Manuscript submitted to
AIMS' Journals
Volume **X**, Number **0X**, XX **200X**

Website: http://AIMsciences.org

pp. **X–XX**

STATISTICAL RANKING USING THE ℓ^1-NORM ON GRAPHS

Braxton Osting[1], Jérôme Darbon[1,2], and Stanley Osher[1]

[1] Department of Mathematics, University of California, Los Angeles
[2] CMLA, ENS Cachan, CNRS, PRES UniverSud

Abstract. We consider the problem of establishing a statistical ranking for a set of alternatives from a dataset which consists of an (inconsistent and incomplete) set of quantitative pairwise comparisons of the alternatives. If we consider the directed graph where vertices represent the alternatives and the pairwise comparison data is a function on the arcs, then the statistical ranking problem is to find a potential function, defined on the vertices, such that the gradient of the potential *optimally* agrees with the pairwise comparisons. Potentials, optimal in the ℓ^2-norm sense, can be found by solving a least-squares problem on the digraph and, recently, the residual has been interpreted using the Hodge decomposition (Jiang *et. al.*, 2010). In this work, we consider an ℓ^1-norm formulation of the statistical ranking problem. We describe a fast graph-cut approach for finding ϵ-optimal solutions, which has been used successfully in image processing and computer vision problems. Applying this method to several datasets, we demonstrate its efficacy at finding solutions with sparse residual.

1. **Introduction.** We consider the statistical rank aggregation problem of establishing an *optimal* statistical ranking for a set of alternatives from a dataset which consists of (i) an inconsistent and incomplete set of quantitative pairwise comparisons of the alternatives and (ii) a set of weights, each associated with a comparison. Inspired by applications in machine learning, social networking, and competitive sports, we focus on the statistical ranking problem for large data sets consisting of cardinal or rated data (as opposed to ordinal or binary datasets). A precise formulation of the rank aggregation problem requires a decision on how to resolve inconsistencies within the data, *i.e.*, the measure of "optimal." Our method seeks solutions for which the error is as sparse as possible.

In what follows, we specify assumptions on the dataset, formulate the rank aggregation problem as an optimization problem, and summarize the results of this paper.

Dataset assumptions. We assume a dataset consisting of the following:

1. A weakly connected[1] directed graph (weak digraph), $D = (V, A)$, consisting of a set of alternatives $V = \{j\}_{j=1}^n$ and a (possibly incomplete) set of ordered alternative pairs, $A = \{k\}_{k=1}^m$, such that arc $k \in V \times V$.

2000 *Mathematics Subject Classification.* 62F07, 65F10, 05C20, 58A14, 05C85.

Key words and phrases. statistical ranking, rank aggregation, Kemeny-Snell ordering, HodgeRank, ℓ^1-norm minimization, graph-cut method.

[1] A digraph is *weakly connected* if replacing its arcs with undirected edges yields a connected graph. If D is disconnected, each weakly connected component is ranked separately.

2. A set of non-negative[2] *pairwise comparisons*, $\{y_k\}_{k=1}^m$, assigning a "degree of preference" to each alternative pair $k \in A$. If $k = ij \in A$, then alternative j is preferred to alternative i as measured by y_k.
3. A set of positive *weights*, $\{w_k\}_{k=1}^m$, where each weight w_k is associated with an arc $k \in A$. A large value of w_k implies, *e.g.*, a high confidence in the pairwise comparison of pair $k \in A$.

We are particularly interested in large datasets ($|V| = n \gg 1$) which are incomplete, *i.e.*, $m = |A| < \binom{n}{2}$.

Digraph operators. We define three operators on the digraph [17]. Let $\mathrm{grad} \in \mathbb{R}^{m\times n}$ be the *arc-vertex incidence matrix* for the digraph $D = (V, A)$,

$$(\mathrm{grad})_{k,j} = \begin{cases} 1 & \text{if } k \in A \text{ and } j = \mathrm{head}(k) \\ -1 & \text{if } k \in A \text{ and } j = \mathrm{tail}(k) \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

We refer to the application of grad to a vector $\phi \in \mathbb{R}^n$ as the *gradient* of ϕ. Define the *divergence*, $\mathrm{div}_w \colon \mathbb{R}^m \to \mathbb{R}^n$,

$$[\mathrm{div}_w\ x]_j := \sum_{k\colon\, j=\mathrm{tail}(k)} w_k x_k \quad - \sum_{k\colon\, j=\mathrm{head}(k)} w_k x_k, \qquad j = 1, \dots, n \tag{2}$$

so that grad and div_w are negative adjoint with respect to the w-inner product $\langle x, y\rangle_w := \sum_k w_k x_k y_k$, *i.e.*,

$$\langle x, \mathrm{grad}\ \phi\rangle_w = \langle w.x, \mathrm{grad}\ \phi\rangle = \langle \mathrm{grad}^t\ (w.x), \phi\rangle = -\langle \mathrm{div}_w\ x, \phi\rangle.$$

The composition operator, $\Delta \in \mathbb{R}^{n\times n}$, defined

$$\Delta = -\mathrm{div}_w\ \circ\ \mathrm{grad}, \quad \text{with entries} \quad [\Delta]_{ij} = \begin{cases} \sum_{k\ni i} w_k & i = j \\ -w_k & i \sim j,\ k = ij \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

is symmetric and referred to as the *w-weighted graph Laplacian*.

In the following, we assume that the dataset (grad, y, w) is given and postpone the discussion of the construction of y and w from raw data to §5.

Consistency of pairwise comparison data. We make the following definitions:

1. The dataset (grad, y) is *acyclic* if there exists a vertex function $\phi \colon V \to \mathbb{R}$ such that
$$\mathrm{sign}(\mathrm{grad}\ \phi) = \mathrm{sign}(y) = 1_m \quad \text{where} \quad 1_m = [1, \dots, 1]^t \in \mathbb{R}^m. \tag{4}$$
2. The dataset (grad, y) is *globally consistent* if it is the gradient of a vertex function, *i.e.*, there exists a vertex function $\phi \colon V \to \mathbb{R}$ such that
$$\mathrm{grad}\ \phi = y. \tag{5}$$
In this case, the vertex function ϕ is called a *potential*. Since a potential (if one exists) is unique modulo an additive constant, we restrict ϕ to the set
$$\mathfrak{A}\mathfrak{d} := \{\phi \in \mathbb{R}^n \colon 1_n^t \phi = 0\} \quad \text{where} \quad 1_n = [1, \dots, 1]^t \in \mathbb{R}^n, \tag{6}$$
and define the projection of $x \in \mathbb{R}^n$ onto $\mathfrak{A}\mathfrak{d}$ by $P_{\mathfrak{A}\mathfrak{d}} x := x - \frac{1}{n}(1_n^t x) 1_n$.

[2] Due to the skew-symmetry of pairwise comparisons, non-negativity is not restrictive.

The dataset (grad, y) being acyclic is equivalent to the digraph $D = (V, A)$ being acyclic. Note that a potential function induces an *ordering relation* on the alternatives $V = \{v_i\}_{i=1}^n$, via $v_j \succ_\phi v_i$ if $\phi_j > \phi_i$. A potential for an acyclic dataset induces an asymmetric ordering which is transitive. Global consistency of a dataset is a harsh requirement, necessarily requiring the dataset be acyclic. The dimension of the space of edge flows is m, while the dimension of imag(grad) is $n-1$, which, depending on the digraph, could be considerably smaller. Global inconsistencies in pairwise comparison data are commonplace in applications (Condorcet paradox); the datasets studied in this work are not globally consistent. However, globally consistent datasets do appear frequently. For example, a dataset for which D is a directed tree, such as those generated from single-elimination tournaments, are globally consistent and hence acyclic.

Formulation of the statistical ranking problem. The *statistical ranking problem* is to find a potential ϕ for an inconsistent dataset (grad, y, w) such that (5) is approximately satisfied in some sense. This is formulated as the following optimization problem:

$$\min_{\phi \in \mathfrak{Ad}} \; J(\phi), \tag{7}$$

where the objective function $J(\phi)$ is a measure of the misfit in (5) and the admissible set, $\mathfrak{Ad}$, is defined in (6). An advantage of the formulation of the ranking problem as an optimization problem is that a small objective value gives a "performance measure" or "certificate of reliability" for a proposed potential. Of particular interest in this work are objective functions of the form

$$J_p(\phi) := \|\text{grad } \phi - y\|_{p,w}^p = \sum_{k=1}^m w_k \left|(\text{grad } \phi)_k - y_k\right|^p, \tag{8}$$

where $\|\cdot\|_{p,w}$ is the w-weighted ℓ^p-norm for $p \geq 1$. In this paper, we discuss the following four objective functions.

