Enhanced statistical rankings via targeted data collection

Braxton Osting BRAXTON@MATH.UCLA.EDU
Department of Mathematics, University of California, Los Angeles

Christoph Brune BRUNE@MATH.UCLA.EDU
Department of Mathematics, University of California, Los Angeles

Stanley Osher SJO@MATH.UCLA.EDU
Department of Mathematics, University of California, Los Angeles

Abstract

Given a graph where vertices represent alternatives and pairwise comparison data, y_{ij}, is given on the edges, the *statistical ranking problem* is to find a potential function, defined on the vertices, such that the gradient of the potential function agrees with pairwise comparisons. We study the dependence of the statistical ranking problem on the available pairwise data, *i.e.*, pairs (i, j) for which the pairwise comparison data y_{ij} is known, and propose a framework to identify data which, when augmented with the current dataset, maximally increases the Fisher information of the ranking. Under certain assumptions, the data collection problem decouples, reducing to a problem of finding an edge set on the graph (with a fixed number of edges) such that the second eigenvalue of the graph Laplacian is maximal. This reduction of the data collection problem to a spectral graph-theoretic question is one of the primary contributions of this work. As an application, we study the Yahoo! Movie user rating dataset and demonstrate that the addition of a small number of well-chosen pairwise comparisons can significantly increase the Fisher informativeness of the ranking.

1. Introduction

In large-scale data analysis and information processing problems, it is important to assess the amount of information contained in a given dataset, since this will influence the quality of the solution. It is then natural to consider the data collection process and to question whether more informative data may be collected or how a given dataset can be augmented to improve informativeness. In this paper, we investigate these questions for the statistical ranking problem.[1]

The problem of statistical ranking arises in a variety of applications, where a collection of alternatives is to be ranked based on pairwise comparisons. Methods for ranking must address a number of inherent difficulties including (i) incomplete data, (ii) inconsistencies in the data, and (iii) imbalanced data. Despite and possibly as a consequence of these difficulties, although ranking from pairwise comparison data is an old problem (David, 1963), there have been several recent contributions to the subject (Langville & Meyer, 2012; Osting et al., 2012b; Hirani et al., 2011; Jiang et al., 2010; Callaghan et al., 2007) with broad applications, *e.g.*, problems in social networking, game theory, and e-commerce.

Let $G = (V, E)$ denote the complete graph, consisting of a set of n nodes, $V = \{j\}_{j=1}^{n}$, representing alternatives, and edges $E = \{k\}_{k=1}^{N}$, where $N := \binom{n}{2} = \frac{n(n-1)}{2}$. Pairwise comparison data consists of (1) a vector $w \in \mathbb{Z}_+^N$ which assigns to each edge $k = ij \in E$ the number of times item i and item j have been compared and (2) a vector $y \in \mathbb{R}^N$ which assigns to each edge $ij \in E$ a quantitative (cardinal) preference of i over j. If the two items, i and j have not been compared, we simply set $w_{ij} = y_{ij} = 0$. When viewed as a statistical inverse problem, the ranking problem is to estimate the overdetermined parameter $\phi \in \mathbb{R}^n$

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

[1] To prevent confusion, we note that the term ranking is used here to indicate a scalar measure of desirability for each item in a collection, sometimes referred to as a rating.

which describes the "desirability" of each alternative given the pairwise comparison data (w, y). The least squares estimator for the ranking problem is defined

$$\hat{\phi}_w = \arg\min_{\phi} \|B\phi - y\|_w, \tag{1}$$

where B is the arc-incidence matrix for G (defined in §2), $\|\cdot\|_w$ denotes the w-weighted ℓ^2 norm, and the dependence of the ranking, $\hat{\phi}_w$, on the available data, w, is emphasized by the subscript.

Generally speaking, the more pairwise comparisons among a fixed number of alternatives, the more informative we expect the ranking, $\hat{\phi}_w$ to be. In this paper, we consider the following question:

> Given a pairwise comparison dataset, (w_0, y_0), and the opportunity to collect ξ additional pairwise comparisons, which data should be targeted to maximally improve the informativeness of the least squares ranking?

We follow the methodology of the optimal design community (Pukelsheim, 2006; Haber et al., 2008), and consider the *Fisher information* for the ranking estimate $\hat{\phi}_w$, defined in (1). Since $\hat{\phi}_w$ is an unbiased estimator, *i.e.*, $\mathbb{E}\hat{\phi}_w = \phi$, the Fisher information is the inverse of the covariance matrix, $\mathrm{Var}(\hat{\phi}_w)$, and thus maximizing the informativeness of the ranking is equivalent to minimizing $\mathrm{Var}(\hat{\phi}_w)$. We are thus led to the following bi-level optimization problem:

$$\min_{w} \quad \|\mathrm{Var}(\hat{\phi}_w)\|_2 \tag{2a}$$

$$\text{such that} \quad \hat{\phi}_w = \arg\min_{\phi} \|B\phi - y\|_w \tag{2b}$$

$$w \in \mathbb{Z}_+^N, \quad w \succeq w_0, \quad \|w - w_0\|_1 \leq \xi. \tag{2c}$$

The constraint in (2c) specifies that only a limited amount of additional data is collected. Choosing to minimize the matrix norm of $\mathrm{Var}(\hat{\phi}_w)$, is referred to as the E-optimal design. Other scalar function choices commonly used as optimal design criteria are $\mathrm{tr}[\mathrm{Var}(\hat{\phi}_w)]$ and $\det[\mathrm{Var}(\hat{\phi}_w)]$ (respectively, A- and D-optimal conditions).

