LEARNING TO RANK WITH COMBINATORIAL HODGE THEORY

XIAOYE JIANG, LEK-HENG LIM, YUAN YAO, AND YINYU YE

Abstract. We propose a number of techniques for learning a global ranking from data that may be incomplete and imbalanced — characteristics that are almost universal to modern datasets coming from e-commerce and internet applications. We are primarily interested in cardinal data based on scores or ratings though our methods also give specific insights on ordinal data. From raw ranking data, we construct pairwise rankings, represented as edge flows on an appropriate graph. Our rank learning method exploits the graph Helmholtzian, which is the graph theoretic analogue of the Helmholtz operator or vector Laplacian, in much the same way the graph Laplacian is an analogue of the Laplace operator or scalar Laplacian. We shall study the graph Helmholtzian using combinatorial Hodge theory, which provides a way to unravel ranking information from edge flows. In particular, we show that every edge flow representing pairwise ranking can be resolved into two orthogonal components, a gradient flow that represents the l_2-optimal global ranking and a cyclic (divergence-free) flow that measures the inconsistency of the global ranking obtained — if this large, then it indicates that the data does not have a good global ranking. This cyclic flow can be further decomposed orthogonally into a triangular cyclic flow (curl) and a 'harmonic' flow that is globally cyclic but locally acyclic; these provides information on whether inconsistency in the ranking data arises locally or globally.

When applied to the problem of rank learning, Hodge decomposition sheds light on whether a given dataset may be globally ranked in a meaningful way or if the data is inherently inconsistent and thus could not have any reasonable global ranking; in the latter case it provides information on the nature of the inconsistency. An obvious advantage over the NP-hardness of Kemeny optimization (which is primarily for ordinal ranking data) is that the discrete Hodge decomposition may be easily computed via a linear least squares regression. We also investigated the l_1-projection of edge flows, showing that this has a dual given by correlation maximization over bounded divergence-free flows, and the l_1-approximate sparse cyclic ranking, showing that this has a dual given by correlation maximization over bounded curl-free flows. We discuss connections with well-known ordinal ranking techniques such as Kemeny optimization and Borda count from social choice theory.

1. Introduction

The problem of ranking in various contexts has become increasingly important in machine learning. Many datasets require some form of ranking to facilitate identification of important entries, extraction of principal attributes, and to perform efficient search and sort operations. Modern internet and e-commerce applications

2000 *Mathematics Subject Classification.* 68T05, 58A14, 90C05, 90C27, 91B12, 91B14.

Key words and phrases. Rank learning, rank aggregation, combinatorial Hodge theory, discrete exterior calculus, combinatorial Laplacian, graph Helmholtzian, Kemeny optimization, Borda count .

have spurred an enormous growth in such datasets: Google's search engine, CiteSeer's citation database, eBay's feedback-reputation mechanism, Netflix's movie recommendation system, all accumulate a large volume of data that needs to be ranked.

These modern datasets typically have one or more of the following features that render traditional ranking methods (such as those in social choice theory) inapplicable or ineffective: (1) unlike traditional ranking problems such as votings and tournaments, the data often contains *cardinal scores* instead of ordinal orderings; (2) the given data is largely *incomplete* with most entries missing a substantial amount of information; (3) the data will almost always be *imbalanced* where the amount of available information varies widely from entry to entry and/or from criterion to criterion; (4) the given data often lives on a large *complex network*, either explicitly or implicitly, and the structure of this underlying network is itself important in the ranking process. These new features have posed new challenges and call for new techniques. In this paper we will look at a method that addresses them to some extent.

A fundamental problem here is to globally rank a set of *alternatives* based on scores given by *voters*. Here the words 'alternatives' and 'voters' are used in a general sense that depends on the context. For example, the alternatives may be websites indexed by Google, scholarly articles indexed by CiteSeer, sellers on eBay, or movies on Netflix; the voters in the corresponding contexts may be other websites, other scholarly articles, buyers, or viewers. The 'voters' could also refer to groups of voters: e.g. websites, articles, buyers, or viewers grouped respectively by topics, authorship, buying patterns, or movie tastes. The 'voters' could even refer to something entirely abstract, such as a collection of different criteria used to judge the alternatives.

The features (1)–(4) can be observed in the aforementioned examples. In the eBay/Netflix context, a buyer/viewer would assign cardinal scores (1 through 5 stars) to sellers/movies instead of ranking them in an ordinal fashion; the eBay/Netflix datasets are highly incomplete since most buyers/viewers would have rated only a very small fraction of the sellers/movies, and also highly imbalanced since a handful of popular sellers/blockbuster movies will have received an overwhelming number of ratings while the vast majority will get only a moderate or small number of ratings. The datasets from Google and CiteSeer have obvious underlying network structures given by hyperlinks and citations respectively. Somewhat less obvious are the network structures underlying the datasets from eBay and Netflix, which come from aggregating the pairwise comparisons of buyers/movies over all sellers/viewers. Indeed, we shall see that in all these ranking problems, graph structures naturally arise from *pairwise comparisons*, irrespective of whether there is an obvious underlying network (e.g. from citation, friendship, or hyperlink relations) or not, and this serves to place ranking problems of seemingly different nature on an equal graph-theoretic footing. The incompleteness and imbalance of the datasets could then be manifested as the (edge) sparsity structure and (vertex) degree distribution of pairwise comparison graphs.

In collaborative filtering applications, one often encounters a personalized ranking problem, when one needs to find a global ranking of alternatives that generates the most consensus within a group of voters who share similar interests/tastes. While the rank learning problem investigated in this paper plays a fundamental

role in such personalized ranking problems, there is also the equally important problem of clustering voters into interest groups, which our methods do not address. We would like to stress that in this paper we only concern ourselves with the ranking problem but not the clustering problem. So while we have made use of the Netflix prize dataset to motivate our studies, our paper should not be viewed as an attempt to solve the Netflix prize problem.

The method that we will use to analyze pairwise rankings, which we represent as edge flows on a graph, comes from *discrete* or *combinatorial Hodge theory*. Among other things, combinatorial Hodge theory provides us with a mean to determine a global ranking that also comes with a 'certificate of reliability' for the validity of this global ranking. While Hodge theory is well-known to pure mathematicians as a corner stone of geometry and topology, and to applied mathematician as an important tool in computational electromagnetics and fluid dynamics, its application to rank learning problems has, to the best of our knowledge, never been studied[1].

In all our proposed methods, the graph in question has as its vertices the alternatives to be ranked, voters' preferences are then quantified and aggregated (we will say how later) into an edge flow on this graph. Hodge theory then yields an orthogonal decomposition of the edge flow into three components: a *gradient flow* that is globally consistent (acyclic), a *harmonic flow* that is locally acyclic but globally cyclic, and a *curl flow* that is locally cyclic. This decomposition is known as the *Hodge decomposition*. The usefulness of the decomposition lies in the fact that the gradient flow component induces a global ranking of the alternatives. Unlike the computationally intractable Kemeny optimal, this may be easily computed via a linear least squares problem. Furthermore, the l_2-norm of the least squares residual, which represents the contribution from the sum of the remaining curl flow and harmonic flow components, quantifies the validity of the global ranking induced by the gradient flow component. If the residual is small, then the gradient flow accounts for most of the variation in the underlying data and therefore the global ranking obtained from it is expected to be a majority consensus. On the other hand, if the residual is large, then the underlying data is plagued with cyclic inconsistencies (i.e. intransitive preference relations of the form $a \succeq b \succeq c \succeq \cdots \succeq z \succeq a$) and one may not assign any reasonable global ranking to it.

We would like to point out here that cyclic inconsistencies are not necessarily due to error or noise in the data but may very well be an inherent characteristic of the data. As the famous impossibility theorems from social choice theory [2, 35] have shown, inconsistency (or, rather, intransitivity) is inevitable in any societal preference aggregation that is sophisticated enough. Social scientists have, through empirical studies, observed that preference judgement of groups or individuals on a list of alternatives do in fact exhibit such irrational or inconsistent behavior. Indeed in any group decision making process, a lack of consensus is the norm rather than the exception in our everyday experience. This is the well-known *Condorcet paradox* [8]: the majority prefers a to b and b to c, but may yet prefer c to a. Even a single individual making his own preference judgements could face such dilemma — if he uses multiple criteria to rank the alternatives. As such, the cyclic inconsistencies is intrinsic to any real world ranking data and should be thoroughly analyzed. Hodge

[1]Nevertheless, Hodge theory has recently found other applications in statistical learning theory [36].

theory again provides a mean to do so. The curl flow and harmonic flow components of an edge flow quantify respectively the local and global cyclic inconsistencies.

Loosely speaking, a dominant curl flow component suggests that the inconsistencies is of a local nature while a dominant harmonic flow component suggests that it is of a global nature. If most of the inconsistencies come from the curl (local) component while the harmonic (global) component is small, then this roughly translates to mean that the ordering of closely ranked alternatives is unreliable but that of very differently ranked alternatives is reliable, i.e. we cannot say with confidence whether the ordering of the 27th, 28th, 29th ranked items makes sense but we can say with confidence that the 4th, 60th, 100th items should be ordered according to their rank. In other words, Condorcet paradox may well apply to items ranked closed together but not to items ranked far apart. For example, if a large number of gourmets (voters) are asked to state their preference on an extensive range of food items (alternatives), there may not be a consensus for their preference with regard to hamburgers, hot dogs, and pizzas and there may not be a consensus for their preference with regard to caviar, foie gras, and truffles; but there may well be a near universal preference for the latter group of food items to the former group. In this case, the inconsistencies will be mostly local and we should expect a large curl flow component. If in addition the harmonic flow component is small, then most of the inconsistencies happen locally and we could interpret this to mean that the global ranking is valid on a coarse scale (ranking different groups of food) but not on a fine scale (ranking similar food items belonging to a particular group).

When studied in conjunction with robust regression and compressed sensing, the three orthogonal subspaces given by Hodge decomposition provide other insights. In this paper we will see two results involving l_1-optimizations where these subspaces provide meaningful and useful interpretations in the primal-dual way: (A) the l_1-projection of an edge flow onto the subspace of gradient flows has a dual problem as the maximal correlation over bounded cyclic flows, i.e. the sum of curl flows and harmonic flows; (B) the l_1-approximation of a sparse cyclic flow, has a dual problem as the maximal correlation over bounded locally acyclic flows. These results indicate that the three orthogonal subspaces could arise even in settings where orthogonality is lost.

1.1. **Organization of this Paper.** In Section 2 we introduce the main problem and discuss how a pairwise comparison graph may be constructed from data comprising cardinal scores by voters on alternatives and how a simple least squares regression may be used to compute the desired solution. We define the *combinatorial curl*, a measure of local (triangular) inconsistency for such data, and also the *combinatorial gradient* and *combinatorial divergence*. Section 3 presents a purely matrix-theoretic view of Hodge theory, but at the expense of some geometric insights. These are covered when we formally introduce Hodge theory in Section 4. We first remind the reader how one may construct a d-dimensional simplicial complex from any given graph (the pairwise comparison graph in our case) by simply filling-in all its k-cliques for $k \leq d$. Then we introduce combinatorial Hodge theory for general d-dimensional simplicial complex but focusing on the $d = 2$ case and its relevance to the ranking problem. In Section 5 we discuss some implications of the Hodge decomposition applied to ranking, with a deeper analysis on the least squares method in Section 2. Section 6 extends the analysis to two closely related l_1-minimization problems, the l_1-projection of pairwise ranking onto gradient flows

and the l_1-approximate sparse cyclic ranking. A discussion of the connections with Kemeny optimization and Borda count in social choice theory can be found in Section 7.

1.2. **Notations.** Let V be a finite set. We will adopt the following notation from combinatorics:

$$\binom{V}{k} := \text{set of all } k\text{-element subset of } V.$$

In particular $\binom{V}{2}$ would be the set of all unordered pairs of elements of V and $\binom{V}{3}$ would be the set of all unordered triples of elements of V (the sets of ordered pairs and ordered triples will be denoted $V \times V$ and $V \times V \times V$ as usual). Ordered and unordered pairs will be delimited by parentheses (i, j) and braces $\{i, j\}$ respectively, and likewise for triples and n-tuples in general.

We will use positive integers to label alternatives and voters. Henceforth, V will always be the set $\{1, \dots, n\}$ and will denote a set of alternatives to be ranked. In our approach to rank learning, these alternatives would be represented as vertices of a graph. $\Lambda = \{1, \dots, m\}$ will denote a set of voters. For $i, j \in V$, we write $i \succeq j$ to mean that alternative i is preferred to alternative j. If we wish to emphasize the preference judgement of a particular voter $\alpha \in \Lambda$, we will write $i \succeq_\alpha j$.

Since our approach mandates that we borrow terminologies from graph theory, vector calculus, linear algebra, algebraic topology, as well as various ranking theoretic terms, we think that it would help to summarize some of the correspondence here.

Graph theory	**Linear algebra**	**Vec. calculus**	**Topology**	**Ranking**
Function on vertices	Vector in $\mathbb{R}^n$	Potential function	0-cochain	Score function
Edge flow	Skew-symmetric matrix in $\mathbb{R}^{n \times n}$	Vector field	1-cochain	Pairwise ranking
Triangular flow	Skew-symmetric hyper--matrix in $\mathbb{R}^{n \times n \times n}$	Tensor field	2-cochain	Triplewise ranking

As the reader will see, the notions of gradient, divergence, curl, Laplace operator, and Helmholtz operator from vector calculus and topology will play important roles in rank learning. One novelty of our approach lies in extending these notions to the other three columns, where most of these have no well-known equivalent. For example, what we will call a *harmonic ranking* is central to the question of whether a global ranking is feasible. This notion is completely natural from the vector calculus or topology point-of-view, they correspond to solutions of the Helmholtz equation or homology classes. However, it will be hard to define harmonic ranking directly in social choice theory without this insight, and we suspect that it is the reason why this notion has never been discussed in existing studies of ranking in social choice theory and other fields.

2. Main Problem

The main problem discussed in this paper is that of learning a global ranking from a dataset comprising a set of alternatives ranked by a set of voters. This is a problem that has received attention in fields including decision science [31, 32], financial economics [4, 25], machine learning [6, 10, 15, 18], social choice [2, 35, 30], statistics [13, 23, 24, 26, 27, 28, 29], among others. Our objective towards rank-learning is two-fold: like everybody else, we want to deduce a global ranking from the data whenever possible; but in addition to that, we also want to detect when

the data does not permit a statistically meaningful global ranking and in which case characterize the data in terms of its local and global inconsistencies.

