Statistical ranking problems and Hodge decompositions of graphs and skew-symmetric matrices

Lek-Heng Lim and Yuan Yao

2008 Workshop on Algorithms for Modern Massive Data Sets

June 28, 2008

(Contains joint work with Qi-xing Huang and Xiaoye Jiang)

Hodge decomposition)

- **Vector calculus:** Helmholtz's decomposition, ie. vector fields on *nice domains* may be resolved into irrotational (curl-free) and solenoidal (divergence-free) component vector fields

$$\mathbf{F} = -\nabla\varphi + \nabla \times \mathbf{A}$$

  φ scalar potential, $\mathbf{A}$ vector potential.

- **Linear algebra:** additive orthogonal decomposition of a skew-symmetric matrix into three skew-symmetric matrices

$$W = W_1 + W_2 + W_3$$

  $W_1 = \mathbf{v}\mathbf{e}^T - \mathbf{e}\mathbf{v}^T$, W_2 clique-consistent, W_3 inconsistent.

- **Graph theory:** orthogonal decomposition of network flows into acyclic and cyclic components.

Ranking on networks (graphs)

- Multicriteria rank/decision systems
  - Amazon or Netflix's recommendation system (user-product)
  - Interest ranking in social networks (person-interest)
  - S&P index (time-price)
  - Voting (voter-candidate)
- Peer review systems
  - publication citation systems (paper-paper)
  - Google's webpage ranking (web-web)
  - eBay's reputation system (customer-customer)

Characteristics

Aforementioned ranking data often

- incomplete: typically about 1% (cf. earlier talks by Agarwal, Banerjee)
- imbalanced: power-law, heavy-tail distributed votes (cf. earlier talks by Fatlousos, Mahoney, Mihail)
- cardinal: given in terms of scores or stochastic choices

Implicitly or explicitly, ranking data may be viewed to live on a simple graph $G = (V, E)$, where

- V: set of alternatives (products, interests, etc) to be ranked
- E: pairs of alternatives to be compared

Example: Netflix customer-product rating

Example (Netflix Customer-Product Rating)

- 480189-by-17770 customer-product rating matrix X
- X is **incomplete**: 98.82% of values missing

However,

- pairwise comparison graph $G = (V, E)$ is very **dense**!
- only 0.22% edges are missed, **almost a complete graph**
- rank aggregation may be carried out without estimating missing values
- **imbalanced**: number of raters on $e \in E$ varies

Caveat: we are not trying to solve the Netflix prize problem

Netflix example continued

The **first order** statistics, mean score for each product, is often inadequate because of the following:

- most customers would rate just a **very small portion** of the products
- different products might have different raters, whence mean scores involve noise due to **arbitrary individual rating scales**
- customers give **ratings** instead or **orderings**

How about **high order** statistics?

From 1st order to 2nd order: pairwise rankings

- *Linear Model*: average score difference between product i and j over all customers who have rated both of them,

$$w_{ij} = \frac{\sum_k (X_{kj} - X_{ki})}{\#\{k : X_{ki}, X_{kj} \text{ exist}\}}.$$

Invariant up to translation.

- *Log-linear Model*: when all the scores are positive, the logarithmic average score ratio,

$$w_{ij} = \frac{\sum_k (\log X_{kj} - \log X_{ki})}{\#\{k : X_{ki}, X_{kj} \text{ exist}\}}.$$

Invariant up to a multiplicative constant.

More invariants

- *Linear Probability Model*: the **probability** that product j is preferred to i in excess of a purely random choice,

$$w_{ij} = \Pr\{k : X_{kj} > X_{ki}\} - \frac{1}{2}.$$

  Invariant up to monotone transformation.

- *Bradley-Terry Model*: logarithmic odd ratio (logit)

$$w_{ij} = \log \frac{\Pr\{k : X_{kj} > X_{ki}\}}{\Pr\{k : X_{kj} < X_{ki}\}}.$$

  Invariant up to monotone transformation.