In §3, we show that for the least squares estimator, the constraint (2b) in the optimization problem (2) decouples, yielding a problem of finding edge weights w for which the w-weighted graph Laplacian has maximal second eigenvalue. In §4, we apply our methods to the Yahoo! Movie user rating dataset and demonstrate that the addition of a small number of well-chosen pairwise comparisons can significantly increase the Fisher informativeness of the ranking.

Related work. Optimal experiment design (Pukelsheim, 2006) is commonly used in inverse problems, *e.g.*, geophysical (Haber et al., 2008) and biomedical imaging (Quinn & Keough, 2002; Seeger & Nickisch, 2011; Chung & Haber, 2012), for the purpose of reducing the amount of data that must be collected (sparsity). In particular, for inverse problems in imaging, Seeger & Nickisch (2011) connect the areas of Bayesian active learning and optimal experimental design, collectively referred to as Bayesian sequential experimental design. The present work is most similar to that found in Osting et al. (2012a), where the methodology of optimal experimental design was applied to the problem of sports scheduling. The primary difference is that the present work focuses on (adaptively) improving a pairwise comparison dataset, whereas the focus of the work of Osting et al. (2012a) is the (static) construction of a dataset. Also, in sports scheduling, it is of interest to restrict to the case $w \in \{0,1\}^N$, rather than the more general case $w \in \mathbb{Z}_+^N$, considered here.

In Jamieson & Nowak (2011) and Ailon (2012), the problem of optimally sampling preference labels is studied for the minimum feedback arc-set in weighted tournaments (MFAST) also known as Kemeny-Young ranking. The dataset considered is ordinal, *i.e.*, only pairwise preference labels are specified, whereas in the present work, the dataset is cardinal, *i.e.*, the preferences are represented as quantitative (real valued) differences between items.

Our approach also differs from that of Glickman (2005), where a Bayesian optimal design approach is used to maximize the expected gain in Kullback-Leibler information from the prior to posterior ranking distributions. We do not assume a prior on the statistical ranking.

We show in §3 that the bilevel optimization problem in (2) is related to the problem of finding multigraphs with large algebraic connectivity. There is a tremendous amount of work on the algebraic connectivity of graphs, originating with Fiedler (1973). The robustness of a network to node/edge failures is highly dependent on the algebraic connectivity of the graph. The rate of convergence of a Markov process on a graph to the uniform distribution is determined by the algebraic connectivity (Sun et al., 2004). Finally, in the "chip-firing game" of Björner, Lovász and Shor, the algebraic connectivity dictates the length of a terminating game (Björner et al., 1991). Consequently, algebraic connectivity is a performance measure for the convergence rate in sensor networks, data fusion, load balancing, and consensus problems (Olfati-Saber et al., 2007).

Outline. In §2, we review spectral properties of the w-weighted graph Laplacian. In §3, the data collection problem (2) is shown to be equivalent to a spectral graph problem. In §4, the results from §3 are applied to study the Yahoo! Movie user rating dataset. We conclude in §5 with a discussion.

2. Eigenvalues of the w-weighted graph Laplacian

In this section, we review properties of the eigenvalues of the w-weighted graph Laplacian; more extensive treatments are given in (Mohar, 1991; Chung, 1997; Biyikoglu et al., 2007).

Let $G = (V, E)$ be the complete graph with $|V| = n$ nodes and let $B \in \mathbb{R}^{N\times n}$ where $N := \binom{n}{2}$ be an arc-vertex incidence matrix for G,

$$B_{k,j} = \begin{cases} 1 & j = \text{head}(k) \\ -1 & j = \text{tail}(k) \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where the orientation of the edges is chosen arbitrarily. Given an edge-weight $w \in \mathbb{Z}_+^N$, the w-weighted graph Laplacian is defined

$$\Delta_w := B^t W B \qquad \text{where } W = \text{diag}(w).$$

The w-weighted degree vector $d \in \mathbb{R}^n$ is defined by $d_i = \sum_j w_{ij}$. Let $M := \|w\|_1 = \frac{1}{2}\|d\|_1$ and d_+ and d_- denote the maximum and minimum w-weighted degrees in the graph. Let $\lambda_i(w)$ for $i = 1, \ldots, n$ denote the eigenvalues of Δ_w. The following are properties of the spectrum of Δ_w:

1. Since Δ_w is symmetric and positive definite, the eigenvalues are real and nonnegative. The spectrum is also bounded above, *e.g.*, $\lambda_n \le 2d_+$.

2. Let E be the edgeset corresponding to the indicator function of w. The second eigenvalue,

$$\lambda_2(w) = \min_{\substack{\|v\|=1 \\ \langle v,1\rangle = 0}} \|Bv\|_{2,w}, \tag{4}$$

is nonzero if and only if the graph (V, E) is connected.

3. Let $\Delta_w v = \lambda v$, $\lambda > 0$, $w_0 = \min_k w_k$ and $w' = w - w_0$. Then

$$\Delta_{w'} v = (\lambda - w_0 n)v. \tag{5}$$

This follows from the fact that $B^t B = n\,\text{Id} - 1_n 1_n^t$.