Let $V = \{1, \dots, n\}$ be the set of alternatives to be ranked and $\Lambda = \{1, \dots, m\}$ be a set of voters. The implicit assumption is that each voter would have *rated*, i.e. assigned cardinal scores or given an ordinal ordering to, a small fraction of the alternatives. But no matter how small the rated portion is, one may always convert such ratings into *pairwise rankings* as follows. For each voter $\alpha \in \Lambda$, the *pairwise ranking matrix* of α is a skew-symmetric matrix $Y^\alpha \in \mathbb{R}^{n\times n}$, i.e. for each ordered pair $(i, j) \in V \times V$, we have

$$Y^\alpha_{ij} = -Y^\alpha_{ji}.$$

Informally, Y^α_{ij} measures the 'degree of preference' of the ith alternative over the jth alternative held by the αth voter. Studies of ranking problems in different disciplines have led to rather different ways of quantifying such 'degree of preference'. In Section 2.2.1, we will see several ways of defining Y^α_{ij} (as score difference, score ratio, and score ordering) coming from decision science, machine learning, social choice theory, and statistics. If the voter α did not compare alternatives i and j, then Y^α_{ij} is considered a missing value and set to be 0 for convenience; this way of handling missing values allows Y^α to be a skew-symmetric matrix for each $\alpha \in \Lambda$. Nevertheless we could have assigned any arbitrary value or a non-numerical symbol to represent missing values, and this would not affect our algorithmic results because of our use of the following weight function.

Define the *weight function* $w : \Lambda \times V \times V \to \mathbb{R}^+$ as the indicator function

$$w^\alpha_{ij} = w(\alpha, i, j) = \begin{cases} 1 & \text{if } \alpha \text{ made a pairwise comparison for } \{i, j\}, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore $w^\alpha_{ij} = 0$ iff Y^α_{ij} is a missing value. Note that $W^\alpha = [w^\alpha_{ij}]$ is a symmetric $\{0, 1\}$-valued matrix; but more generally, w^α_{ij} may be chosen as the capacity (in the graph theoretic sense) if there are multiple comparisons between $\{i, j\}$ by voter α.

Our general paradigm for rank learning is to minimize a weighted sum of pairwise loss of a global ranking on the given data over a model class $\mathcal{M}$ of all global rankings. We begin with a simple sum-of-squares loss function,

$$\min_{X\in\mathcal{M}_G} \sum\nolimits_{\alpha,i,j} w^\alpha_{ij}(X_{ij} - Y^\alpha_{ij})^2, \tag{1}$$

where the *model class* $\mathcal{M}_G$ is a subset of the skew-symmetric matrices,

$$\mathcal{M}_G = \{X \in \mathbb{R}^{n\times n} \mid X_{ij} = s_j - s_i, s : V \to \mathbb{R}\}. \tag{2}$$

Any $X \in \mathcal{M}_G$ induces a global ranking on the alternatives $1, \dots, n$ via the rule $i \succeq j$ iff $s_i \geq s_j$. Note that ties, i.e. $i \succeq j$ and $j \succeq i$, are allowed and this happens precisely when $s_i = s_j$.

For ranking data given in terms of cardinal scores, this simple scheme preserves the magnitudes of the ratings, instead of only the ordering, when we have globally consistent data (see Definition 2.3). Moreover, it may also be computed more easily than many other loss functions (though the computational cost depends also on the choice of $\mathcal{M}$). This simple scheme is not as restrictive as it first seems. For example, the Kemeny optimization in the classical social choice theory may be realized as a special case where $Y^\alpha_{ij} \in \{\pm 1\}$ and $\mathcal{M}$ is the *Kemeny* model class,

$$\mathcal{M}_K := \{X \in \mathbb{R}^{n\times n} \mid X_{ij} = \operatorname{sign}(s_j - s_i), s : V \to \mathbb{R}\}. \tag{3}$$

The function sign : $\mathbb{R} \to \{\pm 1\}$ takes nonnegative numbers to 1 and negative numbers to -1. A binary valued Y_{ij}^{α} is the standard scenario in binary pairwise comparisons [1, 2, 11, 18, 24]; in this context, a global ranking is usually taken to be synonymous as a Kemeny optimal. We will discuss Kemeny optimization in greater details in Section 7.

2.1. **Pairwise Comparison Graph and Flows.** A graph structure arises naturally from ranking data as follows. Let $G = (V, E)$ be an undirected graph whose vertex set is V, the set of alternatives to be ranked, and whose edge set is

$$E = \big\{\{i, j\} \in \tbinom{V}{2} \;\big|\; \textstyle\sum_{\alpha} w_{ij}^{\alpha} > 0\big\}, \tag{4}$$

i.e. the set of pairs $\{i, j\}$ where pairwise comparisons have been made. We call such G a *pairwise comparison graph*. One can further associate weights on the edges as capacity, e.g. $w_{ij} = \sum_{\alpha} w_{ij}^{\alpha}$.

A pairwise ranking can be viewed as *edge flows* on G, i.e. a function $X : V \times V \to \mathbb{R}$ that satisfies

$$X(i, j) = -X(j, i) \quad \text{if } \{i, j\} \in E, \tag{5}$$

$$X(i, j) = 0 \qquad \text{otherwise.} \tag{6}$$

It is clear that a skew-symmetric matrix (in particular, a pairwise ranking matrix) $[X_{ij}]$ induces an edge flow and vice versa. So henceforth we will not distinguish between edge flows and skew-symmetric matrices and will often write X_{ij} in place of $X(i, j)$.

We will now borrow some terminologies from vector calculus. An edge flow of the form $X_{ij} = s_j - s_i$, i.e. $X \in \mathcal{M}_G$, can be regarded as the *gradient* of a function $s : V \to \mathbb{R}$, which will be called a *potential* function. In the context of ranking, a potential function is a *score function* or *utility function* on the set of alternatives, assigning a score $s(i) = s_i$ to alternative i. Note that any such function defines a *global ranking* as discussed after (2). To be precise, we define gradient as follows.

Definition 2.1. *The* ***combinatorial gradient*** *operator maps a potential function $s : V \to \mathbb{R}$ to an edge flow on G as follows*

$$(\operatorname{grad} s)(i, j) = s_j - s_i \quad \textit{for every } \{i, j\} \in E. \tag{7}$$

An edge flow that has this form will be called a ***gradient flow****.*

In other words, the combinatorial gradient takes global rankings to pairwise rankings. Pairwise rankings that arise in this manner will be called *globally consistent* (formally defined in Definition 2.3). Given a globally consistent pairwise ranking X, we can easily solve $\operatorname{grad}(s) = X$ to determine a score function s (up to a constant), and from s we can get a global ranking of the alternatives in the manner described after (2). Observe that $\mathcal{M}_G = \{\operatorname{grad} s \mid s : V \to \mathbb{R}\}$ is the set of all globally consistent pairwise rankings.

For convenience, we will drop the adjective 'combinatorial' from 'combinatorial gradient'. We may sometimes also drop the adjective 'pairwise' in 'globally consistent pairwise ranking' when there is no risk of confusion.

The optimization problem (1) can be rewritten in the form of a weighted l_2-minimization on a pairwise comparison graph

$$\min_{X \in \mathcal{M}_G} \|X - \bar{Y}\|_{2,w}^2 = \min_{X \in \mathcal{M}_G} \Big[\textstyle\sum_{\{i,j\} \in E} w_{ij} (X_{ij} - \bar{Y}_{ij})^2\Big] \tag{8}$$

where

$$w_{ij} := \sum_\alpha w_{ij}^\alpha \quad \text{and} \quad \bar{Y}_{ij} := \frac{\sum_\alpha w_{ij}^\alpha Y_{ij}^\alpha}{\sum_\alpha w_{ij}^\alpha}. \tag{9}$$

An optimizer thus corresponds to an l_2-projection of a pairwise ranking edge flow $\bar{Y}$ onto the space of gradient flows. Note that $W = [w_{ij}] = \sum_\alpha W^\alpha$ is a symmetric nonnegative-valued matrix.

An interesting variation of this scheme is an analogous l_1-projection onto the space of gradient flows,

$$\min_{X \in \mathcal{M}_G} \|X - \bar{Y}\|_{1,w} = \min_{X \in \mathcal{M}_G} \Big[\sum\nolimits_{\{i,j\} \in E} w_{ij} |X_{ij} - \bar{Y}_{ij}| \Big]. \tag{10}$$

Its solutions are more robust to outliers or large deviations in $\bar{Y}_{ij}$ as (10) may be regarded as the *least absolute deviation* (LAD) method in robust regression. We will discuss this problem in greater details in Section 6.1.

Combinatorial Hodge theory will provide a geometric interpretation of the optimizer and residuals of (8) as well as further insights on (10). Before going deeper into the analysis of such optimization problems, we present several examples of pairwise ranking arising from applications.

2.2. **Pairwise Rankings.** Humans are unable to make accurate preference judgement on even moderately large sets. In fact, it has been argued that most people can rank only between 5 to 9 alternatives at a time [33]. This is probably why many rating scales (e.g. the ones used by Amazon, eBay, Netflix, Youtube) are all based on a 5-star scale. Hence one expects large human generated ranking data to be at best partially ordered (with chains of length about 5 to 9, if [33] is accurate). For most people, it is a harder task to rank or rate 20 movies than to compare the movies a pair at a time. In certain settings such as tennis tournaments and wine tasting, only pairwise comparisons are possible. Pairwise comparison methods, which involve the smallest partial rankings, is thus natural for analyzing such ranking data.

Pairwise comparisons also help reduce bias due to the arbitrariness of rating scale by adopting a relative measure. As we will see in Section 2.2.1, pairwise comparisons provide a way to handle missing values, which are expected because of the general lack of incentives or patience for a human to process a large dataset. For these reasons, pairwise comparison methods have been popular in psychology, statistics, and social choice theory [38, 24, 11, 31, 2]. Such methods are also getting increasing attention from the machine learning community as they may be adapted for studying classification problems [17, 15, 18]. We will present two very different instances where pairwise rankings arise: recommendation systems and exchange economic systems.

2.2.1. *Recommendation systems.* The generic scenario in recommendation systems is that there are m voters rating n alternatives. For example, in the Netflix context, viewers will rate a movie on a scale of 5 stars [6]; in financial markets, analysts will rate a stock or a security by 5 classes of recommendations [4]. In these cases, we let $A = [a_{\alpha i}] \in \mathbb{R}^{m \times n}$ represent the voter-alternative matrix. A typically has a large number of missing values; for example, the dataset that Netflix released for its prize competition contains a viewer-movie matrix with 99% of its values missing. The standard problem here is to predict these missing values from the given data but we caution the reader again that this is *not* the problem addressed in our paper.

Instead of estimating the missing values of A, we want to learn a global ranking of the alternatives from A, without having to first estimate the missing values.

Even though the matrix A may be highly incomplete, we may aggregate over all voters to get a pairwise ranking matrix using one of the four following methods.

(1) **Arithmetic mean of score differences:** The score difference refers to $Y_{ij}^{\alpha} = a_{\alpha j} - a_{\alpha i}$. The arithmetic mean over all customers who have rated both i and j is

$$\bar{Y}_{ij} = \frac{\sum_{\alpha}(a_{\alpha j} - a_{\alpha i})}{\#\{\alpha \mid a_{\alpha i}, a_{\alpha j} \text{ exist}\}}.$$

This is translation invariant.

(2) **Geometric mean of score ratios:** Assuming $A > 0$. The score ratio refers to $Y_{ij}^{\alpha} = a_{\alpha j}/a_{\alpha i}$. The (log) geometric mean over all customers who have rated both i and j is

$$\bar{Y}_{ij} = \frac{\sum_{\alpha}(\log(a_{\alpha j}) - \log(a_{\alpha i}))}{\#\{\alpha \mid a_{\alpha i}, a_{\alpha j} \text{ exist}\}}.$$

This is scale invariant.

(3) **Binary comparison:** Here $Y_{ij}^{\alpha} = \operatorname{sign}(a_{\alpha j} - a_{\alpha i})$. Its average is the probability difference that the alternative j is preferred to i than the other way round,

$$\bar{Y}_{ij} = \Pr\{\alpha \mid a_{\alpha j} > a_{\alpha i}\} - \Pr\{\alpha \mid a_{\alpha j} < a_{\alpha i}\}.$$

This is invariant up to a monotone transformation.

(4) **Logarithmic odds ratio:** As in the case of binary comparison, except that we adopt a logarithmic scale

$$\bar{Y}_{ij} = \log \frac{\Pr\{\alpha \mid a_{\alpha j} \geq a_{\alpha i}\}}{\Pr\{\alpha \mid a_{\alpha j} \leq a_{\alpha i}\}}.$$

This is also invariant up to a monotone transformation.

Each of these four statistics is a form of "average pairwise ranking" over all voters. The first model leads to the concept of *position-rules* in social choice theory [30] and is also used in machine learning recently [10]. The second model has appeared in multi-criteria decision theory [31]. The third and fourth models are known as *linear model* [29] and *Bradley-Terry model* [7] respectively in the statistics and psychology literature. There are other plausible choices for defining $\bar{Y}_{ij}$, e.g. [38, 26, 27, 28], but we will not discuss more of them here. It suffices to note that there is a rich variety of techniques to preprocess raw ranking data into the pairwise ranking edge flow $\bar{Y}_{ij}$ that serves as input to our Hodge theoretic method. However, it should also be noted that the l_2- and l_1-optimization on graphs in (8) and (10) may be applied with any of the four choices above since only the knowledge of $\bar{Y}_{ij}$ is required but the sum-of-squares and Kemeny optimization in (1) and (3) require the original score difference or score order data be known for each voter.

2.2.2. *Exchange economic systems.* A purely exchange economic system may be described by a graph $G = (V, E)$ with vertex set $V = \{1, \dots, n\}$ representing the n goods and edge set $E \subseteq \binom{V}{2}$ representing feasible pairwise transactions. If the market is complete in the sense that every pair of goods is exchangeable, then G

is a complete graph. Suppose the exchange rate between the ith and jth goods is given by

$$1 \text{ unit } i = a_{ij} \text{ unit } j, \quad a_{ij} > 0.$$

Then the exchange rate matrix $A = [a_{ij}]$ is a *reciprocal matrix* (possibly with missing values), i.e. $a_{ij} = 1/a_{ji}$ for all $i, j \in V$. The reciprocal matrix was first used in the studies of paired preference aggregation by Saaty [31]; it was also used by Ma [25] to study currency exchange markets. A pricing problem here is to look for a *universal equivalent* which measures the values of goods (this is in fact an abstraction of the concept of money), i.e. $\pi : V \to \mathbb{R}$ such that

$$a_{ij} = \frac{\pi_j}{\pi_i}.$$

In complete markets where G is a complete graph, there exists a universal equivalent if and only if the market is *triangular arbitrage-free*, i.e. $a_{ij}a_{jk} = a_{ik}$ for all distinct $i, j, k \in V$ — since in this case the transaction path $i \to j \to k$ provides no gain nor loss over a direct exchange $i \to k$.

Such purely exchange economic system is equivalent to pairwise ranking via the logarithmic map,

$$X_{ij} = \log a_{ij}.$$

The triangular arbitrage-free condition is then equivalent to the transitivity condition in (12), i.e.

$$X_{ij} + X_{jk} + X_{ki} = 0.$$

So asking if a universal equivalent exists is the same as asking if a global ranking $s : V \to \mathbb{R}$ exists so that $X_{ij} = s_j - s_i$ with $s_i = \log \pi_i$.