Skew-symmetric matrices of pairwise rankings

Recall **skew-symmetric** matrices: $W \in \mathbb{R}^{n \times n}$, $W^T = -W$:

- every $A \in \mathbb{R}^{n \times n}$ decomposable into $A = S + W$, $S = (A + A^T)/2$ symmetric, $W = (A - A^T)/2$ skew-symmetric
- $\mathcal{W} = \{\text{skew-symmetric matrices}\} = \wedge^2(\mathbb{R}) = \mathfrak{o}_n(\mathbb{R})$

All previous models induce (sparse) skew-symmetric matrices of size $|V|$-by-$|V|$

$$w_{ij} = \begin{cases} -w_{ji} & \text{if } \{i, j\} \in E \\ ? & \text{otherwise} \end{cases}$$

where $G = (V, E)$ is a pairwise comparison graph.

Note: such a skew-symmetric matrix induces a pairwise ranking **flow** on graph G.

Pairwise ranking graph for IMDb top 20 movies

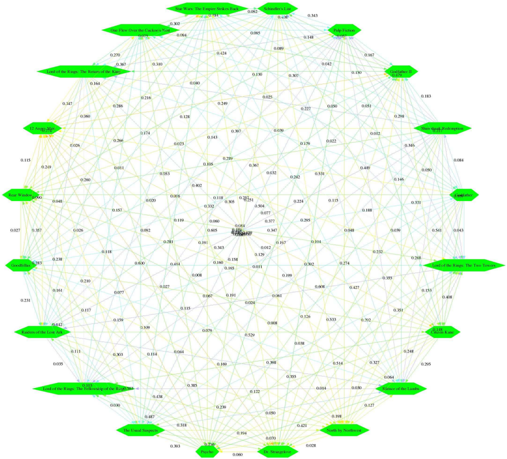

Figure: Pairwise ranking flow of Netflix data restricted to top IMDb movies

Rank aggregation problem

Difficulties:

- Arrow's impossibility theorem
- Kemeny-Snell optimal ordering is NP-hard to compute
- Harmonic analysis on $\mathfrak{S}_n$ is impractical for large n since $|\mathfrak{S}_n| = n!$

Our approach:

Problem

Does there exist a global ranking function, $v : V \to \mathbb{R}$, such that

$$w_{ij} = v_j - v_i =: \delta_0(v)(i,j)?$$

Equivalently, does there exists a scalar field $v : V \to \mathbb{R}$ whose gradient field gives the flow w? ie. is w **integrable**?

Answer: not always!

Multivariate calculus: there are non-integrable vector fields; cf. the film *A Beautiful Mind*:

$$A = \{F : \mathbb{R}^3 \setminus X \to \mathbb{R}^3 \mid F \text{ smooth}\}, \quad B = \{F = \nabla g\},$$
$$\dim(A/B) = ?$$

Similarly here,

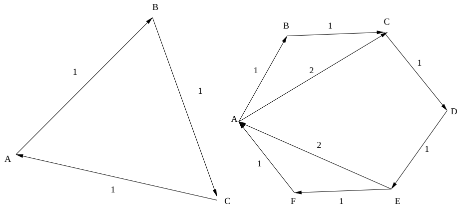

Figure: No global ranking v gives $w_{ij} = v_j - v_i$: (a) triangular cyclic, note $w_{AB} + w_{BC} + w_{CA} \neq 0$; (b) it contains a 4-node cyclic flow $A \to C \to D \to E \to A$, note on a 3-clique $\{A, B, C\}$ (also $\{A, E, F\}$), $w_{AB} + w_{BC} + w_{CA} = 0$

Triangular transitivity

Fact

$W = [w_{ij}]$ skew symmetric associated with graph $G = (V, E)$. If $w_{ij} = v_j - v_i$ for all $\{i, j\} \in E$, then $w_{ij} + w_{jk} + w_{ki} = 0$ for all 3-cliques $\{i, j, k\}$.

Transitivity subspace:

$$\{W \text{ skew symmetric} \mid w_{ij} + w_{jk} + w_{ki} = 0 \text{ for all 3-cliques}\}$$

Example in the last slide, (a) lies outside; (b) lies in this subspace, but not a gradient flow.

Hodge theory: matrix theoretic

A skew-symmetric matrix W associated with G can be decomposed uniquely

$$W = W_1 + W_2 + W_3$$

where

- W_1 satisfies
  - '*integrable*': $W_1(i,j) = v_j - v_i$ for some $v : V \to \mathbb{R}$.
- W_2 satisfies
  - '*curl free*': $W_2(i,j) + W_2(j,k) + W_2(k,i) = 0$ for all (i,j,k) 3-clique;
  - '*divergence free*': $\sum_{j:(i,j)\in E} W_2(i,j) = 0$
- $W_3 \perp W_1$ and $W_3 \perp W_2$.