The w-weighted ℓ^2-norm. The optimization problem (7) with objective function given by (8) with $p = 2$,

$$\min_{\phi \in \mathfrak{Ad}} J_2(\phi) := \|\text{grad } \phi - y\|_{2,w}^2, \tag{9}$$

is studied in [26, 22]. Equation (9) has the following properties:

1. The column rank of grad for a weak digraph is $n-1$ with $\ker(\text{grad}) = \text{span}\{e\}$ (see, *e.g.*, [17, p.103]). Thus, the objective function in (9) is strictly convex on $\mathfrak{Ad}$ and (9) has a unique solution. The solution may be obtained by finding the minimal ℓ^2-norm solution to the normal equations
$$\Delta\, \phi = -\text{div}_w\, y. \tag{10}$$
The solution of (10) is given by $\phi = -\Delta^\dagger\, \text{div}_w\, y$ where $\Delta^\dagger$ denotes the Moore-Penrose pseudoinverse of Δ.
2. Defining the *residual*, $r := y - \text{grad } \phi$, which is the obstruction to global consistency in (9), the optimization problem can be rewritten
$$\min_{r \in \mathbb{R}^m} \; \|r\|_{2,w}^2 \tag{11}$$
$$\text{such that } P_{\text{grad}}^\perp\, r = P_{\text{grad}}^\perp\, y$$
where $P_{\text{grad}}^\perp := (\text{Id}_m - P_{\text{grad}})$ is the w-projection onto $\ker(\text{div}_w)$ and $P_{\text{grad}} := \text{grad}\, \Delta^\dagger\, \text{div}_w$ is the w-projection onto imag(grad). Since $r - y \in \text{imag}(\text{grad})$,

ϕ can then be recovered from r in $O(n)$ operations by solving grad $\phi = r - y$. Results from combinatorial Hodge theory provide a further decomposition of $\text{im}(\text{grad})^\perp$ implying that the residual, $r = \text{grad}\ \phi_\star - y$, can be decomposed into locally cyclic (3 vertex cycle) and locally acyclic (≥ 4 vertex cycle) components [26, 22]. The locally cyclical component is given by curl $\Phi_\star$, where

$$\Phi_\star = \arg \min_{\Phi \in V \times V \times V} \|\text{curl}\ \Phi - y\|_{2,w}. \tag{12}$$

The locally acyclic, harmonic component, $h = y - \text{grad}\ \phi_\star - \text{curl}\ \Phi_\star$, [26] argues, is the least desirable component of the data.

3. Generally, the solution of the normal equation (10) requires $O(n^3)$ computations using standard solvers, which is prohibitively expensive for large datasets. Larger datasets may be considered using Krylov iterative and algebraic multigrid methods [22].

The w-weighted Kemeny-Snell objective function. The w-weighted Kemeny-Snell method for statistical ranking solves

$$\min_{\phi \in \mathfrak{A}\mathfrak{d}} J_K(\phi) := \sum_{k=1}^{m} w_k \left|\text{sign}\ (\text{grad}\ \phi)_k - 1\right| \tag{13}$$

where $\text{sign}(x) = \frac{x}{|x|}$ if $x \neq 0$ and 0 otherwise. The objective function $J_K(\phi)$ seeks the ranking with the (weighted) minimum number of edges which violate the acyclicity condition (4). A dataset is acyclic if there exists a potential ϕ such that $J_K(\phi) = 0$. This formulation is natural if one is only interested in an induced ordering of the alternatives and not a quantitative comparison. However, (13) is equivalent to the feedback arc set problem and hence an NP-hard problem [10].

The w-weighted "ℓ^0-norm". The optimization problem (7) with objective function (8) with $p = 0$, is given by

$$\min_{\phi \in \mathfrak{A}\mathfrak{d}} \ J_0(\phi) := \|\text{grad}\ \phi - y\|_{0,w}. \tag{14}$$

The objective function is a semi-norm, lacking positive homogeneity, and measures the w-weighted support of the residual $r = y - \text{grad}\ \phi$. For $w = 1_m$, $J_0(\phi)$ simply minimizes the number of edges for which the data and grad ϕ disagree. In this case, the optimal solution $\phi_\star$ of (14) will have residual $r_\star = y - \text{grad}\ \phi_\star$ which is as sparse as possible. In fact, (14) generalizes (13) from looking for potentials which have the smallest number of entries for which acyclicity fails to potentials which have the smallest number of entries for which the data is not globally consistent. Like (13), Eq. (14) is NP-hard.

The w-weighted ℓ^1-norm. In this work, we primarily consider (7) with objective function given by (8) with $p = 1$,

$$\min_{\phi \in \mathfrak{A}\mathfrak{d}} J_1(\phi) := \|\text{grad}\ \phi - y\|_{1,w}. \tag{15}$$

Equation (15) has the following properties:

1. In §4, it is shown that (15) can be formulated as a linear program (LP) in $\mathbb{R}^{n+2m}$ with $m + 1$ equality constraints. Convexity implies that any local minima is a global minima. Generally, the optimal set can be an $(n-1)$-dimensional polytope.

2. Defining the *residual*, $r := y - \operatorname{grad} \phi$, the optimization problem (15) can be rewritten analogously to (11) as

$$\min_{r \in \mathbb{R}^m} \|r\|_{1,w} \quad \text{such that } P^{\perp}_{\text{grad}}\, r = P^{\perp}_{\text{grad}}\, y \tag{16}$$

where $P^{\perp}_{\text{grad}}$ is the projection onto $\operatorname{imag}(\text{grad})^{\perp}$. Compressive sensing results imply that if the residual r is sparse, *i.e.*, obeys a power law decay, and if $P^{\perp}_{\text{grad}}$ satisfies certain properties, then the solution of (16) will be a good approximation to the solution of (14), hence sparse [12]. Although we do not verify these properties for $P^{\perp}_{\text{grad}}$ here, we demonstrate using simple examples in §3 and with numerical experiments for real datasets in §5 that the residual for the solution to (15) is sparse. In Prop. 3.3, we give a simple condition which implies that a vertex have at least one incident arc with zero residual.
3. In §4, we describe a graph-cut approach that can be used to efficiently solve (15), which has been applied to image processing and computer vision problems. The method is applied to several example statistical ranking problems in §5.

1.1. **Outline of paper.** In §2, we briefly review related work. In §3, we review the KKT conditions for optimality for the objective function (8) and give several simple examples to demonstrate statistical ranking and illustrate the differences between the ℓ^1- and ℓ^2-norm objective functions, (15) and (9). In §4, a computational method based on graph cuts for solving (15) is described. In §5, we apply the proposed method to three datasets. Finally, in §6, we discuss several future directions.

1.2. **Notation.**

1. Unordered pairs are denoted $\{i, j\}$ and the set of unordered pairs is denoted $\binom{V}{2}$ while ordered pairs are denoted (i, j) or abbreviated ij and the set of ordered pairs by $V \times V$. If i is incident on j or j is incident on i, we denote $i \sim j$. If $k = ij$, then $i = \text{tail}(k)$ and $j = \text{head}(k)$. For node $j \in V$, $\text{id}(j)$ and $\text{od}(j)$ denote the in-degree and out-degree respectively.
2. Greek letters are used for functions $V \to \mathbb{R}$ and Roman letters for functions $A \to \mathbb{R}$.
3. The vectors in $\mathbb{R}^n$ of all ones and zeros are denoted 1_n and 0_n.
4. Pointwise vector multiplication: $(a.b)_i := a_i b_i$
5. The Moore-Penrose pseudoinverse of a matrix A is denoted by $A^{\dagger}$.

2. **Related work.** Rank aggregation has evolved from electoral and social choice foundations with Borda (1781) and Condorcet (1786) to its current state [42, 36]. A comprehensive survey of the literature is beyond the scope of this work; more comprehensive reviews can be found in [14, 31, 10, 26].

Ordinal Data. Early work on ranking focused on ordinal data, where results are often negative. For example, Arrow's impossibility theorem states that when voters have more than 3 options, no voting system can aggregate the ranked preferences of individuals while also meeting 4 (sensible) criteria. Kemeny and Snell showed that

$$d_K(\sigma, \tau) := \#\{k = ij \in A \colon (\sigma_i > \sigma_j \text{ and } \tau_i < \tau_j) \text{ or } (\sigma_i < \sigma_j \text{ and } \tau_i > \tau_j)\}$$

is the unique distance between two orderings, σ and τ, which is a metric, invariant to permutation of the objects, independent of "irrelevant alternatives," and has

minimal positive distance equal to unity [27, §2]. Minimizing the distance, d_K between the orderings induced by a proposed potential ϕ and a given dataset is precisely $J_K(\phi)$ as defined in (13). The equivalence of (13) to the arc feedback set problem and Slater's problem, which are NP-hard, is discussed in [10] along with a survey of relaxation algorithms. [16, 15] extend these ideas to combing ranking results from multiple sources for web-search applications.