4. The first eigenvalue of Δ_w, λ_1, is zero with corresponding eigenvector $v_1 = 1$.

5. The function $\lambda_2(w)$ is non-decreasing in w, *i.e.*, if $w_1 \le w_2$, then $\lambda_2(w_1) \le \lambda_2(w_2)$.

6. For $U \subset V$, define $C(U) := \sum_{i\in U, j\in U^c} w_{ij}$. The second eigenvalue is bounded above:

$$\lambda_2(w) \le \min_{U\subseteq V} \frac{n\ C(U)}{|U||U^c|}. \tag{6}$$

In particular, if $U = \{v\}$ where $v \in V$ is the node with smallest w-weighted degree, i.e., $d_v = d_-$, then $d_v \le \frac{2M}{n}$ and from (6), we obtain

$$\lambda_2(w) \le \frac{n\ d_-}{n-1} \le \frac{2M}{n-1}. \tag{7}$$

7. Consider the weight $w = w_0 + \delta_k$ where δ_k is the indicator function for edge k. Then using Weyl's theorem (Horn & Johnson, 1990), we obtain

$$\begin{aligned} \lambda_2(w) &\le \lambda_2(w_0) + \|B^t \text{diag}(\delta_k) B\| \\ &= \lambda_2(w_0) + 2. \end{aligned} \tag{8}$$

8. Let G be a weighted graph with weights $w \in \mathbb{R}^N$ and let G' be the graph with weights $w' = w + \delta_k$, where δ_k is the indicator function for edge k. Denote the eigenvalues of the w and w'-weighted graph Laplacians by λ_j and λ'_j respectively. Then the eigenvalues λ and λ' interlace, *i.e.*,

$$0 = \lambda_1 = \lambda'_1 \le \lambda_2 \le \lambda'_2 \le \lambda_3 \le \ldots \le \lambda_n \le \lambda'_n.$$

Remark 2.1. *If $w \in \{0,1\}^N$, then λ_2 is referred to as the algebraic connectivity for the graph defined by w (Fiedler, 1973). In this case, M is the number of edges present and $C(U)$ can be interpreted as the number of edges connecting U and U^c. Rather than considering a complete graph with edge weights $w \in \mathbb{Z}_+^N$, one may alternatively view the structure (V, E, w) as a multigraph where w_{ij} is the number of edges connecting nodes i and j. In this case, λ_2 can be interpreted as the algebraic connectivity of the multigraph and $C(U)$ as the number of edges connecting U and U^c.*

3. Targeted data collection for the least squares ranking

Let (w_0, y_0) be a pairwise comparison dataset, as defined in §1 and let $\hat{\phi}_w$ be the least squares ranking (1). In this section, we consider the problem, formulated in

(2), of collecting ξ additional pairwise comparisons to optimally improve the Fisher informativeness of $\hat{\phi}_w$.

Assume that each alternative $j = 1, \ldots, n$ has a "ground truth" ranking, ϕ_j. We label each pair of alternatives by $k = 1, \ldots, \binom{n}{2} \equiv N$ and denote by $B \in \mathbb{R}^{N \times n}$ the arc-vertex incidence matrix for the complete graph, (3). We assume that there exists a pairwise comparison measure, y, such that for each pair of alternatives $\{i, j\}$,

$$y_{ij} = (B\phi)_{ij} + \epsilon_{ij} \tag{9}$$

where $\epsilon \in \mathbb{R}^N$ is a random vector with zero mean, *i.e.*, $\mathbb{E}\epsilon = 0$. Let $w_k \in \mathbb{Z}_+$ denote the number of comparisons between items i and j. We assume that the variance of ϵ_k is given by σ^2/w_k for some constant σ if $w_k \neq 0$ and zero if $w_k = 0$. The variance in the observed comparison is reduced as the number pairwise comparisons increases. The following proposition shows that the optimal data collection problem (2) is equivalent to the problem of finding edge weights, w, for which the w-weighted graph Laplacian has maximal second eigenvalue.

Proposition 3.1. *Let ϵ be a random vector with $\mathbb{E}\epsilon = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2 W^\dagger$ where $W = \mathrm{diag}(w)$ and $w \in \mathbb{Z}_+^N$. Let $\hat{\phi}_w$ be the least squares estimator given in (1) for ϕ in (9). Then the problem of finding a dataset, $w \in \mathbb{Z}_+^N$, with $w \succeq w_0$ and $\|w - w_0\|_1 \leq \xi$ which minimizes $\|\mathrm{Var}(\hat{\phi}_w)\|_2$ is equivalent to the eigenvalue optimization problem*

$$\begin{aligned} &\max_{w} \quad \lambda_2(w) \\ &\text{such that} \quad w \in \mathbb{Z}_+^N, \quad w \succeq w_0, \quad \|w - w_0\|_1 \leq \xi, \end{aligned} \tag{10}$$

where $\lambda_2(w)$ is the second eigenvalue of the w-weighted graph Laplacian, as defined in (4).