2.3. **Measure of Triangular Inconsistency: combinatorial curl.** Upon constructing pairwise rankings from the raw data, we need a statistics to quantify the inconsistency in the pairwise rankings. Again we will borrow a terminology from vector calculus and define a notion of *combinatorial curl* as a measure of triangular inconsistency.

Given a pairwise ranking represented as an edge flow X on a graph $G = (V, E)$, we expect the following 'consistency' property: following a loop $i \to j \to \cdots \to i$ where each edge is in E, the amount of the scores raised should be equal to the amount of the scores lowered, whence after a loop of comparisons we should return to the same score on the same alternative. Since the simplest loop is a triangular loop $i \to j \to k \to i$, this motivates the combinatorial curl as a measure of triangular inconsistency.

We will first define a notion analogous to edge flows. The *triangular flow* on G is a function $\Phi : V \times V \times V \to \mathbb{R}$ that satisfies

$$\begin{aligned}\Phi(i, j, k) = \Phi(j, k, i) = \Phi(k, i, j)\\ = -\Phi(j, i, k) = -\Phi(i, k, j) = -\Phi(k, j, i),\end{aligned}$$

i.e. an odd permutation of the arguments of Φ changes its sign while an even permutation preserves its sign[2] A triangular flow describes *triplewise rankings* in the same way an edge flow describes pairwise rankings.

[2]A triangular flow is an alternating 3-tensor and may be represented as a skew-symmetric hypermatrix $[\Phi_{ijk}] \in \mathbb{R}^{n\times n\times n}$, much like an edge flow is an alternating 2-tensor and may be represented by a skew-symmetric matrix $[X_{ij}] \in \mathbb{R}^{n\times n}$. We will often write Φ_{ijk} in place of $\Phi(i, j, k)$.

Definition 2.2. *Let X be an edge flow on a graph $G = (V, E)$. Let*

$$T(E) := \{\{i, j, k\} \in \binom{V}{n} \mid \{i, j\}, \{j, k\}, \{k, i\} \in E\}$$

be the collection of triangles with every edge in E. We define the ***combinatorial curl*** *operator that maps edge flows to triangular flows by*

$$(\operatorname{curl} X)(i, j, k) = \begin{cases} X_{ij} + X_{jk} + X_{ki} & \text{if } \{i, j, k\} \in T(E), \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

In other words, the combinatorial curl takes pairwise rankings to triplewise rankings. Again, we will drop the adjective 'combinatorial' when there is no risk of confusion. The skew-symmetry of X, i.e. $X_{ij} = -X_{ji}$, guarantees that $\operatorname{curl} X$ is a triangular flow, i.e.

$$\begin{aligned}(\operatorname{curl} X)(i, j, k) = (\operatorname{curl} X)(j, k, i) = (\operatorname{curl} X)(k, i, j) \\ = -(\operatorname{curl} X)(j, i, k) = -(\operatorname{curl} X)(i, k, j) = -(\operatorname{curl} X)(k, j, i).\end{aligned}$$

The curl of a pairwise ranking measures its triangular inconsistency. This extends the *consistency index* by Kendall and Smith [24], which counts the number of circular triads, from ordinal settings to cardinal settings. Note that for binary pairwise ranking where $X_{ij} \in \{\pm 1\}$, the absolute value of the curl, $|(\operatorname{curl} X)(i, j, k)|$, can only take two values, 1 or 3. The triangle $\{i, j, k\} \in T(E)$ contains a cyclic ranking or *circular triad* if and only if $|(\operatorname{curl} X)(i, j, k)| = 3$. If G is a complete graph, the number of circular triads has been shown [24] to be

$$N = \frac{n}{24}(n^2 - 1) - \frac{1}{8}\sum\nolimits_i \Big[\sum\nolimits_j X_{ij}\Big]^2.$$

For ranking data given in terms of cardinal scores and that is generally incomplete, curl plays an extended role in addition to just measuring the triangular inconsistency.

Definition 2.3. *Let $X : V \times V \to \mathbb{R}$ be a pairwise ranking edge flow on a pairwise comparison graph $G = (V, E)$.*

(1) X is called ***consistent*** *or* ***curl-free*** *on $\{i, j, k\} \in T(E)$ if*

$$(\operatorname{curl} X)(i, j, k) = X_{ij} + X_{jk} + X_{ki} = 0.$$

Note that this implies that $\operatorname{curl}(X)(\sigma(i), \sigma(j), \sigma(k)) = 0$ for every permutation σ.

(2) X is called ***globally consistent*** *if it is given by the gradient of a score function, i.e.*

$$X = \operatorname{grad} s \quad \text{for some } s : V \to \mathbb{R}.$$

(3) X is called ***locally consistent*** *or* ***triangularly consistent*** *if it is curl-free on every triangle in $T(E)$, i.e. every 3-clique of G.*

Clearly any gradient flow must be curl-free everywhere, i.e. the well-known identity in vector calculus

$$\operatorname{curl} \circ \operatorname{grad} = 0$$

is also true for combinatorial curl and gradient (a special case of Lemma 4.4). A qualified converse may be deduced from the Hodge decomposition theorem (see also Theorem 5.2): a curl-free flow on a complete graph must necessarily be a

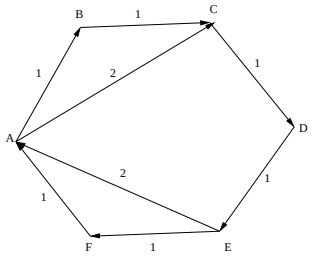

FIGURE 1. A harmonic pairwise ranking, which is locally consistent on every triangles but inconsistent along the loop $A \to B \to C \to D \to E \to F \to A$.

gradient flow, or putting it another way, a locally consistent pairwise ranking must necessarily be a globally consistent pairwise ranking when there are no missing values, i.e. if the pairwise comparison graph is a complete graph (every pair of alternatives has been compared).

When G is an incomplete graph, the condition that X is curl-free on every triangle in the graph will not be enough to guarantee that it is a gradient flow. The reason lies in that curl only takes into account the triangular inconsistency; but since there are missing edges in the pairwise comparison graph G, it is possible that non-triangular cyclic rankings of lengths greater than three can occur. For example, Figure 1 shows a pairwise ranking which is locally consistent on every triangle but globally inconsistent, since it contains a cyclic ranking of length six. Fortunately, Hodge decomposition theorem will tell us, all such cyclic rankings lie in a subspace of *harmonic rankings*, which will be characterized by the kernel of some combinatorial Laplacians.

3. A Matrix Theoretic View of Hodge Decomposition

We will see in this section that edge flows, gradient flows, harmonic flows, and curl flows can all be represented as specially structured skew-symmetric matrices. In this framework, the Hodge decomposition theorem may be viewed as an orthogonal direct sum decomposition of the space of skew-symmetric matrices into three subspaces. A formal treatment of combinatorial Hodge theory will be given in Section 4.

Recall that a matrix $X \in \mathbb{R}^{n\times n}$ is said to be *skew-symmetric* if $X_{ij} = -X_{ji}$ for all $i, j \in V := \{1, \dots, n\}$. One knows from linear algebra that any square matrix A may be written uniquely as a sum of a symmetric and a skew-symmetric matrix,

$$A = \tfrac{1}{2}(A + A^\top) + \tfrac{1}{2}(A - A^\top).$$

We will let[3] this set by $\mathcal{A}$, i.e.

$$\mathcal{A} := \{X \in \mathbb{R}^{n\times n} \mid X^\top = -X\}, \quad \text{and} \quad \mathcal{S} := \{X \in \mathbb{R}^{n\times n} \mid X^\top = X\}$$

It is perhaps interesting to note that semidefinite programming takes place in the cone of symmetric positive definite matrices in $\mathcal{S}$ but the optimization techniques in this takes place in the exterior space $\mathcal{A}$.

[3]The more common notations are $\mathfrak{so}_n(\mathbb{R})$ (Lie algebra of $\mathrm{SO}(n)$) and $\wedge^2(\mathbb{R}^n)$ (second exterior product of $\mathbb{R}^n$) but we avoided these since we use almost no Lie theory and exterior algebra.

One simple way to construct a skew-symmetric matrix is to take a vector $s = [s_1, \dots, s_n]^\top \in \mathbb{R}^n$ and define X by

$$X_{ij} := s_i - s_j.$$

Note that if $X \neq 0$, then $\operatorname{rank}(X) = 2$ since it can be expressed as $se^\top - es^\top$ with $e := [1, \dots, 1]^\top \in \mathbb{R}^n$. These are in a sense the simplest type of skew-symmetric matrices — they have the lowest rank possible for a non-zero skew-symmetric matrix (recall that the rank of a skew-symmetric matrix is necessarily even). In this paper, we will call these *gradient* matrices and denote them collectively by $\mathcal{M}_G$,

$$\mathcal{M}_G := \{X \in \mathcal{A} \mid X_{ij} = s_i - s_j \text{ for some } s \in \mathbb{R}^n\}.$$

For $T \subseteq \binom{V}{3}$, we define the set of T-*consistent* matrices as

$$\mathcal{M}_T := \{X \in \mathcal{A} \mid X_{ij} + X_{jk} + X_{ki} = 0 \text{ for all } \{i, j, k\} \in T\}. \tag{12}$$

We can immediately observe every $X \in \mathcal{M}_G$ is T-*consistent* for any $T \subseteq \binom{V}{3}$, i.e. $\mathcal{M}_G \subseteq \mathcal{M}_T$. Conversely, a matrix X that satisfies

$$X_{ij} + X_{jk} + X_{ki} = 0 \quad \text{for every triple } \{i, j, k\} \in \binom{V}{3}.$$

is necessarily a gradient matrix, i.e.

$$\mathcal{M}_G = \mathcal{M}_{\binom{V}{3}}. \tag{13}$$

Given $T \subseteq \binom{V}{3}$, it is straight forward to verify that both $\mathcal{M}_G$ and $\mathcal{M}_T$ are subspaces of $\mathbb{R}^{n \times n}$. The preceding discussions then imply the following subspace relations:

$$\mathcal{M}_G \subseteq \mathcal{M}_T \subseteq \mathcal{A}. \tag{14}$$

Since these are strict inclusions in general, several complementary subspaces arises naturally. With respect to the usual inner product $\langle X, Y \rangle = \operatorname{tr}(X^\top Y) = \sum_{i,j} X_{ij} Y_{ij}$, we obtain orthogonal complements of $\mathcal{M}_G$ and $\mathcal{M}_T$ in $\mathcal{A}$ as well as the orthogonal complement of $\mathcal{M}_G$ in $\mathcal{M}_T$, which we denote by $\mathcal{M}_H$:

$$\mathcal{A} = \mathcal{M}_G \oplus \mathcal{M}_G^\perp, \qquad \mathcal{A} = \mathcal{M}_T \oplus \mathcal{M}_T^\perp, \qquad \mathcal{M}_T = \mathcal{M}_G \oplus \mathcal{M}_H.$$

We will call the elements of $\mathcal{M}_H$ *harmonic* matrices as we shall see that they are discrete analogues of solutions to the Laplace equation (or, more accurately, the Helmholtz equation). An alternative characterization of $\mathcal{M}_H$ is

$$\mathcal{M}_H = \mathcal{M}_T \cap \mathcal{M}_G^\perp,$$

which may be viewed as a discrete analogue of the condition of being simultaneously curl-free and divergence-free. More generally, this discussion applies to any weighted inner product $\langle X, Y \rangle_w = \sum_{i,j} w_{ij} X_{ij} Y_{ij}$. The five subspaces $M_g, M_T, M_h, M_T^\perp, M_g^\perp$ of $\mathcal{A}$ play a central role in our techniques. As we shall see later, the Helmholtz decomposition in Theorem 4.8 may be viewed as the orthogonal direct sum decomposition

$$\mathcal{A} = \mathcal{M}_G \oplus \mathcal{M}_H \oplus \mathcal{M}_T^\perp.$$

4. Combinatorial Hodge Theory

In this section we will give a brief introduction to combinatorial Hodge theory, paying special attention to its relevance in rank learning. One may wonder why we do not rely on our relatively simple matrix view in Section 3. The reasons are two fold: firstly, important geometric insights are lost when the actual motivations behind the matrix approach are disregarded; and secondly, the matrix picture describes only the case of 2-dimensional simplicial complex but combinatorial Hodge theory extends to any k-dimensional simplicial complex. While so far we did not use any simplicial complex of dimension higher than 2 in our study of rank learning, it is conceivable that higher-dimensional simplicial complex could play a role in future studies.

4.1. **Extension of Pairwise Comparison Graph to Simplicial Complex.** Let $G = (V, E)$ be a pairwise comparison graph. To characterize the triangular inconsistency or curl, one needs to study the triangles formed by the 3-cliques[4], i.e. the set

$$T(E) := \{\{i, j, k\} \in \tbinom{V}{3} \mid \{i, j\}, \{j, k\}, \{k, i\} \in E\}.$$

A combinatorial object of the form (V, E, T) where $E \subseteq \binom{V}{2}$, $T \subseteq \binom{V}{3}$, and $\{i, j\}, \{j, k\}, \{k, i\} \in E$ for all $\{i, j, k\} \in T$ is called a 2-dimensional simplicial complex. This is a generalization of the notion of a graph, which is a 1-dimensional simplicial complex. In particular, given a graph $G = (V, E)$, the 2-dimensional simplicial complex $(V, E, T(E))$ is called the 3-*clique complex* of G.

More generally, a *simplicial complex* (V, Σ) is a vertex set $V = \{1, \dots, n\}$ together with a collection Σ of subsets of V that is closed under inclusion, i.e. if $\tau \in \Sigma$ and $\sigma \subset \tau$, then $\sigma \in \Sigma$. The elements in Σ are called *simplices*. For example, a 0-simplex is just an element $\{i\} \in V$, a 1-simplex is a pair $\{i, j\} \in \binom{V}{2}$, a 2-simplex is a triple $\{i, j, k\} \in \binom{V}{3}$, and so on. For $k \leq |V|$, a k-simplex is a k-element set in $\binom{V}{k}$. $\Sigma_k \subset \binom{V}{k}$ will denote the set of all $(k-1)$-simplices in Σ. In the previous paragraph, $\Sigma_2 = E$, $\Sigma_3 = T$, and $\Sigma = E \cup T$. In general, given any undirected graph $G = (V, E)$, one obtains a $(k-1)$-dimensional simplicial complex $K_G^k := (V, \Sigma_k)$ called the k-*clique complex*[5] of G by 'filling in' all its j-cliques for $j = 1, \dots, k$, or more precisely, by setting $\Sigma = \{j\text{-cliques of } G \mid j = 1, \dots, k\}$. The k-clique complex of G where k is maximal is just called the clique complex of G and denoted K_G.