Cardinal Data. It appears that [11] was the first to consider aggregating cardinal ranking data. Written in the graph language formalism used here, they consider the problem

$$\min_{\phi} \sum_{\lambda\in\Lambda}\sum_{k\in A} w_{k,\lambda}|(\text{grad }\phi)_k - v_{k,\lambda}|. \tag{17}$$

where $v_{k,\lambda}$ is the pairwise preference given to pair $k \in A$ by a user $\lambda \in \Lambda$. In a later paper, [2] showed that the constraint matrix is totally unimodular and formulated the problem as a linear programming problem. This approach differs from the approach advocated in (15) because we first generate aggregate pairwise comparison data, y, from the individual user comparisons $v_{k,\lambda}$, via, *e.g.*, $y_k = \text{mean}_\lambda(v_{k,\lambda})$ and weights w from the number of user comparisons (see §5) and then find a potential for the dataset (grad, y, w). While our approach has the interpretation of grad ϕ being the "ℓ^1 pojection" of the edgeflow y onto imag(grad), it is more difficult to interpret the solution of (17). Note that for a particular choice of constructions for y and w, these two approaches agree for the ℓ^2-norm objective function (9) [26]. More recently, [24] generalizes (17) to, when written in the graph formalism considered here,

$$\min_{\phi} \sum_{\lambda\in\Lambda}\sum_{k\in A} f_{k,\lambda}[(\text{grad }\phi)_k - v_{k,\lambda}],$$

for convex functions $f_{k,\lambda}$. The author refers to this as the "separation" model for ranking and also introduces a further generalization, which also takes as data a desired rating vector, referred to as the "separation-deviation" model. These models are solved using network-flow algorithms and applied to evaluating country credit-risk ratings [23].

As discussed in the introduction, [26] considered (9) and interpreted the residual in terms of the Hodge decomposition. As such, they refer to this ranking methodology as HodgeRank. In [40], the HodgeRank framework developed in [26] is applied to distributing the task of assessing video quality for a large number of videos to a number of viewers using randomly generated graphs. Recently, [22] extended this work to use a graph representation of the data (as considered in this manuscript) rather than skew-symmetric matrices. [22] studies the application of Krylov iterative and algebraic multigrid methods to solving (9) for synthetic data on randomly generated graphs.

Another approach to the ranking problem, described in [21], is a two step process: (1) First extend the incomplete and inconsistent pairwise comparison data to a consistent and complete pairwise comparison dataset using a matrix completion algorithm while constraining the pairwise comparison matrix to be anti-symmetric and to have rank less than or equal to two using the nuclear norm (convex constraints). (2) The statistical ranking is then easily recovered by solving the least squares problem (9) where $y \in \text{im}(\text{grad})$.

It was recently shown that datasets can be constructed for which the HodgeRank method produces arbitrarily different ranking orders from two other well-known methods: the Principal Eigenvector and Tropical Eigenvector methods [39].

Other related work. While the methods studied here establish a ranking for a fixed graph, recent work (see, *e.g.*, [30, 3]) studies the problem of learning, from examples of preferences among alternatives in a graph, a statistical ranking for the remaining objects in the graph.

In [32], the authors used a continuous version of (15) to obtain an "L^1-norm Hodge decomposition" and successfully applied it to Retinex theory, imitating the human visual system which recovers the reflectance under varying illumination conditions. For a rectangle $R \subset \mathbb{R}^2$ and given data (g_1, g_2), they solved the following variational problem:

$$\min_{\phi} \int_R \sqrt{(\phi_x - g_1)^2 + (\phi_y - g_2)^2} \qquad \text{such that} \qquad \int_R \phi = 0. \tag{18}$$

3. **KKT conditions and examples illustrating statistical ranking.** In this section, we review the KKT optimality conditions for equations (9) and (15). We also consider several simple examples of the statistical rank aggregation problem for cyclic and/or inconsistent datasets, (V, A, y, w).

3.1. **KKT optimality conditions.** In what follows, we use the concept of a subdifferential or subderivative from convex analysis [34]. Given a convex function $f\colon \mathbb{R}^n \to \mathbb{R}$, the *subdifferential* of f at a point x_0 is defined

$$\partial f(x_0) = \{v \in \mathbb{R}^n \colon f(x) - f(x_0) \geq \langle v, x - x_0 \rangle \ \forall\ x \in \mathbb{R}^n\}.$$

The subdifferential of the absolute value function $f(x) = |x|$ is given

$$\partial |x| = \begin{cases} \operatorname{sign}(x) & x \neq 0 \\ [-1, 1] & x = 0. \end{cases}$$

Using the identity $\partial(f_1 + f_2)$, we obtain $\partial \|x\|_1 = \sum_{i=1}^n \partial |x_i|$.

Proposition 3.1 (KKT optimality conditions)**.** *Consider the statistical ranking problem* (7) *with objective function given by* $J_p(\phi) := \|\operatorname{grad} \phi - y\|_{p,w}^p$.

1. $\phi_{1,\star}$ *is a global minimizer for* $p = 1$ *if*

$$0 \in \operatorname{div}_w \partial \|\operatorname{grad} \phi_{1,\star} - y\|_{1,w} \tag{19}$$

2. $\phi_{2,\star}$ *is a global minimizer for* $p = 2$ *if*

$$\operatorname{div}_w (\operatorname{grad} \phi_{2,\star} - y) = 0. \tag{20}$$

Equation (19) follows from the identity $\partial g(x) = A^t \partial f(Ax - b)$ where $g(x) = f(Ax - b)$ [34]. Note that $s \in \ker(\operatorname{div}_w)$ implies that s is a w-weighted digraph circulation, *i.e.*, the w-weighted flow into each node equals the w-weighted flow out. The continuous analog of (15), given in (18), has KKT conditions with the interpretation that $r = \operatorname{grad} \phi - g$ has zero curvature on the set $\{x \in R \colon r(x) \neq 0\}$; (19) can be interpreted similarly.

3.2. **Single n-node cycle.** Consider the dataset $V = \{i\}_{i=1}^n$, $A = \{(i, i+1)\}_{i=1}^{n-1} \cup (n, 1)$, $m = n$, $y = 1_m$, and $w = 1_m$. This graph for $n = 4$ is displayed in Fig. 1(left). For this data, the ℓ^p-norm objective function (8) is written

$$J_p(\phi) = \|\operatorname{grad} \phi - 1_m\|_p^p.$$

The optimal solution for $p = 2$ is (uniquely) attained by $\phi_{\star,2} = 0_n$ with objective function value $J_2^\star = n$. The residual for the optimal solution, $r_{\star,2} = 1_m - \operatorname{grad} \phi_{\star,2} = 1_m$, is evenly distributed over all arcs.

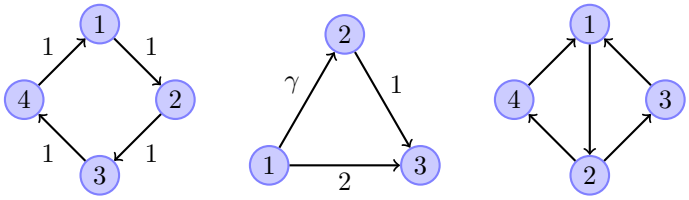

FIGURE 1. Diagrams of the digraphs for the examples considered in §3.

The optimal objective function value for $p = 1$ is $J_1^\star = n$, which is attained by any potential in the $(n-1)$-dimensional polytope defined by

$$P = \{\phi \in \mathbb{R}^n \colon (\text{grad } \phi)_j \leq 1 \ \forall \ j = 1, \ldots, m \text{ and } 1_n^t \phi = 0\}$$

which follows from (19). Note that P contains both 0_n and all circular shifts of the vector $[1, 2, \ldots, n] - \frac{n+1}{2} 1_n$. In the first case, the residual in the objective function is evenly distributed over all arcs, while the error in the latter case is concentrated on a single arc.