Proof. The least squares ranking, written

$$\hat{\phi}_w = \arg \min_{\langle \phi, 1 \rangle = 0} \|B\phi - y\|_{2,w} = (B^t W B)^\dagger B^t W y,$$

is a linear, unbiased ($\mathbb{E}[\hat{\phi}_w] = \phi$) estimator. Here, the dagger (†) is the Moore-Penrose pseudo-inverse. We first compute

$$\begin{aligned} \hat{\phi}_w &= (B^t W B)^\dagger B^t W y = (B^t W B)^\dagger B^t W (B\phi + \epsilon) \\ &= \phi + (B^t W B)^\dagger B^t W \epsilon. \end{aligned}$$

Thus, the variance of $\hat{\phi}_w$ is given by

$$\begin{aligned} \mathrm{Var}(\hat{\phi}_w) &= \mathbb{E}\left[(\hat{\phi}_w - \phi)(\hat{\phi}_w - \phi)^t\right] \\ &= (B^t W B)^\dagger B^t W \mathbb{E}\left[\epsilon \epsilon^t\right] W B (B^t W B)^\dagger. \end{aligned}$$

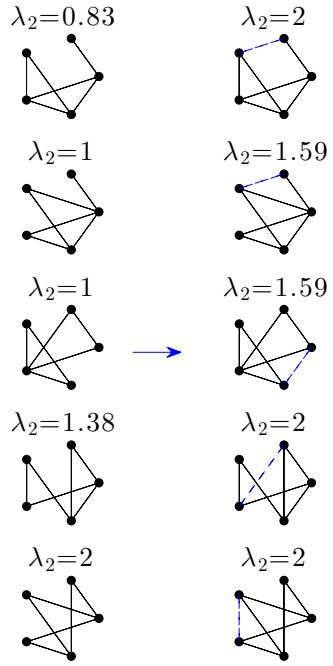


Figure 1. Targeted data collection for small graphs. **(left)** The five topologically distinct connected graphs with $n = 5$ nodes and $m = 6$ edges. **(right)** For each edgeset on the left, we select one additional edge (blue dashes) so that λ_2 for the enhanced graph is maximal. The algebraic connectivity of each graph is indicated. By Prop. 3.1, a ranking on a dataset represented by a graph on the right is more informative than one from a graph on the left. See §3.

Assuming $\mathbb{E}\left[\epsilon \epsilon^t\right] = \sigma^2 W^\dagger$, we obtain $\mathrm{Var}(\hat{\phi}_w) = \sigma^2 (B^t W B)^\dagger \equiv \sigma^2 \Delta_w^\dagger$. Since $\mathrm{Var}(\hat{\phi}_w)$ doesn't depend on the pairwise comparison data, y, the constraint in the optimal data collection problem (2b) decouples. Furthermore, minimizing $\|\mathrm{Var}(\hat{\phi}_w)\|$ is equivalent to maximizing the smallest non-zero eigenvalue of the w-weighted graph Laplacian, $\Delta_\omega = B^t W B$. Provided each item has been compared to at least one other item, the smallest non-zero eigenvalue of Δ_ω is the second one, $\lambda_2(w)$, as defined in (4). □

In Fig. 1, we illustrate Prop. 3.1 by studying (10) where $\|w_0\|_1 = 6$ and $\xi = 1$. Although the graphs in Fig. 1 are small in size, it is already nontrivial to determine which edge should be added to maximally increase the algebraic connectivity. We observe that for graphs with low algebraic connectivity, a significant gain can be achieved, while the results for graphs with relatively high algebraic connectivity are modest. In the lowermost panel in Fig. 1, the algebraic connectivity remains constant as an edge is added. This follows from the fact that the second eigenvalue for the

Algorithm 1 A greedy heuristic for finding edge weights w for which the w-weighted Laplacian has large second eigenvalue (Ghosh & Boyd, 2006b; Wang & Mieghem, 2008). See §2.

Input: An initial edge weight $w_0 \in \mathbb{R}^N$ defined on the complete graph of n nodes and an integer, ξ.

Output: An edge weight, w, such that $\|w - w_0\|_1 = \xi$, and Δ_w has large second eigenvalue.

Set $w = w_0$ (current edge weight)
for $\ell = 1$ **to** ξ, **do**
  Compute the second eigenvector,

$$F = \arg \min_{\substack{\|v\|=1 \\ \langle v,1\rangle=0}} \|Bv\|_w$$

  Find the edge ij which maximizes $(F_i - F_j)^2$
  Set $w = w + \delta_{ij}$
end for

graph on the left has multiplicity 2 and the interlacing property described in §2.

The problem of finding weights $w \in \mathbb{R}^N$ which maximize $\lambda_2(w)$ is a convex optimization problem and can be formulated as a semidefinite program (SDP) (Ghosh & Boyd, 2006a;b). However, for $w \in \mathbb{Z}_+^N$, as in (10), this problem is NP-hard (Mosk-Aoyama, 2008). Solutions to the integer constrained problem may be approximated by solving the unconstrained problem and rounding the solution. This is clearly a lower bound on the optimal solution and, if the values w are large, a reasonable approximation. Another approach, advocated by Ghosh & Boyd (2006b) and Wang & Mieghem (2008), is the greedy algorithm described in Algorithm 1. This algorithm uses the second (Fiedler) eigenvector to iteratively chose an edge for which to increment the corresponding entry of w by one. Solutions generated using this heuristic are found to be adequate for the present work.

Remark 3.2. *From (5), it is tempting to think that only edges with the smallest current weight must be considered in Algorithm 1. However, graphs with $n = 6$ nodes can be generated such that $w = 0$ on some edges and there exists an edge k such that (a) $w_k = 1$, (b) among all edges, $\lambda_2(w)$ increases most significantly when edge k is incremented, and (c) the greedy heuristic in Algorithm 1 selects edge k to increment.*

4. Informativeness of the ranking for the Yahoo! Movie user ratings dataset

In this section, we apply the methodology formulated in §3, to study the Fisher informativeness of the Yahoo! Movie user rating dataset. We show that the addition of targeted edges can significantly improve the informativeness of the movie rating system.

