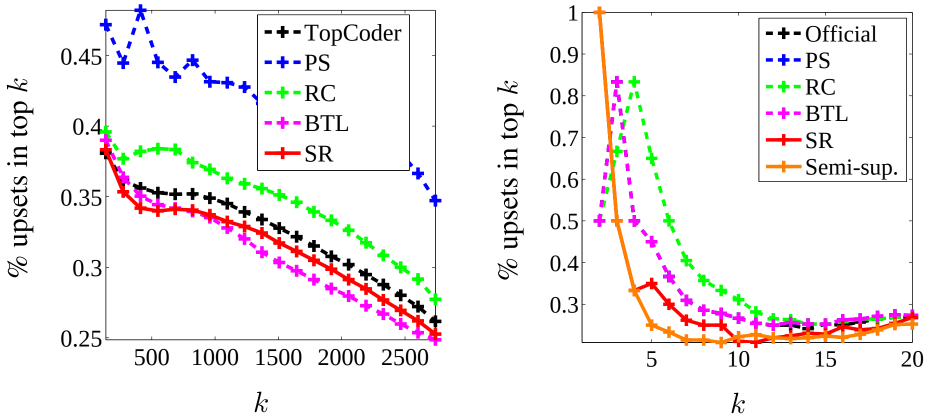

FIGURE 3. Percentage of upsets (i.e. disagreeing comparisons, lower is better) defined in (26), for various values of k and ranking methods, on TopCoder (*left*) and football data (*right*).

(*cf.* figure 4 for seasons 2011-2012 and 2012-2013). Comparisons are defined as the averaged outcome (win, loss, or tie) of home and away games for each pair of teams. As shown in Table 1, the top half of SerialRank ranking is very close to the official ranking calculated by sorting the sum of points for each team

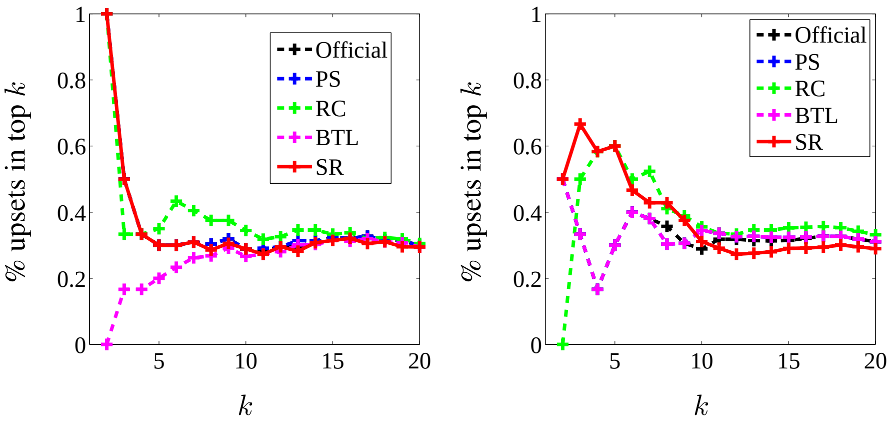

FIGURE 4. Percentage of upsets (i.e. disagreeing comparisons, lower is better) defined in (26), for various values of k and ranking methods, on England Premier League 2011-2012 season (*left*) and 2012-2013 season (*right*).

(3 points for a win, 1 point for a tie). However, there are significant variations in the bottom half, though the number of upsets is roughly the same as for the official ranking. To test semi-supervised ranking, suppose for example that we are not satisfied with the ranking of Aston Villa (last team when ranked by the spectral algorithm), we can explicitly enforce that Aston Villa appears before Cardiff, as in the official ranking. In the ranking based on the corresponding semi-supervised seriation problem, Aston Villa is not last anymore, though the number of disagreeing comparisons remains just as low (*cf.* Figure 3, *right*).

REFERENCES

Achlioptas, D. and McSherry, F. [2007], 'Fast computation of low-rank matrix approximations', *Journal of the ACM* **54**(2).

Ailon, N. [2011], Active learning ranking from pairwise preferences with almost optimal query complexity., *in* 'NIPS', pp. 810–818.