Hodge theory: graph theoretic

Orthogonal decomposition of network flows on G into

$$\text{gradient flow} + \text{globally cyclic} + \text{locally cyclic}$$

where the first two components make up transitive component and

- gradient flow is integrable to give a global ranking
- example (b) is locally (triangularly) acyclic, but cyclic on large scale
- example (a) is locally (triangularly) cyclic

Clique complex of a graph

Extend graph G to a **simplicial complex** $\mathcal{K}(G)$ by attaching triangles

- 0-simplices $\mathcal{K}_0(G)$: V
- 1-simplices $\mathcal{K}_1(G)$: E
- 2-simplices $\mathcal{K}_2(G)$: triangles $\{i, j, k\}$ such that every edge is in E
- k-simplices $\mathcal{K}_k(G)$: $(k+1)$-cliques $\{i_0, \ldots, i_k\}$ of G

For ranking problems, suffices to construct $\mathcal{K}(G)$ up to dimension **2**!

- **global** ranking $v : V \to \mathbb{R}$, 0-**forms**, ie. vectors
- **pairwise** ranking $w(i,j) = -w(j,i)$ for $(i,j) \in E$, 1-**forms**, ie. skew-symmetric matrices

Clique complex

- k-**forms**:

$$C^k(\mathcal{K}(G), \mathbb{R}) = \{u : \mathcal{K}_{k+1}(G) \to \mathbb{R}, u_{i_{\sigma(0)},\ldots,i_{\sigma(k)}} = \operatorname{sign}(\sigma) u_{i_0,\ldots,i_k}\}$$

  for $(i_0, \ldots, i_k) \in \mathcal{K}_{k+1}(G)$, where $\sigma \in \mathfrak{S}_{k+1}$ is a permutation on $(0, \ldots, k)$.
- May put metrics/inner products on $C^k(\mathcal{K}(G), \mathbb{R})$.
- The following metric on 1-forms, is useful for the imbalance issue

$$\langle w_{ij}, \omega_{ij} \rangle_D = \sum_{(i,j)\in E} D_{ij} w_{ij} \omega_{ij}$$

  where

$$D_{ij} = |\{\text{customers who rate both } i \text{ and } j\}|.$$

Discrete exterior derivatives (coboundary maps)

- k-coboundary maps $\delta_k : C^k(\mathcal{K}(G), \mathbb{R}) \to C^{k+1}(\mathcal{K}(G), \mathbb{R})$ are defined as the **alternating difference** operator

$$(\delta_k u)(i_0, \ldots, i_{k+1}) = \sum_{j=0}^{k+1} (-1)^{j+1} u(i_0, \ldots, i_{j-1}, i_{j+1}, \ldots, i_{k+1})$$

- δ_k plays the role of **differentiation**
- $\delta_{k+1} \circ \delta_k = 0$
- In particular,
  - $(\delta_0 v)(i,j) = v_j - v_i =: (\operatorname{grad} v)(i,j)$
  - $(\delta_1 w)(i,j,k) = (\pm)(w_{ij} + w_{jk} + w_{ki}) =: (\operatorname{curl} w)(i,j,k)$ (triangular-trace of skew-symmetric matrix $[w_{ij}]$)

Div, Grad, Curl

For each triangle $\{i, j, k\}$, the **curl**

$$(\text{curl}\, w)(i, j, k) = (\delta_1 w)(i, j, k) = w_{ij} + w_{jk} + w_{ki}$$

measures the total flow-sum along the loop $i \to j \to k \to i$.

- $(\delta_1 w)(i, j, k) = 0$ implies the flow is **path-independent**, which defines the **triangular transitivity subspace**.

For each alternative $i \in V$, the **divergence**

$$(\text{div}\, w)(i) := -(\delta_0^T w)(i) := \sum w_{i*}$$

measures the **inflow-outflow sum** at i.