The dataset. The Yahoo! Movie user rating dataset consists of a $7,642 \times 11,915$ user-movie matrix where each of the $211,197$ nonzero entries (0.23% sparsity density) is a 1 to 13 rating (yah)[2]. Each movie was rated by between 1 and 4,238 users (the average number of reviews per movie is 17.7). Each user rated between 10 and 1,632 movies (the average number of reviews made by each reviewer is 27.6). Of the 70,977,655 (movie) pairs (i, j) where $i > j$, there are 5,742,557 for which a user has given a rating to both movies i and j implying that the pairwise comparisons for the raw dataset are 8.1% complete. The majority of movies in the dataset received relatively few reviews, as reported in Table 1. The movies which received less than 10 rankings were discarded from the dataset, leaving 2,367 movies, each of which were reviewed by an average of 79.8 users. We then removed 11 users who did not review any of the remaining movies. The remaining 7,631 reviewers reviewed between 1 and 1,220 movies (on average they reviewed 24.8 movies).

Construction of pairwise comparison data from movie-user rating data. Let Σ be the set of Yahoo! users, V be the set of all Yahoo! movies and r_i^σ be the rating given to movie $i \in V$ by user $\sigma \in \Sigma$. For each unordered movie pair $\{i, j\} \in V^2$, we define

$$\Sigma_{ij} = \{\sigma \in \Sigma \text{ who rated both movies } i \text{ and } j\}.$$

For each movie pair $\{i, j\} \in V^2$, we define w_{ij} to be the number of users who have viewed both movies i and j, *i.e.*, $w_{ij} = |\Sigma_{ij}|$ and y_k to be the average difference in movie reviews, written

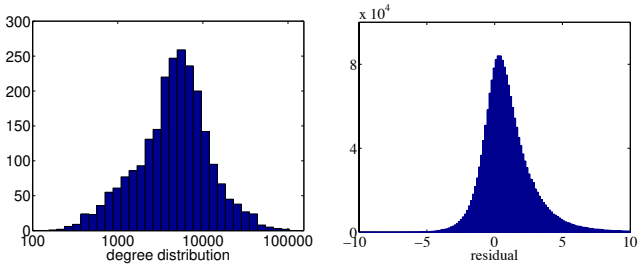
$$y_{ij} = \frac{1}{|\Sigma_{ij}|} \sum_{\sigma \in \Sigma_{ij}} (r_j^\sigma - r_i^\sigma), \text{ where } \{i, j\} \in V^2 \text{ and } i < j. \tag{11}$$

Note that the expression in parenthesis is anti-symmetric in the indices i and j and lies in the interval $[-12, 12]$. The choice $i < j$ corresponds to the choice in arc direction in (3). For the Yahoo!

[2] 34 entries reviewing Yahoo! movie_id 0 were discarded due to absence in movie content description file.

Table 1. Frequency of reviews for items in the Yahoo! Movie user rating dataset.

# times movie reviewed	1	2	3	4	5	6	7	8	9	≥ 10
occurrences	4,901	1,882	897	548	398	316	237	202	167	2,367



$\hat{\phi}_w$	Movie Name
4.46	It's a Wonderful Life (1946)
4.45	Singin' in the Rain (1952)
4.34	Rear Window (1954)
4.11	24: Season 1 (2002)
3.96	The Longest Day (1962)
3.94	The Man Who Shot Liberty Valance (1962)
3.92	Rebecca (1940)
3.87	Friends - The Complete Fourth Season (1997)
3.79	Lady and the Tramp (1955)
3.79	It Happened One Night (1934)

Figure 2. **(top left)** A log-histogram of the w-weighted degree distribution for the graph representing the Yahoo! movie pairwise comparison data. **(top right)** A histogram of the residual, $y - B\hat{\phi}_w$, where $\hat{\phi}_w$ is the least squares ranking. **(bottom)** Top 10 movies and ranking, $\hat{\phi}_w$. See §4.

Movie user rating dataset, we have $n := |V| = 2,367$, $N := \binom{n}{2} = 2,800,161$, $m := \|w\|_0 = 1,884,504$, and $M := \|w\|_1 = 8,322,538$. Thus, there exists at least one comparison for $m/N = 67\%$ of the movie pairs. The mean w-weighted degree of each node is given by $2 \cdot M/n = 3,516$. A log-histogram of the w-weighted degree distribution of the graph representing the pairwise comparison data is given in Fig. 2 (top left).

The least squares ranking. A ranking is obtained by solving the least squares problem, (1), using Matlab's `lsqr` function. The top ten movies found are given in Figure 2. The relative residual norm of the least squares estimator, $\hat{\phi}_w$, is $\frac{\|B\hat{\phi}_w - y\|_w}{\|y\|_w} = 0.53$. In Fig. 2 (top right), we plot a histogram of the residual, $y - B\hat{\phi}_w$. For this pairwise comparison dataset, the normality assumption in Prop. 3.1 is reasonable.

The informativeness of the ranking is $\lambda_2(w) = [\mathrm{Var}(\hat{\phi}_w)]^{-1} = 154.38$. This value is small compared to the upper bound given in (7), $\lambda_2(w) \leq \frac{2M}{n-1} = 7,036$. We next demonstrate that the Fisher information can be significantly improved by the addition of a small number of targeted pairwise comparisons.