Atkins, J., Boman, E., Hendrickson, B. et al. [1998], 'A spectral algorithm for seriation and the consecutive ones problem', *SIAM J. Comput.* **28**(1), 297–310.

Barbeau, E. [1986], 'Perron's result and a decision on admissions tests', *Mathematics Magazine* pp. 12–22.

Blum, A., Konjevod, G., Ravi, R. and Vempala, S. [2000], 'Semidefinite relaxations for minimum bandwidth and other vertex ordering problems', *Theoretical Computer Science* **235**(1), 25–42.

Bradley, R. A. and Terry, M. E. [1952], 'Rank analysis of incomplete block designs: I. the method of paired comparisons', *Biometrika* pp. 324–345.

Feige, U. and Lee, J. R. [2007], 'An improved approximation ratio for the minimum linear arrangement problem', *Information Processing Letters* **101**(1), 26–29.

Fogel, F., Jenatton, R., Bach, F. and d'Aspremont, A. [2013], 'Convex relaxations for permutation problems', *NIPS 2013, arXiv:1306.4805* .

Freund, Y., Iyer, R., Schapire, R. E. and Singer, Y. [2003], 'An efficient boosting algorithm for combining preferences', *The Journal of machine learning research* **4**, 933–969.

Herbrich, R., Minka, T. and Graepel, T. [2006], Trueskill$^{\mathrm{TM}}$: A bayesian skill rating system, *in* 'Advances in Neural Information Processing Systems', pp. 569–576.

Huber, P. J. [1963], 'Pairwise comparison and ranking: optimum properties of the row sum procedure', *The annals of mathematical statistics* pp. 511–520.

Hunter, D. R. [2004], 'MM algorithms for generalized bradley-terry models', *Annals of Statistics* pp. 384–406.

Jamieson, K. G. and Nowak, R. D. [2011], Active ranking using pairwise comparisons., *in* 'NIPS', Vol. 24, pp. 2240–2248.

Joachims, T. [2002], Optimizing search engines using clickthrough data, *in* 'Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 133–142.

Keener, J. P. [1993], 'The perron-frobenius theorem and the ranking of football teams', *SIAM review* **35**(1), 80–93.

Kendall, M. G. and Smith, B. B. [1940], 'On the method of paired comparisons', *Biometrika* **31**(3-4), 324–345.

Kenyon-Mathieu, C. and Schudy, W. [2007], How to rank with few errors, *in* 'Proceedings of the thirty-ninth annual ACM symposium on Theory of computing', ACM, pp. 95–103.

Kleinberg, J. [1999], 'Authoritative sources in a hyperlinked environment', *Journal of the ACM* **46**, 604–632.

Kuczynski, J. and Wozniakowski, H. [1992], 'Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start', *SIAM J. Matrix Anal. Appl* **13**(4), 1094–1122.

Luce, R. [1959], *Individual choice behavior*, Wiley.

Negahban, S., Oh, S. and Shah, D. [2012], Iterative ranking from pairwise comparisons., *in* 'NIPS', pp. 2483–2491.

Page, L., Brin, S., Motwani, R. and Winograd, T. [1998], 'The pagerank citation ranking: Bringing order to the web', *Stanford CS Technical Report* .

Saaty, T. L. [1980], 'The analytic hierarchy process: planning, priority setting, resources allocation', *New York: McGraw* .

Schapire, W. W. C. R. E. and Singer, Y. [1998], Learning to order things, *in* 'Advances in Neural Information Processing Systems 10: Proceedings of the 1997 Conference', Vol. 10, MIT Press, p. 451.

Stewart, G. [2001], *Matrix Algorithms Vol. II: Eigensystems*, Society for Industrial Mathematics.

Stewart, G. and Sun, J. [1990], *Matrix perturbation theory*, Academic Press.

Von Luxburg, U., Belkin, M. and Bousquet, O. [2008], 'Consistency of spectral clustering', *The Annals of Statistics* pp. 555–586.

Wauthier, F. L., Jordan, M. I. and Jojic, N. [2013], Efficient ranking from pairwise comparisons, *in* 'Proceedings of the 30th International Conference on Machine Learning (ICML)'.

7. Appendix

We now detail several complementary technical results.