In this paper, we will mainly concern ourselves with studying the 3-clique complex $K_G^3 = (V, E, T(E))$ where G is a pairwise comparison graph. Note that we could also look at the simplicial complex $(V, E, T_\gamma(E))$ where $T_\gamma := \{\{i, j, k\} \in T(E) \mid |X_{ij} + X_{jk} + X_{ki}| \leq \gamma\}$ where $0 \leq \gamma \leq \infty$. For $\gamma = \infty$, we get K_G^3 but for general γ we get a subcomplex of K_G^3. We have found this to be a useful multiscale characterization of the inconsistencies of pairwise rankings but the detailed discussion will have to be left to a future paper.

4.2. **Cochain, Coboundary Maps, and Combinatorial Laplacians.** We will now introduce some discrete exterior calculus on a simplicial complex where potential functions (global ranking), edge flow (pairwise ranking), triangular flow (triplewise ranking), gradient (global ranking), curl (local inconsistency) become

[4]Recall that a k-clique of G is just a complete subgraph of G with k vertices.

[5]Note that a k-clique is a $(k-1)$-simplex.

just special cases of a much more general framework. We will now also define the notions of *combinatorial divergence* and *combinatorial Laplacians*. A 0-dimensional combinatorial Laplacian is just the usual graph Laplacian but the case of greatest interest to us the 1-dimensional combinatorial Laplacian, or what we will call the *graph Helmholtzian*.

Definition 4.1. *Let K be a simplicial complex and recall that Σ_k denotes its set of k-simplices. A k-dimensional* ***cochain*** *is a function $f : \Sigma_k \to \mathbb{R}$ that is alternating on each of the k-simplex, i.e.*

$$f((i_{\sigma(0)}, \dots, i_{\sigma(k)}) = \operatorname{sign}(\sigma) f(i_0, \dots, i_k),$$

for all $\{i_0, \dots, i_n\} \in \Sigma_k$ and all $\sigma \in \mathfrak{S}_{k+1}$, the permutation group on $k+1$ elements. The set of all k-cochains on K is denoted $C^k(K, \mathbb{R})$.

For simplicity we will often just write C^k for $C^k(K, \mathbb{R})$. In particular, C^0 is the space of potential functions (global ranking), C^1 is the space of edge flows (pairwise rankings), and C^2 is the space of triangular flows (triplewise rankings).

The k-cochain space C^k can be given a choice of inner product. In view of the weighted l_2-minimization for our rank learning problem (8), we will define the following inner product on C^1,

$$\langle X, Y \rangle_w = \sum\nolimits_{\{i,j\} \in E} w_{ij} X_{ij} Y_{ij}, \tag{15}$$

for all edge flows $X, Y \in C^1$. In the context of a pairwise comparison graph G, it may not be immediately clear why this defines an inner product since we have noted after (9) that $W = [w_{ij}]$ is only a nonnegative matrix and it is possible that some entries are 0. However, observe that by definition $w_{ij} = 0$ iff no voters have rated both alternatives i and j and therefore $\{i, j\} \notin E$ by (4) and so any edge flow X will automatically have $X_{ij} = 0$ by (6). Hence we indeed have that $\langle X, X \rangle_w = 0$ iff $X = 0$, as required for an inner product (the other properties are trivial to check).

The operators grad and curl are all special instances of coboundary maps as defined below.

Definition 4.2. *The kth* ***coboundary*** *operator $\delta_k : C^k(K, \mathbb{R}) \to C^{k+1}(K, R)$ is the linear map that takes a k-cochain $f \in C^k$ to a $(k+1)$-cochain $\delta_k f \in C^{k+1}$ defined by*

$$(\delta_k f)(i_0, i_1, \dots, i_{k+1}) = \sum\nolimits_{j=0}^{k+1} (-1)^j f(i_0, \dots, i_{j-1}, i_{j+1}, \dots, i_{k+1}).$$

Note that i_j is omitted from jth term in the sum. Coboundary maps compute an alternating difference with one input left out. In particular, $\delta_0 = \operatorname{grad}$, i.e. $(\delta_0 s)(i, j) = s_j - s_i$, and $\delta_1 = \operatorname{curl}$, i.e. $(\delta_1 X)(i, j, k) = X_{ij} + X_{jk} + X_{ik}$.

Given a choice of an inner product $\langle \cdot, \cdot \rangle_k$ on C^k, we may define the adjoint operator of the coboundary map, $\delta_k^* : C^{k+1} \to C^k$ in the usual manner, i.e. $\langle \delta_k f_k, g_{k+1} \rangle_{k+1} = \langle f_k, \delta_k^* g_{k+1} \rangle_k$.

Definition 4.3. *The* ***combinatorial divergence*** *operator* $\operatorname{div} : C^1(K, \mathbb{R}) \to C^0(K, \mathbb{R})$ *is the adjoint of $\delta_0 = \operatorname{grad}$, i.e.*

$$\operatorname{div} := -\delta_0^*. \tag{16}$$

Divergence will appear in the minimum norm solution to (8) and will be used to characterize $\mathcal{M}_2^\perp$ explicitly. As usual, we will drop the adjective 'combinatorial' when there is no cause for confusion.

For rank learning, it suffices to consider the cases $k = 0, 1, 2$. Let G be a pairwise comparison graph and K_G its clique complex[6]. The cochain maps,

$$C^0(K_G, \mathbb{R}) \xrightarrow{\delta_0} C^1(K_G, \mathbb{R}) \xrightarrow{\delta_1} C^2(K_G, \mathbb{R}) \tag{17}$$

and their adjoints,

$$C^0(K_G, \mathbb{R}) \xleftarrow{\delta_0^*} C^1(K_G, \mathbb{R}) \xleftarrow{\delta_1^*} C^2(K_G, \mathbb{R}), \tag{18}$$

have the following ranking theoretic interpretation with C^0, C^1, C^2 representing the spaces of global rankings, pairwise rankings, and triplewise rankings respectively,

$$\textit{global} \xrightarrow{\text{grad}} \textit{pairwise} \xrightarrow{\text{curl}} \textit{triplewise},$$

$$\textit{global} \xleftarrow{-\text{div}=\text{grad}^*} \textit{pairwise} \xleftarrow{\text{curl}^*} \textit{triplewise}.$$

In summary, the formulas for combinatorial gradient, curl, and divergence are given by

$$\begin{aligned}
(\operatorname{grad} s)(i,j) &= (\delta_0 s)(i,j) = s_j - s_i,\\
(\operatorname{curl} X)(i,j,k) &= (\delta_1 X)(i,j,k) = X_{ij} + X_{jk} + X_{ki},\\
(\operatorname{div} X)(i) &= -(\delta_0^* X)(i) = \textstyle\sum_{j \text{ s.t. } \{i,j\}\in E} w_{ij} X_{ij}
\end{aligned}$$

with respect to the inner product $\langle X, Y\rangle_w = \sum_{\{i,j\}\in E} w_{ij} X_{ij} Y_{ij}$ on C^1.

As an aside, it is perhaps worth pointing out that there is no special name for the adjoint of curl coming from physics because in 3-space, C^1 may be identified with C^2 via a property called *Hodge duality* and in which case curl is a self-adjoint operator, i.e. $\text{curl}^* = \text{curl}$. This will not be true in our case.

If we represent functions on vertices by n-vectors, edge flows by $n \times n$ skew-symmetric matrices, and triangular flows by $n \times n \times n$ skew-symmetric hypermatrices, i.e.

$$\begin{aligned}
C^0 &= \mathbb{R}^n,\\
C^1 &= \{[X_{ij}] \in \mathbb{R}^{n\times n} \mid X_{ij} = -X_{ji}\} = \mathcal{A},\\
C^2 &= \{[\Phi_{ijk}] \in \mathbb{R}^{n\times n\times n} \mid \Phi_{ijk} = \Phi_{jki} = \Phi_{kij} = -\Phi_{jik} = -\Phi_{ikj} = -\Phi_{kji}\},
\end{aligned}$$

then in the language of linear algebra introduced in Section 3, we have the following correspondence

$$\begin{aligned}
\operatorname{im}(\delta_0) &= \operatorname{im}(\text{grad}) = \mathcal{M}_G, \qquad & \ker(\delta_1) &= \ker(\text{curl}) = \mathcal{M}_T,\\
\ker(\delta_0^*) &= \ker(\text{div}) = \mathcal{M}_G^\perp, \qquad & \operatorname{im}(\delta_1^*) &= \operatorname{im}(\text{curl}^*) = \mathcal{M}_T^\perp,
\end{aligned}$$

where $T = T(E)$.

Coboundary maps have the following important closedness property.

Lemma 4.4 (Closedness)**.** $\delta_{k+1} \circ \delta_k = 0$.

For $k = 0$, this and its adjoint are well-known identities in vector calculus,

$$\text{curl} \circ \text{grad} = 0, \qquad \text{div} \circ \text{curl}^* = 0.$$

Ranking theoretically, the first identity simply says that a global ranking must be consistent.

[6]It does not matter whether we consider K_G or K_G^3 or indeed any K_G^k where $k \geq 3$; the higher-dimensional k-simplices where $k \geq 3$ do not play a role in the coboundary maps $\delta_0, \delta_1, \delta_2$.

We will now define combinatorial Laplacians, higher-dimensional analogues of the graph Laplacian.

Definition 4.5. *Let K be a simplicial complex. The k-dimensional* ***combinatorial Laplacian*** *is the operator $\Delta_k : C^k(K,\mathbb{R}) \to C^k(K,\mathbb{R})$ defined by*

$$\Delta_k = \delta_k^* \circ \delta_k + \delta_{k-1} \circ \delta_{k-1}^*. \tag{19}$$

In particular, for $k = 0$,

$$\Delta_0 = \delta_0^* \circ \delta_0 = \operatorname{div} \circ \operatorname{grad}$$

is a discrete analogue of the scalar Laplacian or Laplace operator while for $k = 1$,

$$\Delta_1 = \delta_1^* \circ \delta_1 + \delta_0 \circ \delta_0^* = \operatorname{curl}^* \circ \operatorname{curl} - \operatorname{div} \circ \operatorname{grad}$$

is a discrete analogue of the vector Laplacian or Helmholtz operator. In the context of graph theory, if $K = K_G$, then Δ_0 is called the graph Laplacian [9] while Δ_1 is called the *graph Helmholtzian*.

The combinatorial Laplacian has some well-known, important properties.

Lemma 4.6. *Δ_k is a positive semidefinite operator. Furthermore, the dimension of $\ker(\Delta_k)$ is equal to kth Betti number of K.*

We will call a cochain $f \in \ker(\Delta_k)$ *harmonic* since they are solutions to higher-dimensional analogue of the Laplace equation

$$\Delta_k f = 0.$$

Strictly speaking, the Laplace equation refers to $\Delta_0 f = 0$. The equation $\Delta_1 X = 0$ is really the *Helmholtz equation*. But nonetheless, we will still call an edge flow $X \in \ker(\Delta_1)$ a harmonic flow.

4.3. **Hodge Decomposition Theorem.** We now state the main theorem in combinatorial Hodge theory.

Theorem 4.7 (Hodge Decomposition Theorem)**.** *$C^k(K,\mathbb{R})$ admits an orthogonal decomposition*

$$C^k(K,\mathbb{R}) = \operatorname{im}(\delta_{k-1}) \oplus \ker(\Delta_k) \oplus \operatorname{im}(\delta_k^*).$$

Furthermore,

$$\ker(\Delta_k) = \ker(\delta_k) \cap \ker(\delta_{k-1}^*).$$

An elementary proof targeted at a computer science readership may be found in [16]. For completeness we include a proof here.

Theorem 4.7. We will use Lemma 4.4. First, $C^k = \operatorname{im}(\delta_{k-1}) \oplus \ker(\delta_{k-1}^*)$. Since $\delta_k \delta_{k-1} = 0$, taking adjoint yields $\delta_{k-1}^* \delta_k^* = 0$, which implies that $\operatorname{im}(\delta_k^*) \subseteq \ker(\delta_{k-1}^*)$. Therefore $\ker(\delta_{k-1}^*) = [\operatorname{im}(\delta_k^*) \oplus \ker(\delta_k)] \cap \ker(\delta_{k-1}^*) = [\operatorname{im}(\delta_k^*) \cap \ker(\delta_{k-1}^*)] \oplus [\ker(\delta_k) \cap \ker(\delta_{k-1}^*)] = \operatorname{im}(\delta_k^*) \oplus [\ker(\delta_k) \cap \ker(\delta_{k-1}^*)]$. It remains to show that $\ker(\delta_k) \cap \ker(\delta_{k-1}^*) = \ker(\Delta_k) = \ker(\delta_{k-1}\delta_{k-1}^* + \delta_k^* \delta_k)$. Clearly $\ker(\delta_k) \cap \ker(\delta_{k-1}^*) \subseteq \ker(\Delta_k)$. For any $X = \delta_k^* \Phi \in \operatorname{im}(\delta_k^*)$ where $0 \neq \Phi \in C^{k+1}$, Lemma 4.4 again implies $\delta_{k-1}\delta_{k-1}^* X = \delta_{k-1}\delta_{k-1}^* \delta_k^* \Phi = 0$, but $\delta_k^* \delta_k X = \delta_k^* \delta_k \delta_k^* \Phi \neq 0$, which implies that $\Delta_k X \neq 0$. Similarly for $X \in \operatorname{im}(\delta_0)$. Hence $\ker(\Delta_k) = \ker(\delta_k) \cap \ker(\delta_{k-1}^*)$. □ □ □

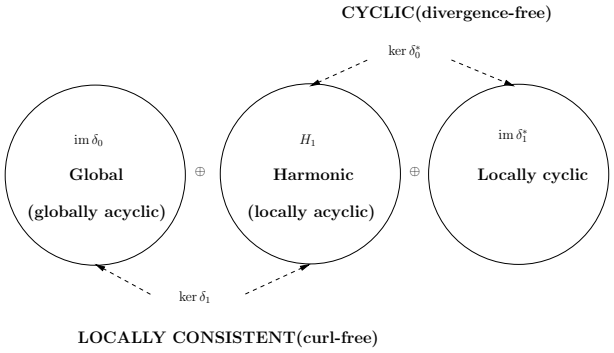

FIGURE 2. Hodge/Helmholtz decomposition of pairwise rankings

While Hodge decomposition holds in general for any simplicial complex and in any dimension k, the case $k = 1$ is more often called the *Helmholtz decomposition theorem*[7]. We will state it here for the special case of interest to us.