3.3. **Globally inconsistent, competing paths.** Consider the dataset $V = (1, 2, 3)$, $A = \{(1, 2), (2, 3), (1, 3)\}$, $y = (\gamma, 1, 2)$ for $\gamma > 0$, and $w = 1_m$ as in Fig. 1(center). This dataset is acyclic, yet globally inconsistent except for $\gamma = 1$. For this data, the ℓ^p objective function (8) is written

$$J_{p,\gamma}(\phi) = |\phi_2 - \phi_1 - \gamma|^p + |\phi_3 - \phi_2 - 1|^p + |\phi_3 - \phi_1 - 2|^p.$$

For $p = 2$, the optimal solution is given by $\phi_{\star,2} = (-(\gamma + 2)/3, (\gamma - 1)/3, 1)$, with optimal objective function value $J_{2,\gamma}^\star = \frac{1}{3}(\gamma - 1)^2$. The residual for the optimal solution is $r_{\star,2} = \frac{1}{3}|\gamma - 1| 1_m$.

For $p = 1$, the optimal objective function value is $J_{1,\gamma}^\star = |\gamma - 1|$. For $0 < \gamma \leq 1$, the optimal solution is given by any vector in the $(n-1)$-polytope:

$$P = \{\phi \in \mathbb{R}^3 \colon \phi_2 - \phi_1 - \gamma \geq 0, \ \ \phi_3 - \phi_2 - 1 \geq 0, \ \ \phi_3 - \phi_1 - 2 \leq 0, \ \text{ and } \ 1_n^t \phi = 0\}.$$

If $\gamma \geq 1$, then the optimal solution lies in the $(n-1)$-polytope defined as above, but with the inequalities reversed. For $\gamma \neq 1$, P contains both $\phi_{\star,2}$ where the residual is evenly distributed over all arcs and also solutions at the vertices of P where the residual is concentrated on a single arc.

3.4. **Two cycles.** Consider the dataset $V = (1, 2, 3, 4)$, $A = \{(1, 2), (2, 3), (3, 1), (2, 4), (4, 1)\}$, $y = 1_m$, and $w = 1_m$ as in Fig. 1(right). For this data, the ℓ^p objective function (8) is written

$$J_p(\phi) = |\phi_2 - \phi_1 - 1|^p + |\phi_3 - \phi_2 - 1|^p + |\phi_1 - \phi_3 - 1|^p + |\phi_4 - \phi_2 - 1|^p + |\phi_1 - \phi_4 - 1|^p.$$

For $p = 2$, the optimal solution is given by $\phi_{\star,2} = (1, -1, 0, 0)/4$ with optimal objective function value $J_2^\star = 9/2$. The residual for the optimal solution is $r_{\star,2} = 1_m - \text{grad } \phi_{\star,2} = (6, 3, 3, 3, 3)/4$.

For $p = 1$, the optimal solution is given by $\phi_{\star,1} = (1, -1, 0, 0)$ with optimal objective function value $J_1^\star = 3$. The residual for the optimal solution is $r_{\star,1} = 1_m - \text{grad } \phi_{\star,1} = (3, 0, 0, 0, 0)$. This example suggests that the ℓ^1 norm might be useful for identifying arcs which can be removed to yield globally consistent data.

3.5. **Further observations.** In §3.2 and §3.3, it is observed that $\phi_{2,\star} = \Delta^\dagger \mathrm{div}_w\, y$ also attains the minimum objective function value for the ℓ^1 problem. In the following proposition, we give sufficient conditions on the data (grad, y, w) for this to occur. These hypotheses include the datasets in §3.2 and §3.3.

Proposition 3.2. *Assume $w = 1_m$. Let $\phi_{2,\star} = \Delta^\dagger \mathrm{div}_w\, y$. If $id(j) + od(j) = 2$ for every node $j \in V$, then $J_1(\phi_{2,\star}) = \min_{\phi \in \mathfrak{Ad}} J_1(\phi)$.*

Proof. Write $r_{2,\star} = y - \mathrm{grad}\ \phi_{2,\star}$. The KKT optimality conditions (20) for the optimality of $\phi_{2,\star}$ for the objective function $J_2(\phi)$ can be expressed: there exists a residual arcflow $r_{2,\star}\colon A \to \mathbb{R}$ such that for each node $j \in V$,

$$\sum_{k\colon\, j=\mathrm{tail}(k)} (r_{2,\star})_k \quad - \sum_{k\colon\, j=\mathrm{head}(k)} (r_{2,\star})_k = 0$$

By assumption, these sums combined have only two terms implying that the terms are either both zero or are arranged in such a way so that the values, and hence signs, cancel. In either case, (19) is satisfied. □

The following proposition suggests why the residual of the ℓ^1-norm problem (15) for datasets with $w = 1_m$ might be sparse.

Proposition 3.3. *Assume $w = 1_m$. Let $\phi_{1,\star}$ denote a minimum of (15) and $r_{1,\star} = y - \mathrm{grad}\ \phi_{1,\star}$. For every $j \in V$ such that $id(j) + od(j)$ is odd, there exists an arc $k \in A$ with either $j = \mathrm{head}(k)$ or $j = \mathrm{tail}(k)$ such that $(r_{1,\star})_k = 0$.*

Proof. The KKT optimality condition (19) implies the existence of an arcflow $s\colon A \to \mathbb{R}$ such that $\mathrm{div} s = 0_n$. Let $j \in V$ be such that $id(j) + od(j)$ is odd and define $A_j = \{k \in A\colon j = \mathrm{head}(k) \text{ or } j = \mathrm{tail}(k)\}$. Suppose $(r_{1,\star})_k \neq 0$ for every $k \in A_j$ which implies $s_k \in \{1, -1\}$ for all $k \in A_j$. But the sum of an odd number of ± 1 cannot be zero. □

In particular, Prop. 3.3 suggests that for the example problem in §3.4, at least one of the residuals is exactly zero.

4. **Computational methods for ranking using ℓ^1-norm regression.** In this section, we demonstrate that (15) can be written as a linear program and describe a network-flow approach for approximately minimizing (15) which utilizes the graph structure of the ranking problem. The network-flow algorithm discussed here has also been employed for imaging and computer vision problems.

4.1. **Fomulation as a linear program.** By introducing the auxillary variables g and h, the ℓ^1-norm optimization problem (15) can be reformulated as the following standard-form linear program (LP) in $\mathbb{R}^{n+2m}$ with $m+1$ equality constraints:

$$\begin{aligned} \min_{(\phi,g,h)\in\mathbb{R}^{n+2m}} \ & (0_n, 1_m, 1_m)^t(\phi; g; h) \\ \text{such that } & [\mathrm{grad},\ -\mathrm{Id}_n,\ \mathrm{Id}_n](\phi; g; h) = y \\ & 1_n^t \phi = 0 \\ & g \geq 0, h \geq 0. \end{aligned} \tag{21}$$

For small datasets, (21) can be solved using black-box LP software. We use `CVX`, a Matlab package for specifying and solving convex programs [19, 18]. For larger datasets however, methods which utilize the underlying graph structure of the optimization problem (15) are more efficient. We discuss such algorithms next.

4.2. **Graph-cut approach for solving (15).** Equation (15) is within the class of convex, cost tension/differential problems [35] that can be optimized using network flow algorithms [5] and, in particular, those described in [4]. Steepest-descent-based algorithms proposed by Murota for computing ϵ-optimal solutions may also be applied [33]. The latter have been in specialized for image processing and computer vision applications [7, 13, 28] using a graph-cut approach [9, 29].

We begin by introducing a steepest-descent algorithm which can be refined to give a second algorithm with better time complexity performance. The main idea of behind these two algorithms is to recast the problem of finding a potential in $\mathbb{R}^n$ to finding a potential on a lattice, for which a descent direction may be obtained by finding an s-t minimum cut. In computer vision, this algorithm is hence known as the graph-cut approach [9, 29].

First algorithm. Given a current potential, $\phi_0 \in \mathbb{R}^n$ and a lattice parameter $\delta > 0$, we introduce a new feasible set, $\mathcal{S}_\delta = \phi_0 + \delta\mathbb{Z}^n$. Our goal is to minimize the objective function $J_1(\phi)$ over this new set, *i.e.*, solve

$$\min_{\phi \in \mathcal{S}_\delta} J_1(\phi) := \|\text{grad } \phi - y\|_{1,w}, \tag{22}$$

The solution to (22) is obtained iteratively by finding, amongst all lattice points within an ℓ^∞-norm ball of radius δ, the potential with smallest objective value, *i.e.*, solving

$$\min_{\sigma \in \{-1,0,1\}^n} J_1(\phi_0 + \delta\sigma). \tag{23}$$

Thus, at each iteration of the solution to (22), each potential component can either retain the current value or increase/decrease by δ.

The solution of (23) can efficiently computed using a graph-cut approach [9] [29]. This is accomplished by constructing an augmented graph, with prescribed edge capacities, for which an s-t minimum cut partitions the graph, yielding a solution to (23), $\sigma_\star$. The solution of the s-t minimum cut problem is efficiently obtained by solving the dual maximum-flow problem. We refer the reader to [7],[13, chp. 3], [28], for instance, for a more comprehensive description of the augmented graph construction and how a descent direction can be found using a s-t minimum cut.