7.1. **Exact recovery results with missing entries.**

Here, as in Section 4, we study the impact of one missing comparison on SerialRank, then extend the result to multiple missing comparisons.

Proposition 7.1. *Given pairwise comparisons $C_{s,t} \in \{-1, 0, 1\}$ between items ranked according to their indices, suppose only one comparison $C_{i,j}$ is missing, with $j - i > 1$ (i.e. $C_{i,j} = 0$), then S^{match} defined in* (3) *remains* strict-R *and the score vector remains strictly monotonic.*

Proof. We use the same proof technique as in proposition 4.2. We write the true score and comparison matrix w and C, while the observations are written $\hat{w}$ and $\hat{C}$ respectively. This means in particular that $\hat{C}_{i,j} = 0$. To simplify notations we denote by S the similarity matrix S^{match} (respectively $\hat{S}$ when the similarity is computed from observations). We first study the impact of the missing comparison $C_{i,j}$ for $i < j$ on the score vector $\hat{w}$. We have

$$\hat{w}_i = \sum_{k=1}^{n} \hat{C}_{k,i} = \sum_{k=1}^{n} C_{k,i} + \hat{C}_{j,i} - C_{j,i} = w_i + 1,$$

similarly $\hat{w}_j = w_j - 1$, whereas for $k \neq i, j$, $\hat{w}_k = w_k$. Hence, w is still strictly increasing if $j > i + 1$. If $j = i + 1$ there is a tie between w_i and w_{i+1}. Now we show that the similarity matrix defined in (3) is a

R-matrix. Writing $\hat{S}$ in terms of S, we get

$$[\hat{C}\hat{C}^T]_{i,t} = \sum_{k\neq j}\left(\hat{C}_{i,k}\hat{C}_{t,k}\right) + \hat{C}_{i,j}\hat{C}_{t,j} = \sum_{k\neq j}\left(C_{i,k}C_{t,k}\right) = \left\{\begin{array}{ll} [CC^T]_{i,t} - 1 & \text{if } t < j \\ \left[CC^T\right]_{i,t} + 1 & \text{if } t > j. \end{array}\right.$$

We thus get

$$\hat{S}_{i,t} = \left\{\begin{array}{ll} S_{i,t} - \frac{1}{2} & \text{if } t < j \\ S_{i,t} + \frac{1}{2} & \text{if } t > j, \end{array}\right.$$

(remember there is a factor $1/2$ in the definition of S). Similarly we get for any $t \neq i$

$$\hat{S}_{j,t} = \left\{\begin{array}{ll} S_{j,t} + \frac{1}{2} & \text{if } t < i \\ S_{j,t} - \frac{1}{2} & \text{if } t > i. \end{array}\right.$$

Finally, for the single corrupted index pair (i,j), we get

$$\hat{S}_{i,j} = \frac{1}{2}\left(n + \sum_{k\neq i,j}\left(\hat{C}_{i,k}\hat{C}_{j,k}\right) + \hat{C}_{i,i}\hat{C}_{j,i} + \hat{C}_{i,j}\hat{C}_{j,j}\right) = S_{i,j} - 0 + 0 = S_{i,j}.$$

For all other coefficients (s,t) such that $s,t \neq i,j$, we have $\hat{S}_{s,t} = S_{s,t}$. Meaning all rows or columns outside of i,j are left unchanged. We first observe that these last equations, together with our assumption that $j - i > 2$, mean that

$$\hat{S}_{s,t} \geq \hat{S}_{s+1,t} \quad \text{and} \quad \hat{S}_{s,t+1} \geq \hat{S}_{s,t}, \quad \text{for any } s < t$$

so $\hat{S}$ remains an R-matrix. To show uniqueness of the retrieved order, we need $j - i > 1$. Indeed, when $j - i > 1$ all these R constraints are strict, which means that $\hat{S}$ is still a strict R-matrix, hence the desired result. ■

We can extend this result to the case where multiple comparisons are missing.