Targeted data collection. We apply the optimal experimental design approach developed in §3 to improve the Fisher information of the least squares ranking. To approximate the solution of (10), we use the greedy algorithm described in Algorithm 1. The second eigenpair of the graph Laplacian is computed using Matlab's `eigs` function, initialized using the eigenvector from the previous iteration. We choose a very modest value of pairwise comparison edges to add, $\xi = .01\% \cdot M = 832$ edges. The results are given in Fig. 3. The addition of the targeted pairwise comparisons leads to an increase in the second eigenvalue of the w-weighted graph Laplacian by a factor of 2.2. The maximum increase for the addition of a single pairwise comparison is ≈ 1, less than the upper bound given in (8). We observe in Fig. 3, that the rate of information increase slows as more pairwise comparisons are added. For a comparison, we also consider the addition of randomly chosen movie pairs. For this modest value of additional edges, ξ, the effect of the informativeness of the ranking is unappreciable.

Finally, we use graph visualization via spectral clustering to illustrate the pairwise comparison and targeted data. In Fig. 4(top) we plot the given pairwise movie comparisons obtained from the Yahoo! user-movie database. In Fig. 4(bottom) we plot the proposed pairwise comparisons, targeted to improve the informativeness of the rating system. To enhance the readability of the graph representation, we plot only 15% randomly selected nodes (356 of $n = 2367$) and the interconnecting edges ($45,327$ of $m = 1,884,504$). Figure 4(top) was generated as follows. First normalized spectral clustering (based on k-means) was used to detect clusters of movies. Next, the Fruchterman-Reingold algorithm was used to generate reasonable positions for the movie clusters and the Kamada-Kawai algorithm was used to place movies within the clusters (Traud et al., 2009). The node placement was obtained using the full dataset. Finally, the weighted graphs were plotted using `wgPlot` (Wu, 2009). Figure 4(bottom) was then generated using the same node placements as in Figure 4(top).

A comparison of the top and bottom panels of Fig. 4 shows that the primary improvement to informativeness arises from the addition of edges which con-

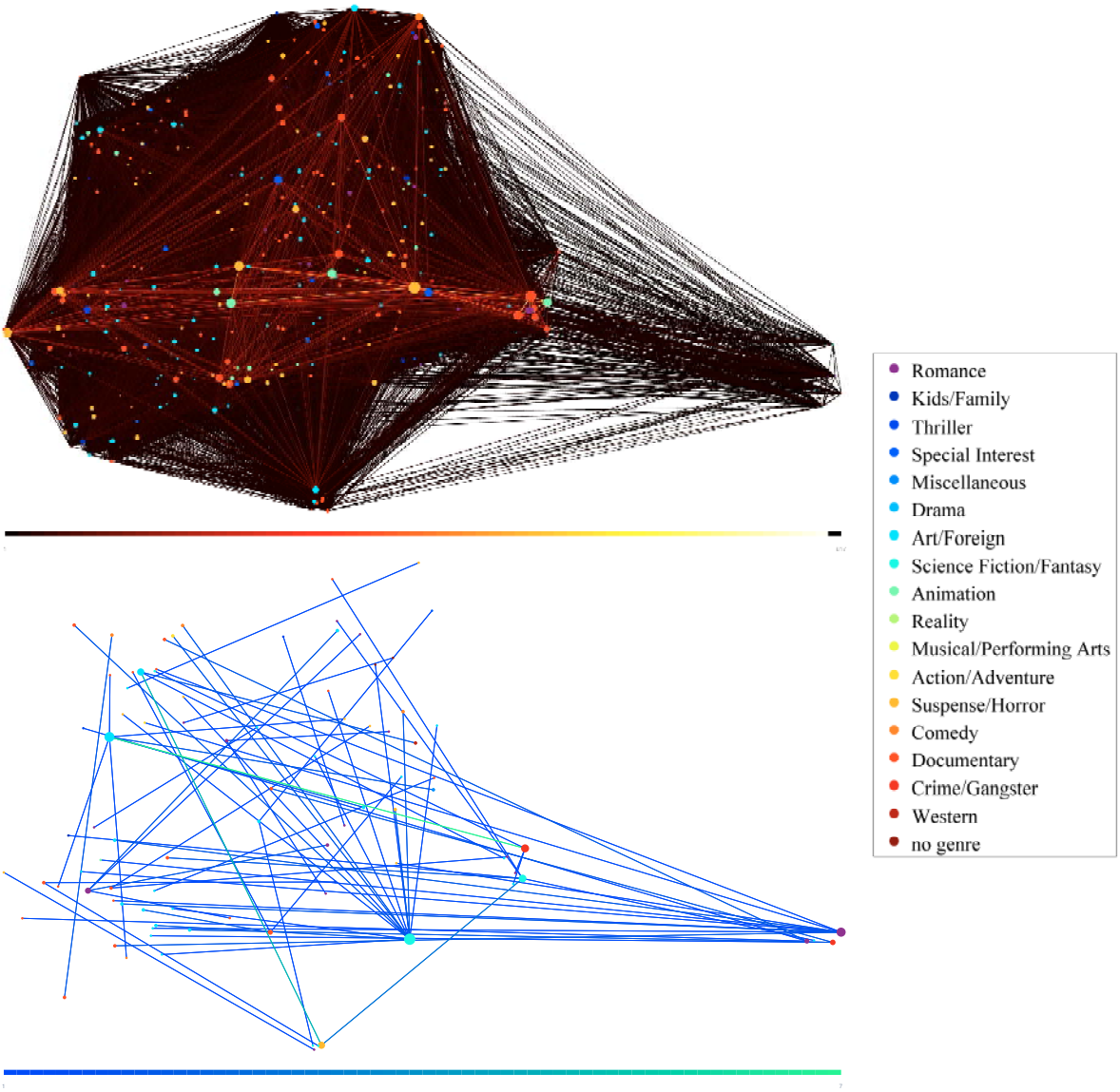
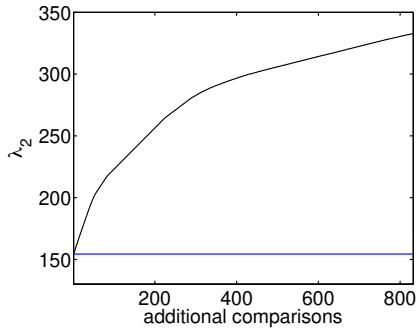