Once the optimal solution $\sigma_\star$ of (23) is computed, the current potential is updated via $\phi_{k+1} = \phi_k + \delta\sigma_\star$ and the procedure is iterated until an objective function value is obtained such that the subsequent iteration has the same value. It may be shown that every iteration reduces the ℓ^∞-norm distance of the current iterate to the optimal solution $\phi_\star$ by δ and thus the number of iterations required for convergence is given by $\|\phi_\star - \phi_0\|_\infty/\delta$ [33]. So, if we denote by $T(n,m)$ the time needed to compute a s-t minimum-cut in a graph of n nodes and m edges, the time complexity of the algorithm is $T(n,m) \cdot \|\phi_\star - \phi_0\|_\infty/\delta$.

Let us emphasize that although this algorithm is effective in practice, it depends on the ℓ^∞-norm distance between the initial potential and global minimum, and thus has only pseudo-polynomial time complexity.

Second algorithm. The second algorithm is a refinement of the first, permitting a polynomial time complexity bound [33]. The idea, similar to the binomial search algorithm, is to iteratively halve the lattice parameter, δ. More precisely, given an initial guess ϕ_0, and a sufficiently large initial lattice parameter δ, (22) is solved for this lattice parameter δ as described above, and then we set $\delta \leftarrow \delta/2$ and repeat until a prescribed precision level ϵ is attained. The method is summarized in Algorithm 1. The number of times (23) needs to be solved for a single lattice

Algorithm 1: For solving the statistical ranking problem, (15).

Data: Initial guess $\phi_0 \in \mathfrak{AD}$, convergence tolerance $\epsilon > 0$, and initial search size parameter $\delta > 0$ (sufficiently large).
Output: ϵ-optimal potential ϕ with $J_1(\phi) = E$.
Set $E = \infty$ (current objective function level)
Set $\phi = \phi_0$
for $k = 0$ **to** ceil($\log_2 \frac{\delta}{\epsilon}$), **do**
  Set $\delta_k = \frac{\delta}{2^k}$
  while $J_1(\phi) < E$, **do**
    Set $E = J_1(\phi)$
    Use graph cut methods to solve:

$$\sigma_\star = \arg \min_{\sigma \in \{-1,0,1\}^n} J_1(\phi + \delta_k \sigma)$$

    Set $\phi = \phi + \delta_k \sigma_\star$

parameter δ is bounded above by n and thus the time complexity of this algorithm is $n \cdot T(n,m) \cdot \log_2 \frac{d}{\epsilon}$ [33].

We use the graph-cut implementation described in [8], which achieves excellent performance in the examples considered in our experiments; see §5.

5. Applications / Experimental studies.

In this section, we conduct a series of computational experiments to demonstrate the differences between solving the statistical ranking problem using the ℓ^1- and ℓ^2-norms. All computations in this section were performed on a 2.4 GHz dual core processor with 2GB memory.

5.1. Jester: the online joke recommender.

In this section, we consider a dataset from an online joke recommendation website [25]. The raw dataset contains 1,761,439 ratings (on a scale from -10 to 10) of 140 jokes from 59,132 users collected between November 2006 and May 2009. In the raw dataset, the jokes are numbered from 1 to 150 with the following 10 numbers omitted: 1, 2, 3, 4, 6, 9, 10, 11, 12, and 14 and the users are numbered from 1 to 63,978 leaving 4,846 users who did not review any jokes. We discard the data associated with the following 12 jokes: 5, 20, 27, 31, 43, 51, 52, 61, 73, 80, 100, and 116, which, due to being removed from the Jester website before the collection period ended, have fewer user reviews. The remaining dataset contains 1,758,234 ratings of 128 jokes from 59,123 users.

Construction of pairwise comparison data from raw data. Let Λ be the set of all joke reviewers and let $u(i,\lambda)$ be the rating given to joke $i \in V$ by reviewer $\lambda \in \Lambda$. Let $E \subset \binom{V}{2}$ denote the set of unordered joke pairs that have been reviewed by at least 1 reviewer. For each unordered joke pair $\{i,j\} = e \in E$, we define

$$\Lambda_e = \{\lambda \in \Lambda \colon \lambda \text{ reviewed both items } i \text{ and } j\}. \tag{24}$$

The unordered pairwise weights and comparison data are then constructed

$$w_e = |\Lambda_e| \quad \text{and} \quad y_e = \frac{1}{|\Lambda_e|} \sum_{\lambda \in \Lambda_e} u(i,\lambda) - u(j,\lambda) \quad \text{where } e = \{i,j\} \in E.$$

We then define the set of ordered pairs A consisting of arcs $a = ij$ such that for $e = \{i,j\} \in E$, $y_e \geq 0$. Lastly, for each $a = ij \in A$, if we denote $e = \{i,j\}$, then we

rank	joke id # ℓ^2	ℓ^1
1	89	89
2	53	53
3	105	105
4	62	62
5	106	106
6	104	104
7	114	50
8	32	114
9	129	32
10	35	68
11	50	35
12	68	129

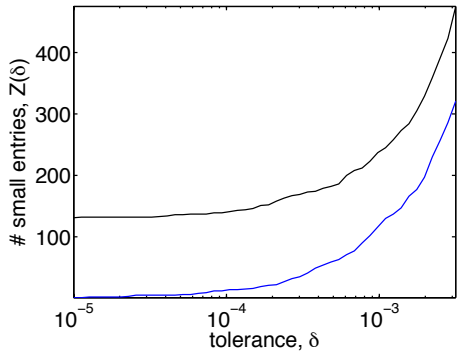

FIGURE 2. See §5.1 (left) A table which give the top 12 joke id numbers (in the original dataset numbering) for optimal rankings obtained using the ℓ^1- and ℓ^2-norms. (right) The number of small entries (25) in the residual of the ℓ^1 (black) and ℓ^2 (blue) solutions.

define the ordered pairwise weight $w_a = w_e$ and comparison $y_a = y_e$. This method is invariant to translation in the ratings $u(i, \lambda)$ for each reviewer $\lambda \in \Lambda$. We remark that there are several other methods of constructing pairwise data from user reviews which are invariant under other transformations [26].

One of the advantageous properties of this dataset is that the pairwise comparison data is complete, *i.e.*, for every unordered pair $\{i, j\}$, either $ij \in A$ or $ji \in A$. In fact, the mean number of users which ranked each joke pair is 6,976. The joke pair with the minimum number of pairwise comparisons is $\{74, 141\}$ with 3,284 comparisons, while the joke pair with the maximum number of pairwise comparisons is $\{7, 8\}$ with 57,456 comparisons. In the following, we exploit the properties that this dataset is relatively small (128 alternatives) and the pairwise comparison data is complete.

Numerical experiments. We begin by considering the solutions to the rank aggregation problem (7) for $p = 1$ and $p = 2$. The relative residual norm for the optimal solutions obtained are

$$\frac{\|\text{grad } \phi_{1,\star} - y\|_{1,w}}{\|y\|_{1,w}} = 0.052 \qquad \text{and} \qquad \frac{\|\text{grad } \phi_{2,\star} - y\|_{2,w}}{\|y\|_{2,w}} = 0.072.$$

We find that the optimal solutions $\phi_{1,\star}$ and $\phi_{2,\star}$ are very close, *e.g.*, $\|\phi_{1,\star} - \phi_{2,\star}\|_2 = 0.26$, although they do induce different orderings on the alternatives (jokes), which we make more precise below. The top-12 jokes for each ranking are given in Figure 2(left).

The *induced ordering relation* on a set of alternatives by the ranking r is defined $i \succ_\phi j$ iff $\phi(i) > \phi(j)$, $i =_\phi j$ iff $\phi(i) = \phi(j)$, and $i \prec_\phi j$ iff $j \succ_\phi i$. Given an ordering relation induced by ϕ on V, let $\tau_r(i)$ denote the order of $i \in V$. If $\tau_\phi(i)$ is large, then alternative i has large ranking ($\phi(i)$ is large).