Proposition 7.2. *Given pairwise comparisons $C_{s,t} \in \{-1,0,1\}$ between items ranked according to their indices, suppose m comparisons indexed $(i_1,j_1),\ldots,(i_m,j_m)$ are missing, i.e. $C_{i_l,j_j} = 0$ for $i = l,\ldots,m$. If the following condition (27) holds true,*

$$|s - t| > 1 \text{ for all } s \neq t \in \{i_1,\ldots,i_m,j_1,\ldots,j_m\} \tag{27}$$

then S^{match} defined in (3) remains strict-R *and the score vector remains strictly monotonic.*

Proof. Proceed similarly as in the proof of proposition 4.3, except that shifts are divided by two. ■

We also get the following corollary.

Corollary 7.3. *Given pairwise comparisons $C_{s,t} \in \{-1,0,1\}$ between items ranked according to their indices, suppose m comparisons indexed $(i_1,j_1),\ldots,(i_m,j_m)$ are either corrupted or missing. If condition (7) holds true then S^{match} defined in (3) remains* strict-R.

Proof. Proceed similarly as the proof of proposition 4.3, except that shifts are divided by two for missing comparisons. ■

7.2. **Numerical experiments with normalized Laplacian.** As shown in figure 5, results are very similar to those of SerialRank with unnormalized Laplacian. We lose a bit of performance in terms of robustness to corrupted comparisons.

7.3. **Spectrum of the unnormalized Laplacian matrix.**

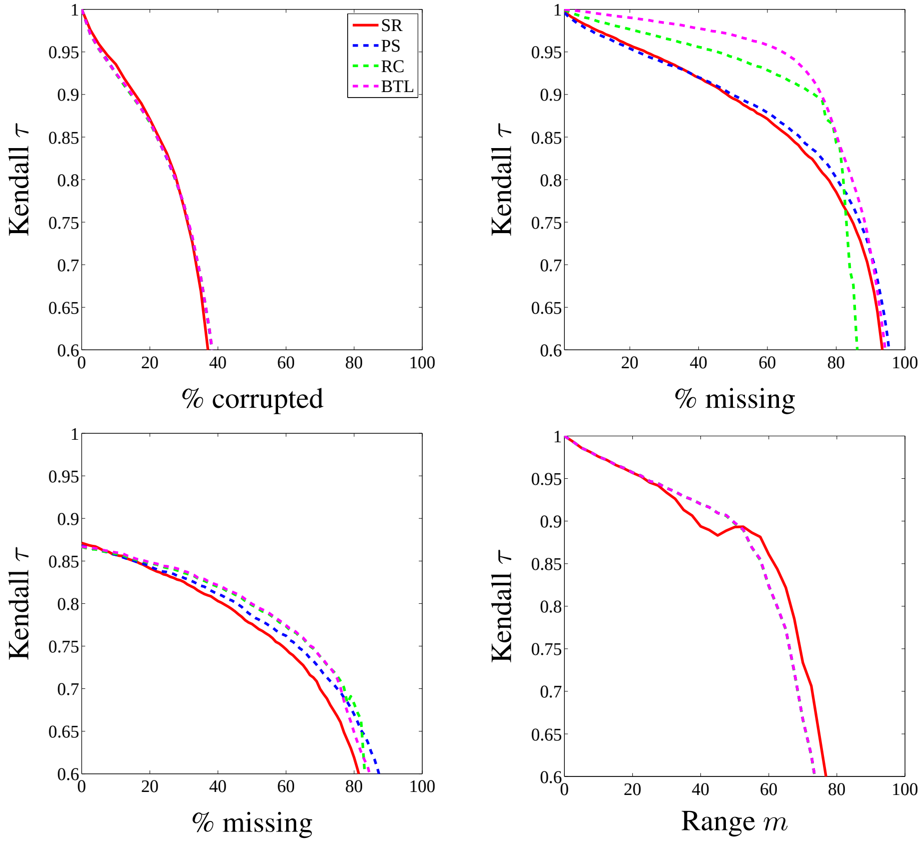

FIGURE 5. Kendall τ (higher is better) for SerialRank with normalized Laplacian (SR, full red line), row-sum (PS, [Wauthier et al., 2013] dashed blue line), rank centrality (RC [Negahban et al., 2012] dashed green line), and maximum likelihood (BTL [Bradley and Terry, 1952], dashed magenta line). In the first synthetic dataset, we vary the proportion of corrupted comparisons *(top left)*, the proportion of observed comparisons *(top right)* and the proportion of observed comparisons, with 20% of comparisons being corrupted *(bottom left)*. We also vary the parameter m in the second synthetic dataset *(bottom right)*.