Theorem 4.8 (Helmholtz Decomposition Theorem)**.** *Let $G = (V, E)$ be an undirected, unweighted graph and K_G be its clique complex. The space of edge flows on G, i.e. $C^1(K_G, \mathbb{R})$, admits an orthogonal decomposition*

$$\begin{aligned} C^1(K_G, \mathbb{R}) &= \operatorname{im}(\delta_0) \oplus \ker(\Delta_1) \oplus \operatorname{im}(\delta_1^*) \\ &= \operatorname{im}(\operatorname{grad}) \oplus \ker(\Delta_1) \oplus \operatorname{im}(\operatorname{curl}^*). \end{aligned} \tag{20}$$

Furthermore,

$$\ker(\Delta_1) = \ker(\delta_1) \cap \ker(\delta_0^*) = \ker(\operatorname{curl}) \cap \ker(\operatorname{div}). \tag{21}$$

The clique complex K_G above may be substituted with any K_G^k with $k \geq 3$ (see Footnote 6). The equation (21) says that an edge flow is harmonic iff it is both curl-free and divergence-free. Figure 4.3 illustrates (20).

To understand the significance of this theorem, we need to discuss the ranking theoretic interpretations of each subspace in the theorem.

(1) $\operatorname{im}(\delta_0) = \operatorname{im}(\operatorname{grad})$ denotes the subspace of pairwise rankings that are the gradient flows of score functions. Thus this subspace comprises the *globally consistent* or *acyclic* pairwise rankings. Given any pairwise ranking from this subspace, we may determine a score function on the alternatives that is unique up to an additive constant[8] and then we may rank all alternatives globally in terms of their scores.
(2) $\ker(\delta_0^*) = \ker(\operatorname{div})$ denotes the subspace of *divergence-free* pairwise rankings, whose total in-flow equals total out-flow for each alternative $i \in V$. Such pairwise rankings may be regarded as *cyclic* rankings, i.e. rankings of the form $i \succeq j \succeq k \succeq \cdots \succeq i$, and they are clearly *inconsistent*. Since $\ker(\operatorname{div}^*) = \operatorname{im}(\operatorname{grad})^\perp$, cyclic rankings have zero projection on global rankings.

[7]For a simply connected manifold, the continuous version of the Helmholtz decomposition theorem is just the fundamental theorem of vector calculus.

[8]Note that $\ker(\delta_0) = \ker(\operatorname{grad})$ is the set of constant functions on V and so $\operatorname{grad}(s) = \operatorname{grad}(s + \text{constant})$.

(3) $\ker(\delta_1) = \ker(\text{curl})$ denotes the subspace of *curl-free* pairwise rankings with zero flow-sum along any triangle in K_G. This corresponds to *locally consistent* (also known as triangularly consistent) pairwise rankings. Note that by the Closedness Lemma $\text{curl} \circ \text{grad} = 0$ and so $\text{im}(\text{grad}) \subseteq \ker(\text{curl})$, whence the globally consistent pairwise rankings induced by gradient flows of score functions only account for a subset of locally consistent rankings. The remaining ones are the locally consistent rankings that are not globally consistent and they are precisely the harmonic rankings discussed below.

(4) $\ker(\Delta_1) = \ker(\text{curl}) \cap \ker(\text{div})$ denotes the subspace of *harmonic* pairwise rankings, or just harmonic rankings in short. It is the space of solutions to the Helmholtz equation. Harmonic rankings are exactly those pairwise rankings that are both *curl-free* and *divergence-free*. These are only locally consistent with zero curl on every triangle in $T(E)$ but are not globally consistent. In other words, while there are no inconsistencies due to small loops of length 3, i.e. $i \succeq j \succeq k \succeq i$, there are inconsistencies along large loops of lengths > 3, i.e. $a \succeq b \succeq c \succeq \cdots \succeq z \succeq a$. So these are also cyclic rankings. Rank aggregation on $\ker(\Delta_1)$ depends on the edge paths traversed in the simplicial complex; along homotopy equivalent paths one obtains consistent rankings. Figure 1 gives an example of harmonic rankings.

(5) $\text{im}(\delta_1^*) = \text{im}(\text{curl}^*)$ denotes the subspace of *locally cyclic* pairwise rankings that have non-zero curls along triangles. By the Closedness Lemma, $\text{im}(\text{curl}^*) \subseteq \ker(\text{div})$ and so this subspace is in general a proper subspace of the divergence-free rankings; the orthogonal complement of $\text{im}(\text{curl}^*)$ in $\ker(\text{div})$ is precisely the space of harmonic rankings $\ker(\Delta_1)$ discussed above.

5. Implications of Hodge Theory

We now state two immediate implications of the Helmholtz decomposition theorem when applied to rank learning. The first implication is that it gives an interpretation of the solution and residual of the optimization problem (8); these are respectively the l_2-projection on gradient flows and divergence-free flows. In the context of rank learning and in the l_2-sense, the solution to (8) gives the nearest globally consistent pairwise ranking to the data while the residual gives the sum total of all inconsistent components (both local and harmonic) in the data. The second implication is the condition that local consistency guarantees global consistency whenever there is no harmonic component in the data (which happens iff the clique complex of the pairwise comparison graph is 'loop-free').

5.1. **Structure Theorem for Global Ranking and the Residual of Inconsistency.** In order to cast our optimization problem (8) in the Hodge theoretic framework, we need to specify relevant inner products on C^0, C^1, C^2. As before, the inner product on the space of edge flows (pairwise rankings) C^1 will be a weighted Euclidean inner product

$$\langle X, Y \rangle_w = \sum\nolimits_{\{i,j\} \in E} w_{ij} X_{ij} Y_{ij}$$

for $X, Y \in C^1$. We will let the inner products on C^0 and C^2 be the unweighted Euclidean inner product

$$\langle r, s \rangle = \sum\nolimits_{i=1}^{n} r_i s_i, \qquad \langle \Theta, \Phi \rangle = \sum\nolimits_{\{i,j,k\} \in T(E)} \Theta_{ijk} \Phi_{ijk}$$

for $r, s \in C^0$ and $\Theta, \Phi \in C^2$. We note that other inner products can be chosen (e.g. the inner products on C^0 and C^2 could have been weighted) with corresponding straightforward modification of (8) but this would not change the essential nature of our methods. We made the above choices mainly to keep our notations uncluttered.

The optimization problem (8) is then equivalent to an l_2-projection of an edge flow representing a pairwise ranking onto im(grad),

$$\min_{s\in C^0} \|\delta_0 s - \bar{Y}\|_{2,w} = \min_{s\in C^0} \|\operatorname{grad} s - \bar{Y}\|_{2,w},$$

The Helmholtz decomposition theorem then leads to the following results about the structures of the solutions and residuals of (8).

Theorem 5.1. *(i) Solutions of* (8) *satisfy the following normal equation*

$$\Delta_0 s = -\operatorname{div} \bar{Y}, \tag{22}$$

where the minimum norm solution is given by

$$s^* = -\Delta_0^\dagger \operatorname{div} \bar{Y} \tag{23}$$

where † *indicates a Moore-Penrose inverse. The divergence in* (23) *is given by*

$$(\operatorname{div} \bar{Y})(i) = \sum\nolimits_{j \text{ s.t. } \{i,j\}\in E} w_{ij} \bar{Y}_{ij},$$

and the matrix representing the graph Laplacian is given by

$$[\Delta_0]_{ij} = \begin{cases} \sum_i w_{ii} & \text{if } j = i, \\ -w_{ij} & \text{if } j \text{ is such that } \{i,j\} \in E, \\ 0 & \text{otherwise.} \end{cases}$$

(ii) The residual $R^* = \delta_0 s^* - \bar{Y}$ is divergence-free, i.e. $\operatorname{div} R^* = 0$. *Moreover, it has a further orthogonal decomposition*

$$R^* = \operatorname{proj}_{\operatorname{im}(\operatorname{curl}^*)} \bar{Y} + \operatorname{proj}_{\ker(\Delta_1)} \bar{Y}, \tag{24}$$

where $\operatorname{proj}_{\operatorname{im}(\operatorname{curl}^*)} \bar{Y}$ *is a local cyclic ranking accounting for local inconsistencies and* $\operatorname{proj}_{\ker(\Delta_1)} \bar{Y}$ *is a harmonic ranking accounting for global inconsistencies. In particular, the first projection is given by*

$$\operatorname{proj}_{\operatorname{im}(\operatorname{curl}^*)} \bar{Y} = \operatorname{curl}^* (\operatorname{curl} \operatorname{curl}^*)^\dagger \operatorname{curl} \bar{Y}. \tag{25}$$

Theorem 5.1. The l_2-projection may be written as

$$\min_{s\in C^0} \|\delta_0 s - \bar{Y}\|_{2,w}^2 = \min_{s\in C^0} \langle \delta_0 s - \bar{Y}, \delta_0 s - \bar{Y} \rangle.$$

The condition for a stationary point then gives the normal equation

$$\delta_0^* \delta_0 s = \delta_0 \bar{Y}.$$

(i) follows from this upon substituting $\Delta_0 = \delta_0^* \delta_0$ and $\operatorname{div} = -\delta_0^*$. (ii) follows from the Helmholtz decomposition theorem. □ □ □

In the special case when the pairwise ranking matrix G is a complete graph and we have an unweighted Euclidean inner product on C^1, the minimum norm solution s^* in (23) satisfies $\sum_i s_i^* = 0$ and is given by

$$s_i^* = -\frac{1}{n} \operatorname{div}(\bar{Y})(i) = -\frac{1}{n} \sum\nolimits_j \bar{Y}_{ij}. \tag{26}$$

In Section 7, we shall see that this is the well-known *Borda count* in social choice theory, a measure that is also widely used in psychology and statistics [24, 26, 27, 28, 11]. Since G is a complete graph only when the ranking data is complete, i.e. every voter has rated every alternative, this is an unrealistic scenario for the type of modern ranking data discussed in Section 1. Among other things, the Hodge theoretic framework generalizes Borda count to scenarios where the ranking data is incomplete or even highly incomplete.

In (ii) the triangular cyclic ranking component is obtained by solving

$$\min_{\Phi \in C^2} \|R^* - \operatorname{curl}^* \Phi\|_{2,w} = \min_{\Phi \in C^2} \|\bar{Y} - \operatorname{curl}^* \Phi\|_{2,w}.$$

The minimum norm solution is given by

$$\Phi^* = (\delta_1 \circ \delta_1^*)^\dagger \delta_1 \bar{Y} = (\operatorname{curl} \circ \operatorname{curl}^*)^\dagger \operatorname{curl} \bar{Y}$$

and the required component is given by $\operatorname{proj}_{\operatorname{im}(\operatorname{curl}^*)} \bar{Y} = \operatorname{curl}^* \Phi^*$. An actual computation requires solving a least squares problem of size $|T| \times |T|$. Since $T \sim O(n^3)$, the least squares problem incurs a computation cost that is of the order $O(n^9)$. Contrast this with the problem of solving for a nearest global ranking (23), which only requires the solution of an $n \times n$ least squares problem and thus has a more manageable $O(n^3)$ cost. Hence we expect it to be harder to isolate the harmonic component of the ranking data than to isolate the global component.

The l_2-residual R^*, being divergence-free, is a cyclic ranking. The subset of C^1 given by

$$\{R^* - \operatorname{curl}^* \Phi \mid \Phi \in C^2\}$$

is called the *homology class* of R^*. The harmonic ranking $\operatorname{proj}_{\ker(\Delta_1)} \bar{Y}$ is just one element in this class. In general, it will be dense in the sense that it will be nonzero on almost every edge in E. This is due to the additional curl-free condition imposed on harmonic rankings. To better understand the structure of R^*, it is often helpful to look for elements in the same homology class with the *sparsest support*, i.e.

$$\min_{\Phi \in C^2} \|R^* - \operatorname{curl}^* \Phi\|_0 = \min_{\Phi \in C^2} \|\operatorname{proj}_{\ker(\Delta_1)} \bar{Y} - \operatorname{curl}^* \Phi\|_0.$$

The widely used convex relaxation replacing the l_0-'norm' by the l_1-norm may be employed [19], i.e.

$$\min_{\Phi \in C^2} \|R^* - \operatorname{curl}^* Z\|_1 := \min_{\Phi \in C^2} \sum\nolimits_{i,j} |R^*_{ij} - (\operatorname{curl}^* t)_{ij}|.$$

A solution $\tilde{\Phi}$ of such an l_1-minimization problem is expected to give a sparse $R^* - \operatorname{curl}^* \tilde{\Phi}$, which we call an l_1-*approximate sparse generator* of R^*, or equivalently, of $\operatorname{proj}_{\ker(\Delta_1)} \bar{Y}$. We will discuss them in detail in Section 6.2. The bottom line here is that we want to find the shortest cycles that represent the global inconsistencies and perhaps remove the corresponding edges in the pairwise comparison graph, in view of what we will discuss next in Section 5.2. One plausible strategy to get a globally consistent ranking is to remove a number of problematic 'conflicting' comparisons from the pairwise comparison graph. Since it is only reasonable to remove as few edges as possible, this translates to finding a homology class with the sparsest support. This is similar to the minimum feedback arc set approach discussed in Section 7.2.

5.2. **Local Consistency versus Global Consistency.** In this section, we discuss a useful result, that local consistency implies global consistency whenever the harmonic component is absent from the ranking data. Whether a harmonic component exists is dependent on the topology of the clique complex K_G^3. We will invoke the recent work of Kahle [20] on such topological properties of random graphs to argue that harmonic components are exceedingly unlikely to occur.

By Lemma 4.6, the dimension of $\ker(\Delta_1)$ is equal to the first Betti number $\beta_1(K)$ of the underlying simplicial complex K. In particular, we know that $\ker(\Delta_1) = 0$ if $\beta_1(K) = 0$, and so the harmonic component of any edge flow on K is automatically absent when $\beta_1(K) = 0$ (roughly speaking, $\beta_1(K) = 0$ means that K does not have any 1-dimensional holes). This leads to the following result.

Theorem 5.2. *Let $K_G^3 = (V, E, T(E))$ be a 3-clique complex of a pairwise comparison graph $G = (V, E)$. If K_G^3 does not contain any 1-loops, i.e. $\beta_1(K_G^3) = 0$, then every locally consistent pairwise ranking is also globally consistent. In other words, if the edge flow $X \in C^1(K_G, \mathbb{R})$ is curl-free, i.e.*

$$\operatorname{curl}(X)(i, j, k) = 0$$

for all $\{i, j, k\} \in T(E)$, then it is a gradient flow, i.e. there exists $s \in C^0(K_G, R)$ such that

$$X = \operatorname{grad} s.$$

Theorem 5.2. This follows from the Helmholtz decomposition theorem since $\operatorname{proj}_{\ker(\Delta_0)} Y = 0$ for all $Y \in C^1$ when $\beta_1(K_G^3) = 0$. □ □ □

When G is a complete graph, then we always have $\beta_1(K_G^3) = 0$ and this justifies the discussion after Definition 2.3 about the equivalence of local and global consistencies for complete pairwise comparison graphs. In general, G will be incomplete due to missing ranking data (not all voters have rated all alternatives) but as long as K_G^3 is loop-free, such a claim still holds. In finance, this theorem translates into the well-known result that "triangular arbitrage-free implies arbitrage-free." The theorem enables us to infer global consistency from a local condition — whether the ranking data is curl-free. We note that being curl-free is a strong condition. If we instead have "triangular transitivity" in the ordinal sense, i.e. $a \succeq b \succeq c$ implies $a \succeq c$, then there is no result analogous to Theorem 5.2.