It is useful to introduce a few metrics to measure the distance between rankings and their induced orderings. We use the following measures of distance to compare potentials for a given dataset:

$$R(\phi_1, \phi_2) = \frac{\|\phi_1 - \phi_2\|_2}{(\|\phi_1\|_2 + \|\phi_2\|_2)/2} \quad T(\phi_1, \phi_2) = \frac{\|\text{grad } (\phi_1 - \phi_2)\|_{2,w}}{(\|\text{grad } \phi_1\|_{2,w} + \|\text{grad } \phi_2\|_{2,w})/2}.$$

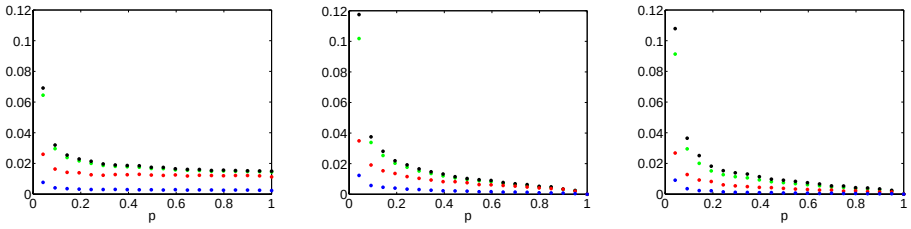

FIGURE 3. See §5.1. A study of the ℓ^1 and ℓ^2 optimal solutions to the statistical ranking problem for varying Erdös-Rényi sub-digraph connectivity parameter, p. (left) $\phi^p_{1,\star}$ vs. $\phi^p_{2,\star}$. (center) $\phi^p_{2,\star}$ vs. $\phi^1_{2,\star}$. (right) $\phi^p_{1,\star}$ vs. $\phi^1_{1,\star}$. In each comparison, four metrics are used: $\overline{R}_p$ (green), $\overline{T}_p$ (black), $\overline{K}_p$(blue), and $\overline{S}_p$ (red).

The *Kendall tau distance* between two potentials ϕ_1 and ϕ_2 is defined

$$K(\phi_1, \phi_2) := \frac{\#\{(i,j) \colon i > j,\ \phi_1(i) < \phi_1(j),\ \text{and } \phi_2(i) > \phi_2(j)\}}{n(n-1)/2}.$$

Following [16], we define the *Spearman footrule distance* between two induced orderings τ_1 and τ_2,

$$S(\tau_1, \tau_2) := \frac{\|\tau_1 - \tau_2\|_1}{n^2/2}.$$

Each of these distances have been normalized to take values on $[0, 1]$.

For the Jester dataset, let $\tau_{1,\star}$ and $\tau_{2,\star}$ denote the ordering indices induced by the rankings $\phi_{1,\star}$ and $\phi_{2,\star}$ respectively. The following table gives the values of these distances.

$R(\phi_{1,\star}, \phi_{2,\star})$	$T(\phi_{1,\star}, \phi_{2,\star})$	$K(\phi_{1,\star}, \phi_{2,\star})$	$S(\tau_{1,\star}, \tau_{2,\star})$
0.015	0.015	0.0023	0.011

The ℓ^1 and ℓ^2 rankings and their respective induced orderings are similar, yet the number of zeros in the solution residuals differ significantly. Define the number of 'small values' in the residual $r = y - \text{grad } \phi$, as measured by

$$Z(\delta) = \#\{j \colon |r_j| < \delta\}. \tag{25}$$

In Fig. 2(right), we plot δ vs. $Z(\delta)$ for the potentials $\phi_{1,\star}$ (black) and $\phi_{1,\star}$ (blue). Indeed the number of zeros in the residual for the ℓ^1-norm ranking is much larger than that for the ℓ^2-norm ranking.

Next, we explore the dependency of the rankings and induced orderings on the connectivity of the graph. For a sequence of decreasing probabilities p, we generate Erdös-Rényi random sub-digraphs $D(p) \subset D$, which randomly discard arcs with a uniform probability of $1 - p$, independent from every other arc. The expected number of arcs in $D(p)$ is $p\binom{n}{2} = p\frac{n(n-1)}{2}$ and thus we consider p to be an incompleteness parameter for the digraph. The threshold for weak connectedness of $D(p)$ is $p_{twc} = \frac{\log n}{n} \approx 0.038$. Using the subset of the pairwise comparison data specified by the random digraph $D(p)$, we compute the solutions to the ℓ^1- and ℓ^2-norm rank aggregation problems using `CVX`, which we denote $\phi^p_{1,\star}$ and $\phi^p_{2,\star}$ respectively. For each value of p, we generate an ensemble of size 20 of random digraphs $D(p)$. For each digraph $D(p)$, we make three comparisons: $\phi^p_{1,\star}$ vs. $\phi^p_{2,\star}$, $\phi^p_{2,\star}$ vs. $\phi^1_{2,\star}$, and $\phi^p_{1,\star}$ vs. $\phi^1_{1,\star}$ using each of the 4 distances defined above: $R(\cdot,\cdot)$, $T(\cdot,\cdot)$, $S(\cdot,\cdot)$, and

largest average user rating	ℓ^2-norm ranking, (9)	ℓ^1-norm ranking, (15)
12.7 Lawrence of Arabia	4.3 It's a Wonderful Life	4.1 It's a Wonderful Life
12.7 The Shawshank Redemption	4.2 Rear Window	4.0 Rear Window
12.6 The Adventures Of Indiana Jones	3.8 Friends (4th Season)	3.5 The Shawshank Redemption
12.4 Say Anything	3.6 Vertigo	3.4 Seven Samurai
12.4 12 Angry Men	3.6 The Shawshank Redemption	3.4 Vertigo
12.4 It's a Wonderful Life	3.5 The Good, the Bad and the Ugly	3.4 The Godfather
12.3 National Lampoon's Animal House	3.4 The Godfather	3.3 The Good, the Bad and the Ugly
12.3 Apocalypse Now Redux	3.3 Serpico	3.2 Psycho
12.4 The Good, the Bad and the Ugly	3.3 To Kill a Mockingbird	3.2 To Kill a Mockingbird
12.2 Braveheart	3.3 The Natural	3.2 12 Angry Men

TABLE 1. Top 10 movies, ranked using 3 different methods. See §5.2.

$K(\cdot,\cdot)$. In Fig. 3, we plot the connectivity parameter p vs. the ensemble average for each of these quantities: $\overline{R}_p$ (green), $\overline{T}_p$ (black), $\overline{K}_p$(blue), and $\overline{S}_p$ (red) for each of the three comparisons: $\phi^p_{1,\star}$ vs. $\phi^p_{2,\star}$ (left), $\phi^p_{2,\star}$ vs. $\phi^1_{2,\star}$ (center), and $\phi^p_{1,\star}$ vs. $\phi^1_{1,\star}$ (right). We find that the potentials are robust in the sense that they change only slightly, even when computed without $\approx 80\%$ of the data. For large datasets, this observation could be used as a heuristic to initialize an optimization method.

5.2. **Yahoo! Movie user ratings.** In this section, we consider the Yahoo! Movie user rating dataset consisting 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 [41][3]. Each movie was rated by between 1 and 4,238 users (the average number of reviews for each 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 which implies that the pairwise comparisons for the raw dataset are 8.1% complete.

The majority of movies in the dataset received relatively few reviews:

# times movie reviewed	1	2	3	4	5	6	7	≥ 8
occurrences	4,901	1,882	897	548	398	316	237	2736

The movies which received less than 20 rankings were discarded from the dataset, leaving 1,477 movies, each of which were reviewed by an average of 119.8 users. We then removed 22 users who did not review any of the remaining movies. The remaining 7620 reviewers reviewed between 1 and 889 movies (on average they reviewed 23.2 movies). The average user rating for each movie was computed and the top ten averages along with the movie names are given in Table 1 (first column).

Construction of pairwise comparison data from raw data. The pairwise comparison data are constructed as described in §5.1. Of the 1,090,026 (movie) pairs (i,j) where $i > j$, there are 907,683 for which a user has given a rating to both movies i and j. Although the pairwise comparison dataset is 83% complete, 85% of the pairwise comparisons were constructed from less than 10 user comparisons. A log-histogram of the number of user-pairwise comparisons for each pair, *i.e.*, $|\Lambda_e|$ where Λ_e is defined in (24), is given in Fig. 4 (left).

Numerical experiments. We begin by computing the solution to the ℓ^1-norm and ℓ^2-norm statistical ranking problems (9) and (15) for this dataset. The graph-cut method described in §4 for solving the ℓ^1-norm statistical ranking problem (15) was

[3]We discarded 34 entries from the dataset which review Yahoo! movie_id 0, which does not appear in the movie content description file.