7.3.1. *Asymptotic Fiedler value and Fiedler vector.* We use results on the convergence of Laplacian operators to provide a description of the spectrum of the unnormalized Laplacian in SerialRank. Following the same analysis as in [Von Luxburg et al., 2008] we can prove that asymptotically, once normalized by n^2, apart from the first and second eigenvalue, the spectrum of the Laplacian matrix is contained in the interval $[0.5, 0.75]$. Moreover, we can characterize the eigenfunctions of the limit Laplacian operator by a differential equation, enabling to have an asymptotic approximation for the Fiedler vector.

Taking the same notations as in [Von Luxburg et al., 2008] we have here $k(x, y) = 1 - |x - y|$. The degree function is

$$d(x) = \int_0^1 k(x, y) d\,\mathbf{Prob}(y) = \int_0^1 k(x, y) d(y)$$

(samples are uniformly ranked). Simple calculations give

$$d(x) = -x^2 + x + 1/2.$$

We deduce that the range of d is $[0.5, 0.75]$. Interesting eigenvectors (*i.e.* here the second eigenvector) are not in this range. We can also characterize eigenfunctions f by

$$\begin{aligned}
& Uf(x) = \lambda f(x) \;\; \forall x \in [0,1] \\
\Leftrightarrow \;\; & Mdf(x) - Sf(x) = \lambda f(x) \\
\Leftrightarrow \;\; & d(x)f(x) - \int_0^1 k(x,y)f(y)d(y) = \lambda f(x) \\
\Leftrightarrow \;\; & f(x)(-x^2 + x + 1/2) - \int_0^1 (1 - |x-y|)f(y)d(y) = \lambda f(x)
\end{aligned}$$

Differentiating twice we get

$$f''(x)(1/2 - \lambda + x - x^2) + 2f'(x)(1-2x) = 0. \tag{28}$$

The asymptotic expression for the Fiedler vector is then a solution to this differential equation, with $\lambda < 0.5$. Let γ_1 and γ_2 be the roots of $(1/2 - \lambda + x - x^2)$ (with $\gamma_1 < \gamma_2$). We can suppose that $x \in (\gamma_1, \gamma_2)$ since the degree function is nonnegative. Simple calculations show that

$$f'(x) = \frac{A}{(x-\gamma_1)^2(x-\gamma_2)^2}$$

is solution to (28), where A is a constant. Now we note that

$$\begin{aligned}
\frac{1}{(x-\gamma_1)^2(x-\gamma_2)^2} \;\; = \;\; & \frac{1}{(\gamma_1-\gamma_2)^2(\gamma_2-x)^2} + \frac{1}{(\gamma_1-\gamma_2)^2(\gamma_1-x)^2} \\
& - \frac{2}{(\gamma_1-\gamma_2)^3(\gamma_2-x)} + \frac{2}{(\gamma_1-\gamma_2)^3(\gamma_1-x)}.
\end{aligned}$$

We deduce that the solution f to (28) satisfies

$$f(x) = B + \frac{A}{(\gamma_1-\gamma_2)^2}\left(\frac{1}{\gamma_1 - x} + \frac{1}{\gamma_2 - x}\right) - \frac{2A}{(\gamma_1-\gamma_2)^3}\left(\log(x-\gamma_1) - \log(\gamma_2 - x)\right),$$

where A and B are two constants. Since f is orthogonal to the unitary function for $x \in (0,1)$, we must have $f(1/2) = 0$, hence B=0 (we use the fact that $\gamma_1 = \frac{1-\sqrt{1+4\alpha}}{2}$ and $\gamma_2 = \frac{1+\sqrt{1+4\alpha}}{2}$, where $\alpha = 1/2 - \lambda$).

As shown in figure 6 , the asymptotic expression for the Fiedler vector is very accurate numerically, even for small values of n. The asymptotic Fiedler value is also very accurate (2 digits precision for $n = 10$, once normalized by n^2).