Figure 4. **(top)** A 15% randomly chosen subset of the pairwise comparison graph for the Yahoo! user-movie database. Nodes represent movies, node size reflects weighted degree (*i.e.*, number of comparisons with other movies), and node color indicates genre (see legend). Edges represent weighted pairwise comparisons colored by edge weights (*i.e.*, number of comparisons). **(bottom)** Pairwise comparisons targeted for collection to improve the informativeness of the least squares ranking. Targeted comparisons are colored by weight (multiplicity). See §4.

nects two relatively weakly connected components of the graph (see Remark 2.1). With 4 exceptions, each targeted movie pair is only incremented once; it isn't generally advantageous to add a pair multiple times (see Remark 3.2).

5. Discussion and future directions

We have applied methods from optimal experiment design to provide a new framework for targeted data collection for more informative rankings. At the heart of this framework is a bi-level optimization problem (2) where the inner problem is to determine the least-squares estimate for the ranking and the outer problem is to identify the data which minimizes the variance of the ranking. We show that for a Gaussian error model (9), the outer and inner problems decouple, yielding a problem of finding an edgeweight $w \in \mathbb{Z}_+^N$, such that the w-weighted graph Laplacian has small second eigenvalue (10). This can be interpreted as finding a multigraph with large algebraic connectivity.



method of data collection	additional comparisons original $\xi = 0$	.005% M $\xi = 416$	.01% M $\xi = 832$
random	154.38	154.38	154.38
optimal	154.38	**298.44**	**332.78**
upper bound (7)	7,035	7,035	7,036

Figure 3. **(top)** The informativeness of the ranking, $\lambda_2(w)$, as a small number ($.01\% \cdot M$) of targeted pairwise comparisons (black) and randomly selected pairwise comparisons (blue) are added. **(bottom)** The value of $\lambda_2(w)$ for this augmented dataset and the upper bound on λ_2 given in (7). The change in informativeness for randomly added data is unappreciable compared to a 2.2 fold increase for targeted data. See §4.

There are several applications in, *e.g.*, social networking, game theory, and e-commerce, where improved data collection could potentially benefit ranking. In particular, for the Yahoo! Movie user ratings dataset (considered in §4), we have shown that the informativeness of ranking can be increased by a factor of 2.2 if just .01% of additional, optimally-targeted data is added to the dataset. In contrast, if the same amount of random data is added, there is only a very small effect on the informativeness of the ranking.

In this paper, we have focused on targeted data collection for improved rankings, neglecting several important factors including the cost of data collection and potential constraints on what data may be collected. There are two simple extensions to our method which may be employed to accommodate these additional factors. The cost of data collection could be incorporated by either adding a penalization term in (10) or by incorporating additional weights into the norm used to compute λ_2 in (10). Data collection constraints may be handled by explicitly forbidding certain edge weights to be incremented in the greedy Algorithm 1 for targeting data collection.

The least-squares ranking estimate (1) is referred to as HodgeRank by some authors (Jiang et al., 2010; Xu et al., 2011), since the Hodge decomposition implies that the residual in (1), $r = B\phi - y$, can be further decomposed into two orthogonal components: (1) a divergence-free component which consists of 3-cycles and (2) a harmonic component which consists of longer cycles (Jiang et al., 2010; Hirani et al., 2011). In fact, Jiang et al. (2010) argues that a dataset which has a large harmonic component is inherently inconsistent and does not have a reasonable ranking. The harmonic component lies in the kernel of the graph Helmholtzian with dimension given by the first Betti number of the associated simplical complex. Optimal reduction of the first Betti number may provide an alternative approach to improving the informativeness of the least squares ranking.

Finally, we mention two extensions of the present work. It would be interesting to consider optimal data collection for nonlinear ranking methods, including robust estimators (Osting et al., 2012b), random walker methods (Callaghan et al., 2007), Perron-Frobenius eigenvalue methods (Keener, 1993; Langville & Meyer, 2012), and Elo methods (Elo, 1978; Glickman, 1995; Langville & Meyer, 2012). Secondly, the implementation of a data collection method should be more carefully modeled for particular applications. For instance, for the Yahoo! Movie user rating dataset, the pairwise comparison data is constructed from user rating data and thus any targeted pairwise comparison addition must be solicited from a user. Since the number of pairwise comparisons for which a particular reviewer adds when a new movie is reviewed is equal to the number of previous reviews that user has contributed, it may make sense to solicit additional reviews from users with many previous reviews. That is, one must consider the propagation of information from the user reviews to the pairwise comparison data in (11).