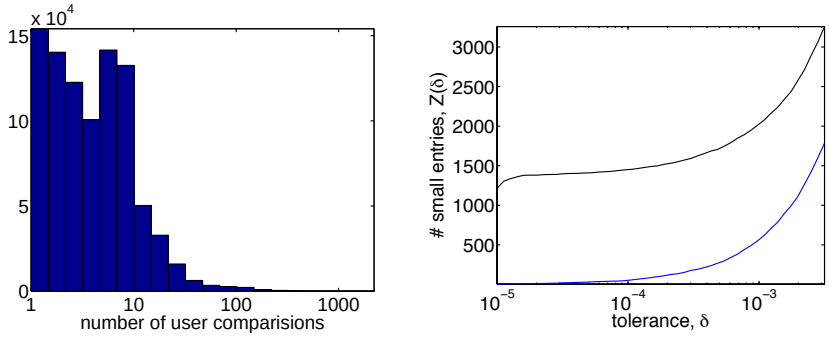

FIGURE 4. See §5.2. (left) A log-histogram of the number of user comparisons for each movie pair. (right) The number of small entries (25) in the residual of the ℓ^1 (black) and ℓ^2 (blue) solutions.

implemented in C. The optimal solution obtained, $\phi_{1,\star}$, has relative residual norm

$$\frac{\|\text{grad } \phi_{1,\star} - y\|_{1,w}}{\|y\|_{1,w}} = 0.415.$$

The solution to the ℓ^2-norm statistical ranking problem (9) was obtained using Matlab's `lsqr` function. The relative residual norm of the optimal solution obtained, $\phi_{2,\star}$, is

$$\frac{\|\text{grad } \phi_{2,\star} - y\|_{2,w}}{\|y\|_{2,w}} = 0.488.$$

The larger relative residual norms for these potentials indicate that this dataset is less consistent than the dataset considered in §5.1.

The top ten movies using the ordering induced by $\phi_{1,\star}$ and $\phi_{2,\star}$ are given in Table 1 along with the top ten movies obtained by sorting the average user ratings. As in §5.1, we find $\phi_{1,\star}$ and $\phi_{2,\star}$ are similar:

$R(\phi_{1,\star}, \phi_{2,\star})$	$T(\phi_{1,\star}, \phi_{2,\star})$	$K(\phi_{1,\star}, \phi_{2,\star})$	$S(\tau_{1,\star}, \tau_{2,\star})$
0.088	0.068	0.016	0.032

In Fig. 4, we plot δ vs. $Z(\delta)$ (as defined in (25)) for the potentials $\phi_{1,\star}$ (black) and $\phi_{2,\star}$ (blue). The number of small values in the residual for the ℓ^1-norm ranking is much larger than that for the ℓ^2-norm ranking.

5.3. **Association of Tennis Professionals (ATP) Match Data.** The Association of Tennis Professionals (ATP) is an organization which organizes the ATP World Tour, the primary tennis circuit for male professional tennis players [6]. On October 5, 2011 we collected 2011 ATP tennis match records from the Tennis Datenbank website [38] using a Scrapy web crawler [37]. The dataset was collected by starting with the set of tennis players containing only Rafael Nadal and subsequently adding players to the set who have played members of the set. In this manner, a set of 609 players was collected, each of which have played at least one other member of the set. For each of these players, the match history from October 9, 2010 to October 5, 2011 was then downloaded, yielding 10,784 matches (where both players were contained in the player set). We then remove players from the dataset who played less than 27 matches. The remaining dataset consists of 4,074 matches played between 218 players. For each match, the collected dataset contains

the outcome, match score, court type (hard, clay, grass, etc...), tournament round, date, and venue information. At the end of each tennis match, there is a winner and a loser; there are no ties.

Construction of pairwise comparison data from raw data. Let Λ be the set of all tennis matches and let $E \subset \binom{V}{2}$ denote the set of unordered player pairs that have met in a match at least once. For each unordered player pair $\{i,j\} = e \in E$, we define

$$\Lambda_e = \{\lambda \in \Lambda : \text{match } \lambda \text{ is between players } i \text{ and } j\}.$$

For each player pair $e = \{i,j\} \in E$, we let w_e be the number of times players i and j have met in a match, *i.e.*, $w_e = |\Lambda_e|$.

For the pairwise comparison values y_k, we seek a construction that estimates the relative player strengths, *i.e.*, provides a "victory margin" for matches, while appealing to simplicity. To accomplish this, we propose using match scores rather than just win-loss results. For player pair $\{i,j\} = e \in E$, the unordered pairwise comparison data are constructed

$$y_e = \frac{1}{|\Lambda_e|} \sum_{\lambda \in \Lambda_e} \left(2 \frac{\#\{\text{sets } i \text{ beat } j \text{ in match } \lambda\}}{\#\{\text{sets in match } \lambda\}} - 1 \right) \quad \text{where } e = \{i,j\} \in E.$$

Note that the expression in parenthesis is anti-symmetric in the indices i and j and lies in the interval $[-1,1]$. Then, following §5.1, we define the set of ordered pairs A consisting of arcs $a = ij$ such that for $e = \{i,j\} \in E$, $y_e \geq 0$. Lastly, for each $a = ij \in A$, if we denote $e = \{i,j\}$, then we define the ordered pairwise weight $w_a = w_e$ and comparison $y_a = y_e$.

The resulting digraph, $D = (V, A)$, is fairly sparse, $\frac{m}{n(n-1)/2} = 0.1722$, which is approximately 7 times the Erdös-Rényi threshold for weak connectedness ($p_{twc} = \frac{log(n)}{n} \approx 0.025$). For the players who did meet in a tennis match, they did so only 1.19 (= mean w) times on average. The following table gives histogram data for the number of matches between players:

# times player pairs met	0	1	2	3	4	5	6	7	8	≥ 9
occurrences	19,603	3,395	560	77	13	3	0	1	1	0

Numerical experiments. The relative residual norm for the optimal solutions obtained to the rank aggregation problem (7) for $p = 1$ and $p = 2$ are

$$\frac{\|\text{grad } \phi_{1,\star} - y\|_{1,w}}{\|y\|_{1,w}} = 0.79 \qquad \text{and} \qquad \frac{\|\text{grad } \phi_{2,\star} - y\|_{2,w}}{\|y\|_{2,w}} = 0.86.$$

This dataset is less consistent than the datasets considered in §5.1 and §5.2. This may partially reflect the fact that each pairwise comparison is typically based on only a single match, while for other datasets, pairwise comparisons where based on an average over many reviews. The ranking methods considered in this paper also neglect temporal variations in the data, which may be more pronounced for this particular dataset.

The top-10 tennis players for each ranking and also those given by the Association of Tennis Professionals (ATP) are given in Table 2. The ATP player ranking is based on points accrued from (typically 19) tournament results from the previous calendar year. Points are awarded based on the tournament category (Grand Slam, Barclays ATP World Tour Finals, ATP World Tour Masters 1000, etc...) and the round in which the player advances [6]. Indeed, all three ranking methodologies produce very similar rankings.

ATP Entry Ranking	ℓ^1-norm ranking, (15)	ℓ^2-norm ranking, (9)
1 Djokovic, Novak (SRB)	1.34 Djokovic, Novak (SRB)	1.20 Djokovic, Novak (SRB)
2 Nadal, Rafael (ESP)	1.33 Federer, Roger (SUI)	1.16 Federer, Roger (SUI)
3 Federer, Roger (SUI)	1.28 Nadal, Rafael (ESP)	1.09 Nadal, Rafael (ESP)
4 Murray, Andy (GBR)	1.09 Soderling, Robin (SWE)	0.91 Murray, Andy (GBR)
5 Ferrer, David (ESP)	1.06 Murray, Andy (GBR)	0.86 Soderling, Robin (SWE)
6 Soderling, Robin (SWE)	0.95 Ferrer, David (ESP)	0.83 Del Potro, J. M. (ARG)
7 Tsonga, Jo-Wilfried (FRA)	0.92 Fish, Mardy (USA)	0.78 Ferrer, David (ESP)
8 Fish, Mardy (USA)	0.92 Del Potro, J. M. (ARG)	0.72 Tsonga, Jo-Wilfried (FRA)
9 Monfils, Gael (FRA)	0.84 Tsonga, Jo-Wilfried (FRA)	0.70 Fish, Mardy (USA)
10 Berdych, Tomas (CZE)	0.75 Monfils, Gael (FRA)	0.68 Monfils, Gael (FRA)

TABLE 2. 2011 tennis player rankings given by the Association of Tennis Professionals (ATP) as of October 5, 2011 and obtained by solving (15) and (9); see §5.3.

The number of small values in the residual $r = y - \text{grad}\ \phi_\star$, as measured by $Z(\delta)$ defined in (25) for $\delta = 4 \times 10^{-4}$ is 0 for the ℓ^2-norm ranking and 260 for the ℓ^1-norm ranking.

6. **Discussion and future directions.** In this paper, we have considered the statistical rank aggregation problem of ranking a set of alternatives from a dataset consisting of pairwise comparisons of the alternatives. Our approach uses the ℓ^1-norm objective function (15), which provides an alternative to recent work utilizing the ℓ^2-norm objective function (9) [26, 22]. These two objective functions have different interpretations of the residuals, each of which may be useful for a particular application. There are many possible future extensions of the work conducted in this paper.