At least for Erdös-Rényi random graphs, the Betti number β_1 could only be non-zero when the edges are neither too sparse nor too dense. The following result by Kahle [20] quantifies this statement. He showed that β_1 undergoes two phase transitions from 0 to nonzero and back to 0 as the density of edges grows.

Theorem 5.3 (Kahle 2006)**.** *For an Erdös-Rényi random graph $G(n, p)$ on n vertices where the edges are independently generated with probability p, its clique complex K_G almost always has $\beta_1(K_G) = 0$, except when*

$$\frac{1}{n^2} \ll p \ll \frac{1}{n}. \tag{27}$$

Without getting into a discussion about whether Erdös-Rényi random graphs are good models for pairwise ranking comparison graphs of real-world ranking data, we note that the Netflix pairwise comparison graph has a high probability of having $\beta_1(K_G) = 0$ if Kahle's result applies. Although the original customer-product rating matrix of the Netflix prize dataset is highly incomplete (more than 99% missing

values), its pairwise comparison graph is very dense (less than 0.22% missing edges). In other words, p (probability of an edge) and n (number of vertices) are both large and so (27) is not satisfied.

6. l_1-aspects of Hodge Theoretic Rank Learning

Hodge theory is by and large an l_2-theory: inner products on cochains, adjoints of coboundary operators, orthogonality of Hodge decomposition, are all naturally associated with (weighted or unweighted) l_2-norms. In this section, we will take an oblique approach and study the l_1-aspects of combinatorial Hodge theory in the context of rank learning, with robustness and parsimony (or sparsity) being our two obvious motivations. We will study two l_1-norm minimization problems: (1) the l_1-projection on gradient flows (globally consistent rankings), which we show to have a dual problem as correlation maximization over bounded divergence-free flows (cyclic rankings); (2) an l_1-relaxed approximation of sparse divergence-free flows (cyclic rankings) of the residual of the l_2-projection, which we show to have a dual problem as correlation maximization over bounded curl-free flows (locally consistent rankings). We observe that the primal versus dual relation is revealed as an 'im(grad) versus ker(div)' relation in first case and an 'im(curl*) versus ker(curl)' relation in the second case.

6.1. **Robust Rank Learning: l_1-projection on gradient flows.** We have briefly mentioned this problem in (10), Section 2, as an l_1-variation of the least squares model (8) for rank learning. Here we will derive a duality result for (10). As before, we assume a pairwise comparison graph $G = (V, E)$ and an edge flow $\bar{Y} \in C^1(K_G, \mathbb{R})$ that comes from our ranking data. Consider the following minimization problem,

$$\begin{array}{ll} \min & \|X - \bar{Y}\|_{1,w} \\ \text{s.t.} & X = \operatorname{grad} s, \\ & X = -X^\top. \end{array} \tag{28}$$

which may be regarded as the l_1-projection[9] of an edge flow $\bar{Y}$ onto the space of gradient flows,

$$\min_{s \in C^0} \|\operatorname{grad} s - \bar{Y}\|_{1,w} := \sum\nolimits_{\{i,j\} \in E} w_{ij} |s_j - s_i - \bar{Y}_{ij}|. \tag{29}$$

In other words, we attempt to find the nearest globally consistent ranking $\operatorname{grad} s$ to the pairwise ranking $\bar{Y}$ as measured by the l_1-norm. Such a norm is often employed in robust regression since its solutions will be relatively more robust to outliers or large deviations in the ranking data $\bar{Y}$ when compared to the l_2-norm in (8) [37, 10]. The computational cost paid in going from (8) to (28) is in replacing a linear least squares problem with a linear programming problem.

Recall that the minimum norm l_2-minimizer is given by $s^* = -(\Delta_0)^\dagger \operatorname{div} \bar{Y}$ and the l_2-residual is given by $R^* = \bar{Y} - \operatorname{grad} s^*$. Then

$$\min_{s \in C^0} \|\operatorname{grad} s - \bar{Y}\|_{1,w} = \min_{s' \in C^0} \|\operatorname{grad} s' - R^*\|_{1,w}$$

[9]The projection of a point X onto a closed subset S in a finite-dimensional norm space is simply the unique point $X_S \in S$ that is nearest to X in the norm.

where $s' = s - s^*$. It follows that the l_1-minimizers in (29) may be characterized by[10]

$$\operatorname{argmin}_{s\in C^0} \|\operatorname{grad} s - \bar{Y}\|_{1,w} = \operatorname{argmin}_{s\in C^0} \|\operatorname{grad} s - \bar{Y}\|_{2,w} + \operatorname{argmin}_{s'\in C^0} \|\operatorname{grad} s' - R^*\|_{1,w}.$$

The deviation from the minimum norm l_2-minimizer s^* is a "median gradient flow" extracted from the cyclic residual R^*, which moves the l_1-residual $\bar{Y} - \operatorname{grad}(s_1 + s_2)$ *outside* the space of divergence-free flows; here

$$s_1 \in \operatorname{argmin}_{s'\in C^0} \|\operatorname{grad} s' - R^*\|_{1,w} \quad \text{and} \quad s_2 \in \operatorname{argmin}_{s'\in C^0} \|\operatorname{grad} s' - R^*\|_{2,w}.$$

On the other hand, in the dual problem to (28), we search for a solution *inside* the space of divergence-free flows. More precisely, the dual form of the l_1-projection (28) searches within a space of bounded divergence-free flows for a flow that is maximally correlated with $\bar{Y}$. Before we state this theorem, we note that the inner product defined in (15) for skew-symmetric matrices representing edge flows,

$$\langle X, Y\rangle_w := \sum\nolimits_{\{i,j\}\in E} w_{ij} X_{ij} Y_{ij},$$

also defines an inner product over $\mathbb{R}^{n\times n}$ if the symmetric weight matrix $W = [w_{ij}]$ has no zero entries, i.e. $w_{ij} > 0$. We will assume that this is the case in the following theorem.

Theorem 6.1. *The l_1-projection problem* (28) *has the following dual problem,*

$$\begin{array}{ll} \max & \langle X, \bar{Y}\rangle_w \\ \text{s.t.} & |X_{ij}| \le 1, \\ & \operatorname{div} X = 0, \\ & X = -X^\top. \end{array} \tag{30}$$

Theorem 6.1. Let $n = |V|$. Consider

$$\min_{s\in C^0} \|\operatorname{grad} s - \bar{Y}\|_{1,w} = \min_{s\in C^0} \sum\nolimits_{\{i,j\}\in E} w_{ij} |(\delta_0 s)_{ij} - \bar{Y}_{ij}|.$$

Let $\gamma_{ij} = |(\operatorname{grad} s)_{ij} - \bar{Y}_{ij}|$. This leads to the following equivalent problem,

$$\begin{array}{ll} \min & \langle \mathbf{1}, \gamma\rangle_w = \sum_{\{i,j\}\in E} w_{ij}\gamma_{ij} \\ \text{s.t.} & \gamma \ge 0, \\ & \operatorname{grad} s - \bar{Y} \le \gamma, \\ & \operatorname{grad} s - \bar{Y} \ge -\gamma, \end{array}$$

where $\gamma = [\gamma_{ij}] \in \mathbb{R}^{n\times n}$ and $\mathbf{1} \in \mathbb{R}^{n\times n}$ is the matrix whose all entries are 1. We will let $W = [w_{ij}] \in \mathbb{R}^{n\times n}$ be the matrix of weights (assumed to be all positive). Note that these are all symmetric matrices. The inequalities in the constraints are in the elementwise sense.

[10]Recall that argmin refers to the *set* of all minimizers. The addition of sets here is just the usual Minkowski sum.

The Lagrangian becomes

$$\begin{aligned} L(s,\gamma,\lambda^{\pm},\mu) &= \sum\nolimits_{\{i,j\}\in E}[w_{ij}\gamma_{ij} - \lambda_{ij}^{+}(\gamma_{ij} - (\operatorname{grad} s)_{ij} + \bar{Y}_{ij}) \\ &\qquad\qquad - \lambda_{ij}^{-}(\gamma_{ij} + (\operatorname{grad} s)_{ij} - \bar{Y}_{ij}) - \mu_{ij}\gamma_{ij}] \\ &= \langle \mathbf{1}, \gamma\rangle_w - \langle V \circ \lambda^{+}, \gamma - \operatorname{grad} s + \bar{Y}\rangle_w \\ &\qquad\qquad - \langle V \circ \lambda^{-}, \gamma + \operatorname{grad} s - \bar{Y}\rangle_w - \langle V \circ \mu, \gamma\rangle_w \end{aligned}$$

where $V := [v_{ij}] \in \mathbb{R}^{n\times n}$ is defined by

$$v_{ij} = \begin{cases} w_{ij}^{-1} & \text{if } w_{ij} \neq 0 \\ 0 & \text{if } w_{ij} = 0, \end{cases}$$

and $A \circ B := [a_{ij}b_{ij}]$ denotes the Hadamard product, i.e. elementwise product, of two matrices A, B of same dimensions. $\lambda^{+}, \lambda^{-}, \mu \geq 0$ denotes the dual variables.

The saddle-point condition yields

$$\frac{\partial L}{\partial s} = \operatorname{grad}^{*}\left[V \circ (\lambda^{+} - \lambda^{-})\right] = 0,$$

(recall that $\operatorname{grad}^{*} = -\operatorname{div}$) and

$$\frac{\partial L}{\partial \gamma} = W \circ \mathbf{1} - \lambda^{+} - \lambda^{-} - \mu = 0.$$

Notice that from the complementary condition we have,

$$\lambda_{ij}^{+} \cdot \lambda_{ij}^{-} = 0.$$

Define a skew-symmetric matrix by

$$X_{ij} = -X_{ji} = \begin{cases} w_{ij}^{-1}(\lambda_{ij}^{-} - \lambda_{ij}^{+}) & \text{if } \{i,j\} \in E, \\ 0 & \text{otherwise.} \end{cases}$$

This leads to the following dual problem

$$\begin{array}{ll} \max & \langle X, \bar{Y}\rangle_w = \sum_{\{i,j\}\in E} w_{ij}X_{ij}\bar{Y}_{ij} \\ \text{s.t.} & |X_{ij}| \leq 1, \\ & \operatorname{div}(X) = 0, \\ & X = -X^{\top}, \end{array}$$

as required. □ □ □

Theorem 6.1 shows that for l_1-projections, the dual problem searches in the orthogonal complement of the primal domain. The primal search space is the space of gradient flows im(grad) while the dual search space is the space of divergence-free flows ker(div). Recall that for l_2-projections, gradient flow corresponds to the solution while divergence-free flow corresponds to the residual. So the solution-residual split in the l_2-setting is in this sense analogous to the primal-dual split in l_1-setting.

An optimal l_1-minimizer of (28) can only be decided up to a constant from the complementary conditions,

$$0 < |X_{ij}| < 1 \Rightarrow s_j - s_i = \bar{Y}_{ij}.$$

The constraint $\sum_i s_i = 0$ may be imposed to remove this extra degree of freedom.

6.2. **Conflict resolution: l_1-minimization for approximate sparse cyclic rankings.** In the discussion at the end of Section 5.1 (just preceding Section 5.2), we mentioned that an l_1-approximate sparse cyclic ranking for R^* may be formulated as the following l_1-minimization problem,

$$\begin{array}{ll} \min & \|X - R^*\|_1 \\ \text{s.t.} & X = \operatorname{curl}^* \Phi, \\ & X = -X^\top. \end{array} \tag{31}$$

This is equivalent to

$$\min_{\Phi \in C^2} \|\operatorname{curl}^* \Phi - R^*\|_1 := \sum\nolimits_{\{i,j\} \in E} |(\operatorname{curl}^* \Phi)_{ij} - R^*_{ij}|,$$

which is in turn equivalent to

$$\min_{\Phi \in C^2} \|\operatorname{curl}^* \Phi - \operatorname{proj}_{\ker(\Delta_1)} \bar{Y}\|_1,$$

where $\operatorname{proj}_{\ker \Delta_1} \bar{Y}$ is the harmonic component in R^*. The chief motivation for this minimization problem has been explained at the end of Section 5.1 — we would like to identify the edges of conflicting pairs in a pairwise comparison graph so that we may have the option of removing them to get a globally consistent ranking.

While both (28) and (31) are l_1-norm minimizations over some pairwise ranking flow. The obvious difference between them lies in that the former searches over $\operatorname{im}(\operatorname{grad})$, the space of gradient flows, i.e. where $X = \operatorname{grad} s$, while the latter searches over $\operatorname{im}(\operatorname{curl}^*)$, the space of curl flows, i.e. where $X = \operatorname{curl}^* \Phi$. The number of free parameters in $\operatorname{grad} s$ is just $|V| = n$ but on the other hand the number of free parameters in $\operatorname{curl}^* \Phi$ is $|T(E)|$, which is typically of the order $O(n^3)$. Therefore we expect to be able to get a residual for (31) that is much sparser than the residual for (28) simply because we are searching over a much larger space. As an illustration, Figure 3 shows the results of these two optimization problems on the same data.

The next theorem shows that the dual problem of (31) also maximizes correlation with the given pairwise ranking flow R^* but over bounded curl-free flows instead of bounded divergence-free flows as in (30).