7.3.2. *Bounding the eigengap.* We now give two simple propositions on the Fiedler value and the third eigenvalue of the Laplacian matrix, which enable us to bound the eigengap between the second and the third eigenvalues.

Proposition 7.4. *Given all comparisons indexed by their true ranking, let λ_2 be the Fiedler value of S^{match}, we have*

$$\lambda_2 \leq \frac{2}{5}(n^2+1).$$

Proof. Consider the vector x whose elements are uniformly spaced and such that $x^T\mathbf{1} = 0$ and $\|x\|_2 = 1$. x is a feasible solution to the Fiedler eigenvalue minimization problem. Therefore,

$$\lambda_2 \leq x^T L x.$$

Simple calculations give $x^T L x = \frac{2}{5}(n^2+1)$. ■

Numerically the bound is very close to the true Fiedler value: $\lambda_2/n^2 \approx 0.39$ and $2/5 = 0.4$.

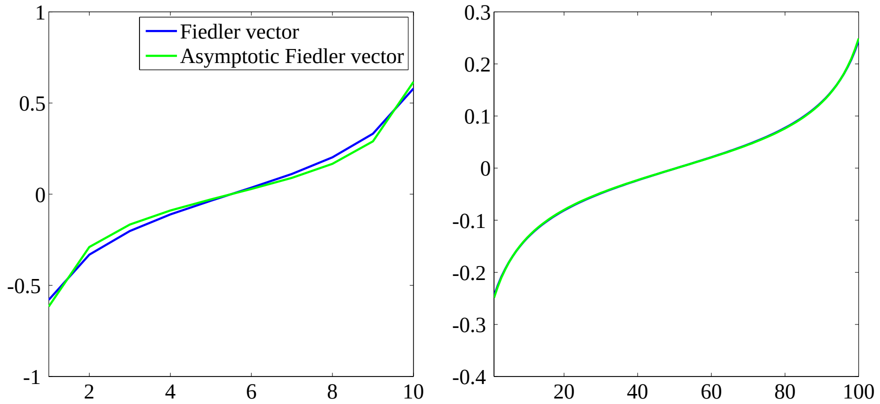

FIGURE 6. Comparison between the asymptotic analytical expression of the Fiedler vector and the numeric values obtained from eigenvalue decomposition, for $n = 10$ (*left*) and $n = 100$ (*right*).

Proposition 7.5. *Given all comparisons indexed by their true ranking, the vector $v = [\alpha, -\beta, \ldots, -\beta, \alpha]^T$ where α and β are such that $v^T\mathbf{1} = 0$ and $\|v\|_2 = 1$ is an eigenvector of the Laplacian matrix L of S^{match} The corresponding eigenvalue is $\lambda = n(n+1)/2$.*

Proof. Check that $Lv = \lambda v$. ■

7.4. **Other choices of similarities.** The results in this paper shows that forming a similarity matrix (R-matrix) from pairwise preferences will produce a valid ranking algorithm. In what follows, we detail a few options extending the results of Section 2.2.

7.4.1. *Cardinal comparisons.* When input comparisons take continuous values between -1 and 1, several choice of similarities can be made. First possibility is to use S^{glm}. An other option is to directly provide $1 - \mathrm{abs(C)}$ as a similarity to SerialRank. This option has a much better computational cost.

7.4.2. *Adjusting contrast in S^{match}.* Instead of providing S^{match} to SerialRank, we can change the "contrast" of the similarity, *i.e.* take the similarity whose elements are powers of the elements of S^{match}.

$$S^{\mathrm{contrast}}_{i,j} = (S^{\mathrm{match}}_{i,j})^{\alpha}.$$

This construction gives slightly better results in terms of robustness to noise on synthetic datasets.

C.M.A.P., ÉCOLE POLYTECHNIQUE,
PALAISEAU, FRANCE.
E-mail address: fajwel.fogel@cmap.polytechnique.fr

CNRS & D.I., UMR 8548,
ÉCOLE NORMALE SUPÉRIEURE, PARIS, FRANCE.
E-mail address: aspremon@ens.fr

MICROSOFT RESEARCH,
CAMBRIDGE, UK.
E-mail address: milanv@microsoft.com