Our focus for the computational methods in this paper, described in §4 and applied to datasets in §5, was to demonstrate proof of concept for ranking using the proposed ℓ^1-norm objective function, (15). For the moderately-sized datasets studied in §5, our implementation for solving (15) finds solutions in seconds on a laptop computer, comparable to the time required to solve (9) using Matlab's `lsqr` function. We believe that significant improvements can be made to our algorithm and are particularly interested in the use of differential inclusion methods [1].

We have demonstrated, using simple examples in §3 and with numerical experiments for real datasets in §5, that the residual for the solution to (15) is sparse. In the introduction, we noted that this optimization problem, when expressed as in (16), resembles the framework of compressive sensing problems [12]. It would be of great interest to determine if the properties needed for these results to apply hold.

We are also interested in exploring the sensitivity of the ranking on the pairwise values y, arc weights w, and digraph topology D. Furthermore, it may be possible to apply robust optimization methods [20] to reduce the influence of "noise" in the dataset, by including information about the uncertainty of y, w, and D into the ranking model.

Finally, as discussed in the introduction, the Hodge decomposition further decomposes the residual in (9), into locally cyclic (3 vertex cycles) and locally acyclic ($\geq$ 4 vertex cycles) components. We are interested in the problem of employing fast graph-cut methods to approximate the solution of (12), to locate the locally cyclical component of the pairwise data, by solving

$$\min_{\Phi \in V \times V \times V} \|\text{curl}\ \Phi - y\|_{1,w}.$$

This formulation is also motivated in [26].

Acknowledgements. We thank Yuan Yao for suggesting this problem and Lawrence Carin for directing us to the Yahoo! Webscope dataset. B. Osting is supported by NSF DMS-1103959. J. Darbon is supported by ONR-N00014-11-1-0749. S. Osher is supported by ONR N00014-08-1-1119, N00014-10-0221, and NSF DMS-0914561.

REFERENCES

[1] J.P. Aubin and A. Cellina, *Differential inclusions*, Springer-Verlag, 1984.
[2] I. Ali, W. D. Cook, and M. Kress, *Ordinal ranking and intensity of preference: A linear programming approach*, Management Science **32** (1986), no. 12, 1642–1647.
[3] S. Agarwal, *Learning to rank on graphs*, Machine Learning **81** (2010), 333–357.
[4] R. Ahuja, D. Hochbaum, and J. Orlin, *Solving the convex cost integer dual network flow problem*, Management Science **49** (2003), 950–964.
[5] R.K Ahuja, T.L. Magnanti, and J.B. Orlin, *Network flows: Theory, algorithms and applications*, Prentice Hall, 1993.
[6] *2011 ATP world tour media guide*, http://www.atpworldtour.com/Press/Rankings-and-Stats.aspx, accessed: 10/5/2011.
[7] J.M. Bioucas-Dias and G. Valadao, *Phase unwrapping via graph cuts*, IEEE Transactions on Image Processing **16** (2007), no. 3, 698–709.
[8] Y. Boykov and V. Kolmogorov, *An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision*, IEEE Transactions on Pattern Analysis and Machine Intelligence **26** (2004), no. 9, 1124–1137.
[9] Y. Boykov, O. Veksler, and R. Zabih, *Fast approximate energy minimization via graph cuts*, IEEE Transactions on Pattern Analysis and Machine Intelligence **23** (2001), no. 11, 1222–1239.
[10] I. Charon and O. Hudry, *A survey on the linear ordering problem for weighted or unweighted tournaments*, 4OR **5** (2007), 5–60.
[11] W. D. Cook and M. Kress, *Ordinal ranking with intensity preference*, Management Science **31** (1985), no. 1, 26–32.
[12] E. Candes and T. Tao, *Near optimal signal recovery from random projections: Univeral encoding strategies?*, IEEE Transactions on Information Theory **52** (2006), no. 12, 5406–5425.
[13] J. Darbon, *Composants logiciels et algorithmes de minimisation exacte d'énergies dédiés au traitement des images*, Ph.D. thesis, Ecole Nationale Supérieure des Télćommunications, October 2005.
[14] H. A. David, *The method of paired comparisons*, Charles Griffin & Co., 1963.
[15] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, *Rank aggregation revisited.*
[16] ———, *Rank aggregation methods for the web*, Proceedings International Conference World Wide Web (WWW'01), vol. 10, 2001, pp. 613–622.
[17] L. R. Foulds, *Graph theory applications*, Springer, 1992.
[18] M. Grant and S. Boyd, *Graph implementations for nonsmooth convex programs*, Recent Advances in Learning and Control (V. Blondel, S. Boyd, and H. Kimura, eds.), Lecture Notes in Control and Information Sciences, Springer-Verlag Limited, 2008, http://stanford.edu/~boyd/graph_dcp.html, pp. 95–110.
[19] ———, *CVX: Matlab software for disciplined convex programming, version 1.21*, http://cvxr.com/cvx, April 2011.
[20] D. Goldfarb and G. Iyengar, *Robust portfolio selection problems*, Mathematics of Operations Research **28** (2003), no. 1, 1–38.
[21] D. F. Gleich and L.-H. Lim, *Rank aggregation via nuclear norm minimization*, arXiv:1102.4821, 2011.
[22] A. N. Hirani, K. Kalyanaraman, and S. Watts, *Least squares ranking on graphs*, arXiv:1011.1716v4, 2011.
[23] D. S. Hochbaum and E. Moreno-Centeno, *Country credit-risk rating aggregation via the separation-deviation model*, Optimization Methods and Software **23** (2008), no. 5, 741–762.

[24] D. S. Hochbaum, *The separation and separation-deviation methodology for group decision making and aggregate ranking*, TutORials in Operations Research **7** (2010), 116–141.
[25] *Jester: The online joke recommender*, http://eigentaste.berkeley.edu/dataset/jester_dataset_2.zip, accessed: 9/9/2011.
[26] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye, *Statistical ranking and combinatorial Hodge theory*, Math. Program. Ser. B **127** (2010), no. 1, 203–244.
[27] J. G. Kemeny and J. L. Snell, *Mathematical models in the social sciences*, The MIT Press, 1962.
[28] V. Kolmogorov and A. Shioura, *New algorithms for convex cost tension problem with application to computer vision*, Discrete Optimization **6** (2009), no. 4, 378–393.
[29] V. Kolmogorov and R. Zabih, *What energy functions can be minimized via graph cuts?*, IEEE Transactions on Pattern Analysis and Machine Intelligence **2** (2004), no. 26, 147–159.
[30] T.-Y. Liu, *Learning to rank for information retrieval*, Foundations and Trends in Information Retrieval **3** (2009), no. 3, 225–331.
[31] J. I. Marden, *Analyzing and modeling rank data*, Monographs on Statistics and Applied Probability, No. 64, Chapman & Hall, 1995.
[32] W. Ma, J.-M. Morel, S. Osher, and A. Chien, *An l1-based model for retinex theory and its application to medical images*, IEEE Conference on Computer Vision and Pattern Recognition (2011), 153–160.
[33] K. Murota, *Discrete convex optimization*, SIAM Society for Industrial and Applied Mathematics, 2003.
[34] R. T. Rockafellar, *Convex analysis*, Princeton University Press, 1970.
[35] R.T. Rockafellar, *Network flows and monotropic optimization*, John Wiley & Sons, 1984.
[36] D. G. Saari, *Mathematical structure of voting paradoxes*, Economic Theory **15** (2000), 1–53.
[37] *Scrapy web crawling framework*, http://scrapy.org/, accessed: 10/5/2011.
[38] *Tennis datenbank website*, http://tennis.wettpoint.com/en/, accessed: 10/5/2011.
[39] N. M. Tran, *Pairwise ranking: choice of method can produce arbitrarily different rank order*, arXiv:1103.1110v1, 2011.
[40] Q. Xu, Y. Yao, T. Jiang, Q. Huang, B. Yan, and W. Lin, *Random parial paired comparison for subjective video quality assessment via HodgeRank*, ACM Multimedia, 2011.
[41] *Yahoo! Webscope dataset: ydata-ymovies-user-movie-ratings-content-v1_0*, http://webscope.sandbox.yahoo.com, accessed: 10/5/2011.
[42] H. P. Young, *Condorcet's theory of voting*, The American Political Science Review **82** (1988), no. 4, 1231–1244.

Received January 16, 2012; revised xxxx 20xx.

E-mail address: braxton@math.ucla.edu, jerome@math.ucla.edu, sjo@math.ucla.edu