Theorem 6.2. *Let the inner product be as defined in* (15)*, i.e.*

$$\langle X, Y \rangle_w := \sum\nolimits_{\{i,j\} \in E} w_{ij} X_{ij} Y_{ij}.$$

The dual problem of the l_1-minimization (31) *is*

$$\begin{array}{ll} \max & \langle X, R^* \rangle_w \\ \text{s.t.} & |X_{ij}| \le w_{ij}^{-1}, \\ & \operatorname{curl} X = 0, \\ & X = -X^\top. \end{array} \tag{32}$$

Theorem 6.2. The proof is similar to Theorem 6.1 where grad is replaced by curl^*. Since we have the unweighted l_1-norm in this case, we get the upper bound $|X_{ij}| \le w_{ij}^{-1}$. □ □ □

As we can see, curl in Theorem 6.2 plays the role of div in Theorem 6.1 in the dual problem and curl^* in Theorem 6.2 plays the role of grad in Theorem 6.1 in the primal problem. There is a slight difference on the upper bounds for $|X_{ij}|$, due to the fact that (28) uses a weighted l_1-norm while (31) uses an unweighted l_1-norm. In both theorems, the primal and dual search spaces are orthogonal complements of

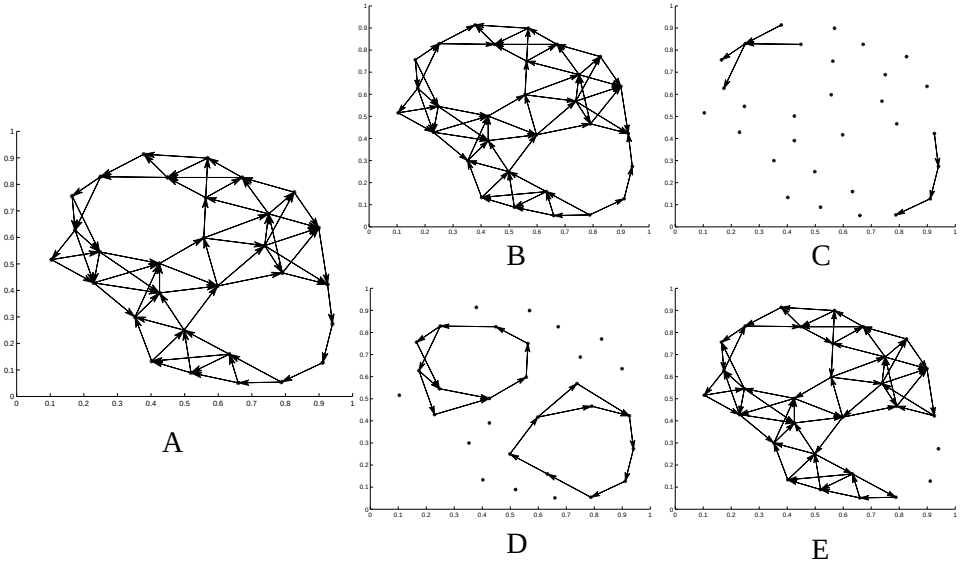

FIGURE 3. Comparisons of the two l_1-optimizations, (28) and (31), with the same harmonic ranking. For simplicity we set weights $w_{ij} = 1$. The arrows in the picture indicate the edge flow direction of pairwise rankings. A. a harmonic ranking flow h; B. the l_1-projection on the gradient flows by (28) (i.e. $\operatorname{grad} s_0$ where $s_0 = \operatorname{argmin}_s \|\operatorname{grad} s - h\|_1$); C. the l_1-projection residual in (28) (i.e. $h - \operatorname{grad} s_0$); D. the approximate sparse cycles by (31) (i.e. $h - \operatorname{curl}^\top \Phi_0$ where $\Phi_0 = \operatorname{argmin}_\Phi \|\operatorname{curl}^\top \Phi - h\|_1$); E. the l_1-projection on locally cyclic flows by (31) (i.e. $\operatorname{curl}^\top \Phi_0$).

each other as given by the Helmholtz decomposition theorem. These two problems thus exhibit a kind of "structural duality".

7. Connections to Social Choice Theory

Social choice theory is almost undoubtedly the discipline most closely associated with the study of ranking, having a long history dating back to Condorcet's famous treatise in 1785 [8] and a large body of work that led to at least two Nobel prizes [3, 12, 35].

The famous impossibility theorems of Arrow [2] and Sen [34] in social choice theory formalized the inherent difficulty of achieving a global ranking of alternatives by aggregating over the voters. However it is still possible to perform an approximate rank aggregation in reasonable, systematic manners. Among the various proposed methods, the best known ones are those by Condorcet [8], Borda [5], and Kemeny [21]. In particular, the Kemeny approach is often regarded as the best approximate rank aggregation method under some assumptions [40, 39]. It is however NP-hard to compute and the assumptions may be unnatural in the context of modern ranking problems.

We have described earlier how the minimization of (8) over the gradient flow model class

$$\mathcal{M}_G = \{X \in C^1 \mid X = \operatorname{grad} s, s : V \to \mathbb{R}\}$$

leads to a Hodge theoretic generalization of Borda count but the minimization of (8) over the Kemeny model class

$$\mathcal{M}_K = \{X \in C^1 \mid X_{ij} = \operatorname{sign}(s_j - s_i), s : V \to \mathbb{R}\}$$

leads to Kemeny optimization. In this section, we will discuss the connection in greater detail.

The following are some desirable properties of ranking data that have been widely studied, used, and assumed in social choice theory. A ranking problem is called *complete* if each voter in Λ gives a total ordering or permutation of all alternatives in V; this implies that $w_{ij}^{\alpha} > 0$ for all $\alpha \in \Lambda$ and all distinct $i, j \in V$, in the terminology of Section 2. It is *balanced* if the pairwise comparison graph $G = (V, E)$ is k-regular with equal weights $w_{ij} = c$ for all $\{i, j\} \in E$. A complete and balanced ranking induces a complete graph with equal weights on all edges. Moreover, it is *binary* if every pairwise comparison is allowed only two values, say, ± 1 without loss of generality. So $Y_{ij}^{\alpha} = 1$ if voter α prefers alternative j to alternative i, and $Y_{ij}^{\alpha} = -1$ otherwise. Ties are disallowed to keep the discussion simple.

Classical social choice theory often assumes complete, balanced, and binary rankings. However, these are all unrealistic assumptions for modern data coming from internet and e-commerce applications. Take the Netflix dataset for illustration, a typical user α of Netflix would have rated at most a very small fraction of the entire Netflix inventory. Indeed, as we have mentioned in Section 2.2.1, the viewer-movie rating matrix has 99% missing values. Moreover, while blockbuster movies would receive a disproportionately large number of ratings, since just about every viewer has watched them, the more obscure or special interest movies would receive very few ratings. In other words, the Netflix dataset is highly incomplete and highly imbalanced. Therefore its pairwise comparison graph is expected to have a sparse edge structure if we ignore pairs of movies where few comparisons have been made[11] (a consequence of incompleteness) and a vertex degree distribution that is far from constant (a consequence of imbalance). We should qualify the the second statement. Of course, a k-regular, unweighted graph (or constantly weighted graph) has a constant vertex degree distribution — this is exactly the definition of balance. However, it is not so clear what we should consider as a 'vertex degree distribution' when the pairwise ranking edge flow is taken into account (note that an edge flow assigns weights to the edges and the graph in questions effectively becomes a weighted graph).

Lastly, as we have discussed in Section 2.2, most modern ranking datasets including the Netflix one are given in terms of ratings or scores on the alternatives by the voters (e.g. one through five stars). While it is possible to ignore the cardinal nature of the dataset and just use its ordinal information to construct a binary pairwise ranking, we would be losing valuable information — for example, a 5-star versus 1-star comparison is indistinguishable from a 3-star versus 2-star comparison when one only takes the ordinal information into account.

[11]This will not be true if we do not perform such thresholding. As we noted earlier, the Netflix pairwise comparison graph is almost a complete graph missing only 0.22% of its edges although the Netflix dataset has 99% of its values missing.

Therefore, one is ill-advised to apply methods from classical social choice theory to modern ranking data directly. However we will see in the next section that our Hodge theoretic extension of Borda count adapts to these new features in modern datasets, i.e. incomplete, imbalanced, cardinal data, but still restricts to the usual Borda count in social choice theory when for data that is complete, balanced, and ordinal/binary.

The reader may wonder why the impossibility theorems of social choice theory do not invalidate our Hodge theoretic approach. One reason is given in the previous paragraph, namely, we work under different assumptions: our ranking data is incomplete, imbalanced, cardinal, and so these impossibility results do not apply. In particular, these impossibility theorems are about *intransitivity*, i.e. whether one might have $i \succeq j \succeq k \succeq i$, which is an ordinal condition; but our approach deals with *inconsistency*, i.e. whether one might have $X_{ij} + X_{jk} + X_{ki} \neq 0$, which is a cardinal condition. The second and more important reason is that we do not merely seek a global ranking but also a local ranking and a harmonic ranking, with the latter two components accounting for the cyclic inconsistencies in the ranking data. We acknowledge at the outset that not all datasets can be reasonably assigned a global ranking but can sometimes be cyclic in nature. So we instead seek to analyze ranking data by examining its three constituting components: global, local, harmonic. The magnitude of the cyclic (local + harmonic) component then quantifies the inconsistencies that impede a global ranking. We do not always regard the cyclic component, which measures the cardinal equivalent of the impossibilities in social choice theory, as noise. In our framework, the data may be 'explained' by a global ranking only when the cyclic component is small; if that is not case, then the cyclic component is an integral part of the ranking data and one has no reason to think that the global component would be any more informative than the cyclic component.

7.1. **Kemeny Optimization and Borda Count.** The basic idea of Kemeny's rule [21, 22] is to minimize the number of pairwise mismatches from a given ordering of the alternatives to a voting profile, i.e. the collection of total orders on the alternatives by each voter. The minimizers are called the *Kemeny optima* and are often regarded as the most reasonable candidates for a global ranking of the alternatives. To be precise, we define the binary pairwise ranking associated with a permutation $\sigma \in \mathfrak{S}_n$, the permutation group on n elements, to be $Y_{ij}^{\sigma} = \operatorname{sign}(\sigma(i) - \sigma(j))$. Given two total orders or permutations on the alternatives $V = \{1, \dots, n\}$, $\sigma, \tau \in \mathfrak{S}_n$, the *Kemeny distance* (also known as *Kemeny-Snell* or *Kendall τ distance*) is defined to be

$$d_K(\sigma, \tau) := \frac{1}{2} \sum\nolimits_{i<j} |Y_{ij}^{\sigma} - Y_{ij}^{\tau}| = \frac{1}{4} \sum\nolimits_{i,j} |Y_{ij}^{\sigma} - Y_{ij}^{\tau}|,$$

i.e. the number of pairwise mismatches between σ and τ. Given a voting profile as a set of permutations on V by m voters, $\{\tau_i \in \mathfrak{S}_n \mid i = 1, \dots, m\}$, the following combinatorial minimization problem

$$\min_{\sigma \in \mathfrak{S}_n} \sum\nolimits_{i=1}^{m} d_K(\sigma, \tau_i) \tag{33}$$

is called *Kemeny optimization* and is known to be NP-hard [14] with respect to n when $m \geq 4$. For binary-valued rankings with $Y_{ij}^{\alpha} \in \{\pm 1\}$, the optimization

problem

$$\min_{X\in\mathcal{M}_K} \sum\nolimits_{\alpha,i,j} w_{ij}^{\alpha}(X_{ij}-Y_{ij}^{\alpha})^2, \tag{34}$$

counts up to a constant the number of pairwise mismatches from a total order. Hence for a complete, balanced, and binary-valued ranking problem, our minimization problem (8) becomes Kemeny optimization if we replace the subspace $\mathcal{M}_G$ by the discrete subset $\mathcal{M}_K$.

Another well-known method for rank aggregation is the *Borda count* [5], which assigns a voter's top ith alternative a *position-based score* of $n-i$; then the global ranking on V is then derived from the sum of its scores over all voters. This is equivalent to saying that the global ranking of the ith alternative is derived from the score

$$s_B(i) = -\sum\nolimits_{\alpha,j=1}^{m,n} Y_{ij}^{\alpha}, \tag{35}$$

i.e. the alternative that has the most pairwise comparisons in favor of it from all voters will be ranked first, and so on. As we have found in (26), the minimum norm solution of the l_2-projection onto gradient flows is given by

$$s^*(i) = -\frac{1}{n}\sum\nolimits_k \bar{Y}_{ik} = -c\sum\nolimits_{\alpha,k} Y_{ik}^{\alpha},$$

where c is an positive constant. Hence for a complete, balanced, and binary ranking problem, the Hodge theoretic approach yields the Borda count up to a positive multiplicative constant, which has no effect on the ordering of alternatives by scores.

7.2. **Comparative Studies.** The following theorem gives three equivalent characterizations of (34) when $Y_{ij}^{\alpha}\in\{\pm 1\}$. Note that here we do not assume that the data is complete and balanced.

Theorem 7.1. *Suppose that $Y_{ij}^{\alpha}\in\{\pm 1\}$. The following optimization problems are all equivalent:*

(i) The weighted least squares problem,

$$\min_{X\in\mathcal{M}_K} \sum\nolimits_{\alpha,i,j} w_{ij}^{\alpha}(X_{ij}-Y_{ij}^{\alpha})^2,$$

where

$$\mathcal{M}_K = \{X\in\mathcal{A} \mid X_{ij} = \operatorname{sign}(s_j - s_i), s: V\to\mathbb{R}\}$$

(ii) The linear programming problem,

$$\max_{X\in\mathcal{K}_1}\langle X,\bar{Y}\rangle = \max_{X\in\mathcal{K}_1}\sum\nolimits_{\{i,j\}\in E} w_{ij}X_{ij}\bar{Y}_{ij} \tag{36}$$

where $\mathcal{K}_1$ is the set

$$\Big\{\sum\nolimits_{\sigma\in S_n}\mu_\sigma P^\sigma \;\Big|\; \sum\nolimits_\sigma \mu_\sigma = 1,\quad \mu_\sigma\geq 0,\quad P_{ij}^{\sigma} = \operatorname{sign}(\sigma(j)-\sigma(i))\Big\}.$$

(iii) The weighted l_1-minimization problem,

$$\min_{X\in\mathcal{K}_2}\|X-\bar{Y}\|_{1,w} = \min_{X\in\mathcal{K}_2}\sum\nolimits_{\{i,j\}\in E} w_{ij}|X_{ij}-\bar{Y}_{ij}| \tag{37}$$

where $\mathcal{K}_2$ is the set

$$\{X\in\mathcal{A} \mid (s_j-s_i)X_{ij}\geq 0 \text{ for some } s: V\to\mathbb{R} \text{ and } \{i,j\}\in E\}.$$

(iv) The minimum feedback arc set of the weighted directed graph $G_{W\circ\bar{Y}} = (V, \overrightarrow{E}, W\circ\bar{Y})$, whose vertex set is V, directed edge $(i,j) \in \overrightarrow{E} \subseteq V \times V$ iff $\bar{Y}_{ij} > 0$ with weight $w_{ij}\bar{Y}_{ij}$.

The set $\mathcal{K}_1$ is the convex hull of skew-symmetric permutation matrices P^σ as defined in [40]. The set $\mathcal{K}_2$ is called the *transitive pairwise region* by Saari [30], which comprises $n!$ cones corresponding to each of the $n!$ permutations on V.

Theorem 7.1. Assuming (i). Since $X_{ij} \in \{\pm 1\}$, we obtain

$$\begin{aligned}\sum\nolimits_{\alpha,i,j} w_{ij}^\alpha (X_{ij} - Y_{ij}^\alpha)^2 &= \sum\nolimits_{\alpha,i,j} w_{ij}^\alpha \left[X_{ij}^2 - 2X_{ij}Y_{ij}^\alpha + (Y_{ij}^\alpha)^2\right] \\ &= c - 2\sum\nolimits_{i,j} X_{ij} \sum\nolimits_\alpha w_{ij}^\alpha Y_{ij}^\alpha \\ &= c - 2\sum\nolimits_{i,j} w_{ij} X_{ij} \bar{Y}_{ij}\end{aligned}$$

where c is a constant that does not depend on X. So the problem becomes

$$\max_{X\in\mathcal{M}_K} \sum\nolimits_{\{i,j\}\in E} w_{ij} X_{ij} \bar{Y}_{ij}. \tag{38}$$

Since $\mathcal{M}_K$ is a discrete set containing $n!$ points, a linear programming problem over $\mathcal{M}_K$ is equivalent to searching over its convex hull, i.e. $\mathcal{K}_1$, which gives (ii).