Acknowledgments

We thank Lawrence Carin, Jérôme Darbon, Mark L. Green, and Yuan Yao for useful discussions. B. Osting is supported by NSF DMS-1103959. C. Brune is supported by ONR grants N00014-10-10221 and N00014-12-10040. S. Osher is supported by ONR N00014-08-1-1119, N00014-10-10221, and NSF DMS-0914561.

References

Yahoo! Webscope dataset: ydata-ymovies-user-movie-ratings-content-v1_0. http://webscope.sandbox.yahoo.com. accessed: 10/5/2011.

Ailon, N. An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity. *J. Mach. Learn. Res.*, 13:137–

164, 2012.

Biyikoglu, T., Leydold, J., and Stadler, P. F. *Laplacian Eigenvectors of Graphs*. Springer, 2007.

Björner, A., Lovász, L., and Shor, P. W. Chip-firing games on graphs. *European J. Combin*, 12(4), 1991.

Callaghan, T., Mucha, P. J., and Porter, M. A. Random walker ranking for NCAA Division I-A football. 2007.

Chung, F. R. K. *Spectral Graph Theory*. AMS, 1997.

Chung, M. and Haber, E. Experimental design for biological systems. *Siam J. Control Optim.*, 50(1): 471–489, 2012.

David, H. A. *The Method of Paired Comparisons*. Charles Griffin & Co., 1963.

Elo, A. *The Rating of Chessplayers, Past and Present*. Arco Pub., 1978.

Fiedler, M. Algebraic connectivity of graphs. *Czech. Math. J.*, 23:298–305, 1973.

Ghosh, A. and Boyd, S. Upper bounds on algebraic connectivity via convex optimization. *Linear Algebra and its Applications*, 418:693–707, 2006a.

Ghosh, A. and Boyd, S. Growing well-connected graphs. *Proc. IEEE Conf. Decision & Control*, 2006b.

Glickman, M. E. A comprehensive guide to chess ratings. *American Chess Journal*, 3, 1995.

Glickman, M. E. Adaptive paired comparison design. *Journal of Statistical Planning and Inference*, 127: 279–293, 2005.

Haber, E., Horesh, L., and Tenorio, L. Numerical methods for experimental design of large-scale linear ill-posed inverse problems. *Inverse Problems*, 24: 055012, 2008.

Hirani, A. N., Kalyanaraman, K., and Watts, S. Least squares ranking on graphs. arXiv:1011.1716v4, 2011.

Horn, R.A. and Johnson, C.R. *Matrix Analysis*. Cambridge University Press, 1990.

Jamieson, K. G. and Nowak, R. D. Active ranking using pairwise comparisons. In *Neural Information Processing Systems (NIPS)*, pp. 2240–2248, 2011.

Jiang, X., Lim, L.-H., Yao, Y., and Ye, Y. Statistical ranking and combinatorial Hodge theory. *Math. Program. Ser. B*, 127(1):203–244, 2010.

Keener, J. P. The Perron-Frobenius theorem and the ranking of football teams. *SIAM Review*, 35(1):80–93, 1993.

Langville, A. N. and Meyer, C. D. *Who's #1?: The Science of Rating and Ranking*. Princeton University Press, 2012.

Mohar, B. The Laplacian spectrum of graphs. In *Graph Theory, Combinatorics, and Applications*, volume 2, pp. 871–898. Wiley, 1991.

Mosk-Aoyama, D. Maximum algebraic connectivity augmentation is NP-hard. *Operations Research Letters*, 36(6):677–679, 2008.

Olfati-Saber, R., Fax, A., and Murray, R. M. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.

Osting, B., Brune, C., and Osher, S. Optimal data collection for improved rankings expose well-connected graphs. submitted, 2012a.

Osting, B., Darbon, J., and Osher, S. Statistical ranking using the ℓ^1-norm on graphs. submitted, 2012b.

Pukelsheim, F. *Optimal Design of Experiments*. SIAM, 2006.

Quinn, G. P. and Keough, M. J. Experimental design and analysis for biologists. *Cambridge University Press, Cambridge, UK*, 2002.

Seeger, M. and Nickisch, H. Large scale Bayesian inference and experimental design for sparse linear models. *SIAM Journal of Imaging Sciences*, 4(1):166–199, 2011.

Sun, J., Boyd, S., Xiao, L., and Diaconis, P. The fastest mixing Markov process on a graph and a connection to a maximum variance unfolding problem. *SIAM Review*, 48:2006, 2004.

Traud, A. L., Frost, C., Mucha, P. J., and Porter, M. A. Visualization of communities in networks. *Chaos*, 19:041104, 2009.

Wang, H. and Mieghem, P. Van. Algebraic connectivity optimization via link addition. In *Bionetics 2008, Hyogo, Japan*, 2008.

Wu, Mike. wgPlot. http://www.mathworks.com/matlabcentral/fileexchange/24035, 2009.

Xu, Q., Yao, Y., Jiang, T., Huang, Q., Yan, B., and Lin, W. Random partial paired comparison for subjective video quality assessment via HodgeRank. In *ACM Multimedia*, 2011.