- $(\delta_0^T w)(i) = 0$ implies alternative i is preference-neutral in all pairwise comparisons.
- divergence-free flow $\delta_0^T w = 0$ is **cyclic**

Boundary of a boundary is empty

Fundamental tenet of topology: $\delta_{k+1} \circ \delta_k = 0$.
For $k = 0$,

$$C^0 \xrightarrow{\delta_0} C^1 \xrightarrow{\delta_1} C^2,$$

ie.

$$\text{Global} \xrightarrow{\text{grad}} \text{Pairwise} \xrightarrow{\text{curl}} \text{Triplewise}$$

and so

$$\text{Global} \xleftarrow{\text{grad}^*(=:-\,\text{div})} \text{Pairwise} \xleftarrow{\text{curl}^*} \text{Triplewise}.$$

We have

$$\text{curl} \circ \text{grad}(\text{Global Rankings}) = 0.$$

This implies

- global rankings are transitive/consistent,
- no need to consider rankings beyond triplewise.

Combinatorial Laplacians

- k-dimensional **combinatorial Laplacian**, $\Delta_k : C^k \to C^k$ by

$$\Delta_k = \delta_{k-1}\delta_{k-1}^* + \delta_k^*\delta_k, \qquad k > 0$$

- $k = 0$, **graph Laplacian** or **vertex Laplacian**

$$\Delta_0 = \delta_0^*\delta_0$$

- $k = 1$, **vector Laplcian** (first term is edge Laplacian)

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

- Important Properties:
  - Δ_k positive semidefinite
  - $\ker(\Delta_k) = \ker(\delta_{k-1}^*) \cap \ker(\delta_k)$ **harmonic forms**
  - Hodge decomposition

Hodge decomposition for combinatorial Laplacians

- Every combinatorial Laplacians Δ_k has an associated Hodge decomposition.
- For $k = 1$, this is the decomposition (of discrete vector fields/skew symmetric matrices/network flows) that we have been discussing.

Theorem (Hodge decomposition for pairwise ranking)

The space of pairwise rankings, $C^1(\mathcal{K}(G), \mathbb{R})$, admits an orthogonal decomposition into three

$$C^1(\mathcal{K}(G), \mathbb{R}) = \operatorname{im}(\delta_0) \oplus H_1 \oplus \operatorname{im}(\delta_1^*)$$

where

$$H_1 = \ker(\delta_1) \cap \ker(\delta_0^*) = \ker(\Delta_1).$$

Illustration

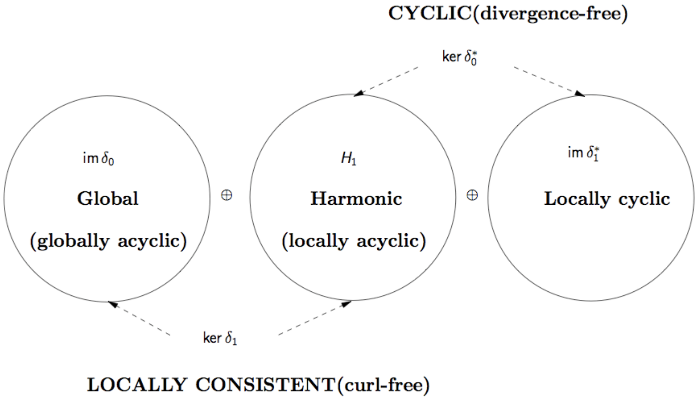

Figure: Hodge decomposition for pairwise rankings

Harmonic rankings: locally consistent but globally inconsistent

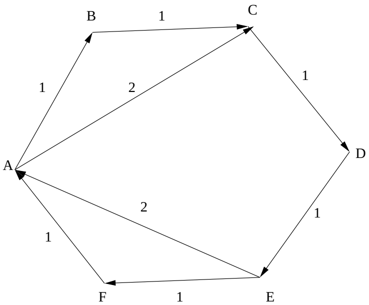

Figure: A locally consistent but globally cyclic harmonic ranking.

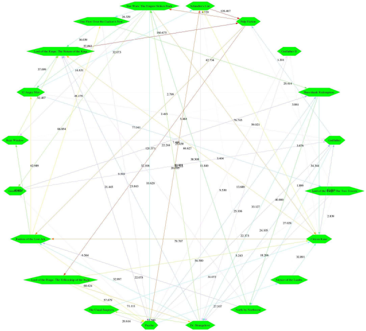

Figure: A harmonic ranking from truncated Netflix movie-movie network

Rank aggregation as projection

Rank aggregation problem reduced essentially to linear least squares

Corollary

Every pairwise ranking admits a unique orthogonal decomposition,

$$w = \operatorname{proj}_{\operatorname{im}(\delta_0)} w + \operatorname{proj}_{\ker(\delta_0^*)} w$$

i.e.