(iv) can also be derived from (38). Consider a weighted directed graph $G_{W\circ\bar{Y}}$ where an edge $(i,j) \in \overrightarrow{E}$ iff $\bar{Y}_{ij} > 0$, and in which case has weight $|w_{ij}\bar{Y}_{ij}|$. (38) is equivalent to finding a directed acyclic graph by reverting a set of edge directions whose weight sum is minimized. This is exactly the minimum feedback arc set problem.

Finally, we show that (iii) is also equivalent to the minimum feedback arc set problem. For any $X \in \mathcal{K}_2$, the transitive region, there is an associated weighted directed acyclic graph $G_{W\circ X}$ where an edge $(i,j) \in \overrightarrow{E}$ iff $X_{ij} > 0$, and in which case has weight $|w_{ij}X_{ij}|$. Note that an optimizer of (37) has either $X_{ij}^* = -X_{ji}^* = \bar{Y}_{ij}$ or $X_{ij}^* = -X_{ji}^* = 0$ on an edge $\{i,j\} \in E$, which is equivalent to the problem of finding a directed acyclic graph by deleting a set of edges from $G_{W\circ\bar{Y}}$ such that the sum of their weights is minimized. Again, this is exactly the minimum feedback arc set problem. □ □ □

It is known that the minimum feedback arc set problem in (iv) is NP-hard, and therefore, so are the other three. Moreover, (iii) provides us with some geometric insights when we view it alongside with (8), the l_2-projection onto gradient flows $\mathcal{M}_G = \{X \in \mathcal{A} \mid X_{ij} = s_j - s_i, s : V \to \mathbb{R}\}$ — which we have seen to be a Hodge theoretic extension of Borda count. We will illustrate their differences and similarities pictorially via the following example borrowed from Saari [30].

Consider the simplest case of three-item comparison with $V = \{i,j,k\}$. For simplicity, we will assume that $w_{ij} = w_{jk} = w_{ki} = 1$ and $\bar{Y}_{ij}, \bar{Y}_{jk}, \bar{Y}_{ki} \in [-1,1]$. Figure 4 shows the unit cube in $\mathbb{R}^3$. We will label the coordinates in $\mathbb{R}^3$ as $[X_{ij}, X_{jk}, X_{ki}]$ (instead of the usual $[x,y,z]$). The shaded plane corresponds to the set where $X_{ij} + X_{jk} + X_{ki} = 0$ in the unit cube. Note that this set is equal to the model class $\mathcal{M}_G$ because of (13). On the other hand, the transitive pairwise region $\mathcal{K}_2$ consists of the six orthants within the cube with vertices $\{\pm 1, \pm, 1, \pm 1\} - \{[1,1,1], [-1,-1,-1]\}$. We will write

$$I(X) := \sum\nolimits_{\alpha,i,j} w_{ij}^\alpha (X_{ij} - Y_{ij}^\alpha)^2.$$

The Hodge theoretic optimization (8) is the l_2-projection onto the plane $X_{ij}+X_{jk}+X_{ki}=0$, while by (iii), the Kemeny optimization (34) is the l_1-projection onto the aforementioned six orthants representing the transitive pairwise region $\mathcal{K}_2$.

In the general setting of social choice theory, the following theorem from [30] characterizes the order relations between the Kemeny optimization and the Borda count.

Theorem 7.2 (Saari-Merlin 2000)**.** *The Kemeny winner (the most preferred) is always strictly above the Kemeny loser (the least preferred) under the Borda count; similarly the Borda winner is always strictly above the Borda loser under the Kemeny rule. There is no other constraint in the sense that the two methods may generate arbitrary different total orders except for those constraints.*

The Kemeny rule has several desirable properties in social choice theory which the Borda count lacks [40]. The Kemeny rule satisfies the Condorcet rule, in the sense that if one candidate in V wins all pairwise comparisons against other candidates in V, then it must be the overall winner. A Condorcet winner is any alternative i such that $\sum_j \operatorname{sign}(\sum_\alpha Y_{ij}^\alpha) = n$. Note that the Condorcet winner may not exist in general but Kemeny or Borda winners always exist. However, if a Condorcet winner exists, then it must be the Kemeny winner. On the other hand, Borda count can only ensure that the Condorcet winner is ranked strictly above the Condorcet loser (least-preferred). Another major advantage of the Kemeny rule is its consistency in global rankings under the elimination of candidates in V. The Borda count and many other position-based rules fail to meet this condition. In fact, the Kemeny rule is the unique rule that meets all three of following: (1) satisfies the Condorcet rule, (2) consistency under elimination, and (3) a natural property called neutral (that we will not discuss here). See [40] for further details.

Despite the many important features that the Kemeny rule has, its high computational cost (NP-hard) makes simpler rules like Borda count attractive in practice, especially when there is large number of alternatives to be ranked. Moreover, in cardinal rankings where it is desirable to preserve the magnitude of score differences [10] and not just the order relation, using the Hodge theoretic variant of Borda count with the model class $\mathcal{M}_G$ becomes more relevant than Kemeny optimization with model class $\mathcal{M}_K$.

8. Summary and Conclusion

We have introduced combinatorial Hodge theory into the analysis of some rank learning techniques based on minimizing pairwise ranking errors over a model space. The incompleteness and imbalance of modern datasets make such an approach attractive. Such datasets often have a sparse structure arising from the incompleteness of the data (and/or from thresholding) that can be captured by the Hodge theoretic approach towards rank learning discussed in this paper. We have seen that combinatorial Hodge theory reveals the close relationship between the graph structure and ranking structure. In particular, it provides a decomposition of an edge flow representing a pairwise ranking into three orthogonal components, a gradient flow, a locally cyclic flow, and a locally acyclic flow, representing a global ranking, a triangular cyclic ranking, and a harmonic ranking respectively. Consistency of the ranking data is governed to a large extent by the structure of its pairwise comparison graph; this is in turn revealed by the Hodge decomposition

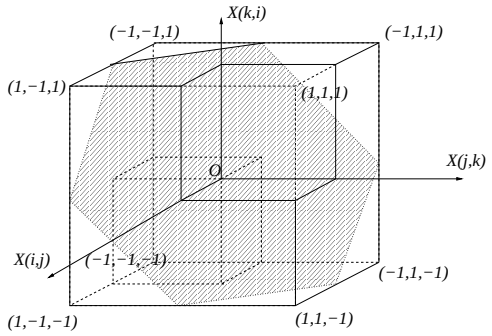

FIGURE 4. The shaded region is the subspace $X_{ij}+X_{jk}+X_{ki} = 0$. The transitive region consists of six orthants whose corresponding vertices belong to $\{\pm 1, \pm, 1, \pm 1\} - \{[1, 1, 1], [-1, -1, -1]\}$. The Borda count or $\min_{X \in \mathcal{M}_G} I(X)$ is the l_2-projection onto the shaded plane while the Kemeny optimization or $\min_{X \in \mathcal{M}_K} I(X)$ is the l_1-projection onto the transitive region.

associated with the combinatorial Laplacians of the clique complex of the graph. The sparsity structure of a pairwise comparison graph imposes certain constraints on the topology and geometry and its clique complex, which in turn decides the properties of our rank learning algorithms.

We propose a Hodge theoretic approach towards learning the global ranking component of a dataset. This is done via a l_2-projection of a pairwise ranking onto the space of gradient flows, which yields a globally consistent ranking. We saw that among other connections to classical social choice theory, the score recovered from this globally consistent ranking is a generalization of the well-known Borda count to the ranking data that is cardinal, imbalanced, and incomplete. The residual left is the l_2-projection onto the space of divergence-free flows. This sheds light on the nature of the inconsistencies and in addition one may use an l_1-approximate sparse cyclic rankings to identify the edges where conflicts among voters occur. The l_1-minimization problem for this is shown to have a dual as correlation maximization over the bounded curl-free flows. On the other hand, the l_1-projection on the gradient flows, which we view as a robust variant of the l_2-version, has as dual problem a correlation maximization over bounded cyclic flows.

Our results suggest that combinatorial Hodge theory could be a promising tool for the supervised learning of ranking, especially for modern datasets with cardinal, incomplete, and imbalanced information.

REFERENCES

[1] N. Ailon, M. Charikar, and A. Newman, "Aggregating inconsistent information: ranking and clustering," *Proc. ACM Symposium Theory Comput.* (STOC '05), **37** (2005), pp. 684–693.

[2] K.J. Arrow, "A difficulty in the concept of social welfare," *J. Polit. Econ.*, **58** (1950), no. 4, pp. 328–346.

[3] K. Arrow, "General economic equilibrium: purpose, analytic techniques, collective choice," Nobel Memorial Lecture, December 12, 1972, pp. 109–131 in: Assar Lindbeck (Ed.), *Nobel Lectures: Economic Sciences 1969–1980*, World Scientific, Singapore, 1992.

[4] B. M. Barber, R. Lehavy, M McNichols, and B. Trueman, "Can investors profit from the prophets? security analyst recommendations and stock returns," *J. Finance*, **56** (2001), no. 2, pp. 531-563.
[5] J.-C. de Borda, "Mémoire sur les élections au scrutin," *Histoire de'Académie Royale des Sciences*, **102** (1781), pp. 657–665.
[6] R.M. Bell and Y. Koren, "Scalable collaborative filtering with jointly derived neighborhood interpolation weights," *Proc. IEEE Internat. Conf. Data Mining* (ICDM '07), **7** (2007), pp. 43-52.
[7] R. Bradley and M. Terry, "The rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, **39** (1952), pp. 324–345.
[8] A.-N. de Condorcet, *Éssai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, Imprimerie Royale, Paris, 1785.
[9] F. Chung, *Spectral graph theory*, CBMS Regional Conference Series in Mathematics, **92**, AMS, Providence, RI, 1997.
[10] C. Cortes, M. Mohri, and A. Rastogi, "Magnitude-preserving ranking algorithms," *Proc. Internat. Conf. Mach. Learn.* (ICML '07), **24** (2007), pp. 169–176.
[11] H.A. David, *The method of paired comparisons*, 2nd Ed., Griffin's Statistical Monographs and Courses, **41**, Oxford University Press, New York, NY, 1988.
[12] G. Debreu, "Stochastic choice and cardinal utility," *Econometrica*, **26** (1958), no. 3, pp. 440–444.
[13] P. Diaconis, "A generalization of spectral analysis with application to ranked data," *Ann. Statist.*, **17** (1989), no. 3, pp. 949–979.
[14] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," *Proc. Internat. Conf. World Wide Web* (WWW '01), **10** (2001), pp. 613–622.
[15] Y. Freund, R. Iyer, R. Shapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *J. Mach. Learn. Res.*, **4** (2004), no. 6, pp. 933–969.
[16] J. Friedman, "Computing Betti numbers via combinatorial Laplacians," *Algorithmica*, **21** (1998), no. 4, pp. 331–346.
[17] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *Ann. Statis.*, **26** (1998), no. 2, pp. 451–471.
[18] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," pp. 115–132, *Advances in large margin classifiers*. MIT Press, Cambridge, MA, 2000.
[19] A. Tahbaz-Salehi and A. Jadbabaie, "Distributed coverage verification in sensor networks without location information," *Proc. IEEE Conf. Decis. Control*, **47** (2008), to appear.
[20] M. Kahle, "Topology of random clique complexes," *Discrete Math.*, (2008), to appear.
[21] J.G. Kemeny, "Mathematics without numbers," *Daedalus*, **88** (1959), pp. 571-591.
[22] J.G. Kemeny and L.J. Snell, "Preference ranking: an axiomatic approach," pp. 9–23 in J.G. Kemeny and L.J. Snell (Eds.), *Mathematical models in the social sciences*, MIT Press, Cambridge, MA, 1973.
[23] M. Kendall and J.D. Gibbons, *Rank correlation methods*, 5th Ed., Oxford University Press, Oxford, 1990.
[24] M.G. Kendall and B.B. Smith, "On the method of paired comparisons," *Biometrika*, **31** (1940), no. 3–4, pp. 324–345.
[25] M. Ma, "A matrix approach to asset pricing in foreign exchange market," (2006), *preprint* (`http://ssrn.com/abstract=921755`).
[26] F. Mosteller, "Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations," *Psychometrika*, **16** (1951), no. 1, pp. 3–9.
[27] F. Mosteller, "Remarks on the method of paired comparisons: II. The effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed," *Psychometrika*, **16** (1951), no. 2, pp. 203–206.
[28] F. Mosteller, "Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed," *Psychometrika*, **16** (1951), no. 2, pp. 207–218.
[29] G.E. Noether, "Remarks about a paired comparison model," *Psychometrika*, **25** (1960), pp. 357–367.

[30] D.G. Saari and V.R. Merlin, "A geometric examination of Kemeny's rule," *Soc. Choice Welf.*, **17** (2000), no. 3, pp. 403–438.
[31] T.L. Saaty, "A scaling method for priorities in hierarchical structures," *J. Math. Psych.*, **15** (1977), no. 3, pp. 234–281.
[32] T.L. Saaty, "Inconsistency and rank preservation," *J. Math. Psych.*, **28** (1984), no. 2, pp. 205–214.
[33] T.L. Saaty and M.S. Ozdemir, "Why the magic number seven plus or minus two," *Math. Comput. Modelling*, **38** (2003), no. 3–4, pp. 233–244.
[34] A. Sen, "The impossibility of a Paretian liberal," *J. Polit. Econ.*, **78** (1970), no. 1, pp. 152–157.
[35] A.K. Sen, "The possibility of social choice," Nobel Lecture, December 8, 1998, pp. 178–215 in: Torsten Persson (Ed.), *Nobel Lectures: Economic Sciences 1996–2000*, World Scientific, Singapore, 2003.
[36] S. Smale and N. Smale, "Hodge decomposition and learning theory," *preprint*, (2008).
[37] S.C. Narula and J.F. Wellington, "The minimum sum of absolute errors regression: a state of the art survey," *Internat. Statist. Rev.* **50** (1982), no. 3: pp. 317–326.
[38] L.L. Thurstone, "The method of paired comparisons for social values," *J. Abnorm. Soc. Psychol.*, **21** (1927), pp. 384–400.
[39] H.P. Young, "Condorcet's theory of voting," *Am. Polit. Sci. Rev.*, **82** (1988), pp. 1231–1244.
[40] H.P. Young and A. Levenglick, "A consistent extension of Condorcet's election principle," *SIAM J. Appl. Math.*, **35** (1978), no. 2, pp. 285–300.

Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, 94305
E-mail address: xiaoyej@stanford.edu

Department of Mathematics, University of California, Berkeley, CA 94720
E-mail address: lekheng@berkeley.edu

Department of Mathematics, Stanford University, Stanford, CA 94305
E-mail address: yuany@stanford.edu

Department of Management Science and Engineering, Stanford University, Stanford, CA, 94305
E-mail address: yinyu-ye@stanford.edu