$$\textit{pairwise} = \operatorname{grad}(\textit{global}) + \textit{cyclic}$$

Particularly the first projection grad(*global*) *gives a global ranking*

$$x^* = (\delta_0^* \delta_0)^\dagger \delta_0^* w = -(\Delta_0)^\dagger \operatorname{div}(w)$$

$O(n^3)$ flops complexity with great algorithms (dense: Givens/Householder QR, Golub-Reinsch SVD; sparse: CGLS, LSQR; sampling: DMM '06)

Erdős-Rényi random graph

Heuristical justification from Erdős-Rényi random graphs (cf. earlier talks by Chung, Mihail, Saberi)

Theorem (Kahle '07)

For an Erdős-Rényi random graph $G(n, p)$ with n vertices and edges forming independently with probability p, its clique complex χ_G will have zero 1-homology almost always, except when

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

Since the full Netflix movie-movie comparison graph is almost complete (0.22% missing edges), one may expect the chance of nontrivial harmonic ranking is small.

Which pairwise ranking model might be better?

Use curl distribution:

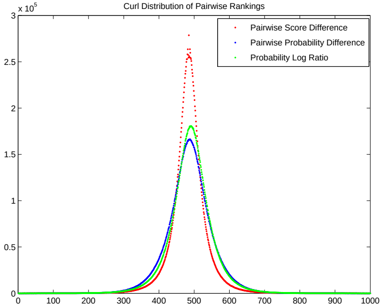

Figure: Curl distribution of three pairwise rankings, based on most popular 500 movies. The pairwise score difference in red have the thinnest tail.

Comparisons of Netflix global rankings

	Mean Score	Score Difference	Probability Difference	Logarithmic Odd Ratio
Mean Score	1.0000	0.9758	0.9731	0.9746
Score Difference		1.0000	0.9976	0.9977
Probability Difference			1.0000	0.9992
Logarithmic Odd Ratio				1.0000
Cyclic Residue	-	6.03%	7.16%	7.15%

Table: Kendall's Rank Correlation Coefficients between different global rankings for Netflix. Note that the pairwise score difference has the smallest relative residue.

Why pairwise ranking works for Netflix?

- Pairwise rankings are good approximations of **gradient flows** on movie-movie networks
- In fact, Netflix data in the large scale behaves like a 1-dimensional curve in high dimensional space
- To visualize this, we use a spectral embedding approach

Spectral embedding

Technique proposed by Goel, Diaconis, and Holmes (cf. earlier talk by Jordan).

- Map every movie to a point in S^5 by

$$\text{movie } m \rightarrow (\sqrt{p_1(m)}, \ldots, \sqrt{p_5(m)})$$

  where $p_k(m)$ is the probability that movie m is rated as star $k \in \{1, \ldots, 5\}$. Obtain a movie-by-star matrix Y.
- Do SVD on Y, which is equivalent to do eigenvalue decomposition on the linear kernel

$$K(s,t) = \langle s, t \rangle^d, \qquad d = 1$$

- $K(s,t)$ is nonnegative, whence the first eigenvector captures the **centricity** (density) of data and the second captures a **tangent field** of the manifold.

SVD embedding

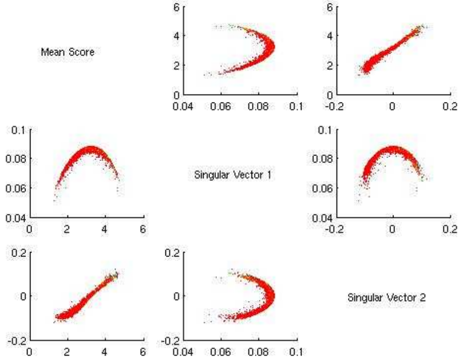

Figure: The second singular vector is monotonic to the mean score, indicating the intrinsic parameter of the horseshoe curve is driven by the mean score

Conclusions

- Ranking as **1-dimensional scaling** of data
- Pairwise ranking as approximate gradient fields or **flows on graphs**
- Hodge Theory provides an **orthogonal decomposition** for pairwise ranking flows,
- This decomposition helps characterize the *local (triangular) vs. global consistency* of pairwise rankings, and gives a natural *rank aggregation* scheme