Ranking from pairwise comparisons: a near-linear time minimax optimal algorithm for learning BTL weights

Julien Hendrickx (UCLouvain)

We consider the problem of ranking and learning the qualities $w_1, \ldots, w_n$ of a collection of items by performing noisy comparisons among them. We assume that there is a fixed "comparison graph", and every neighboring pair of items is compared k times.

We focus more specifically on the popular Bradley-Terry-Luce model, where comparisons are i.i.d. events, and the probability for item i to win the comparison against j is $w_i/(w_i + w_j)$.

We propose a near-linear time algorithm allowing us to recover the weights with an accuracy that outperforms all existing algorithms, and show that this accuracy is actually within a constant factor of information-theoretic lower bounds, that we also develop. This accuracy is related to the average resistance of the comparison graph.

Our algorithm is based on a weighted least square, with weights determined from empirical outcomes of the comparisons.

We further discuss the extension to other models of comparisons, and comparisons involving multiple items.

Ranking from pairwise comparisons: a near-linear time minimax optimal algorithm for learning BTL weights

Julien Hendrickx – Lille – 10 March 2023

What if Ligue 1 has to stop now?

Who is champion? *What is the ranking?*

→ who goes to L2, to European league etc.

<u>Possible solution:</u> use current standing

			pts	J.	G.	N.	P.	p.	c.	+/-	G. N. P.
1	Paris-SG		**63**	26	20	3	3	66	25	+41	G P G G G
2	Marseille		**55**	26	17	4	5	49	25	+24	P G G P G
3	Monaco		**51**	26	15	6	5	55	36	+19	G G G P N
4	Lens		**51**	26	14	9	3	40	21	+19	N P G N N
5	Rennes		**46**	26	14	4	8	45	29	+16	P P G G P
6	Lille		**45**	26	13	6	7	46	33	+13	G G P G N
7	Nice		**42**	26	11	9	6	34	22	+12	G G N G N
8	Reims	2 ▲	**40**	26	9	13	4	34	26	+8	N G N G G
9	Lorient	1 ▼	**40**	26	11	7	8	38	36	+2	N P G P N
10	Lyon	1 ▼	**39**	26	11	6	9	39	28	+11	G G P G N
11	Clermont	1 ▲	**34**	26	9	7	10	26	34	-8	P P P N G
12	Toulouse	1 ▼	**32**	26	9	5	12	41	46	-5	P G P P P

What if Ligue 1 has to stop now?

Who is champion? *What is the ranking?*

→ who goes to L2, to European league etc.

<u>*Possible solution:*</u> use current standing

			pts	J.	G.	N.	P.	p.	c.	+/-	G. N. P.
1	Paris-SG		**63**	26	20	3	3	66	25	+41	G P G G G
2	Marseille		**55**	26	17	4	5	49	25	+24	P G G P G
3	Monaco		**51**	26	15	6	5	55	36	+19	G G G P N
4	Lens		**51**	26	14	9	3	40	21	+19	N P G N N
5	Rennes		**46**	26	14	4	8	45	29	+16	P P G G P
6	Lille		**45**	26	13	6	7	46	33	+13	G G P G N
7	Nice		**42**	26	11	9	6	34	22	+12	G G N G N
8	Reims	2 ▲	**40**	26	9	13	4	34	26	+8	N G N G G
9	Lorient	1 ▼	**40**	26	11	7	8	38	36	+2	N P G P N
10	Lyon	1 ▼	**39**	26	11	6	9	39	28	+11	G G P G N
11	Clermont	1 ▲	**34**	26	9	7	10	26	34	-8	P P P N G
12	Toulouse	1 ▼	**32**	26	9	5	12	41	46	-5	P G P P P

Nice and Reims similar

But 2 weeks ago

STADE DE REIMS 3 0 TOULOUSE FC

good

MONACO 0 3 OGC NICE

Much stronger achievement

- Nice should get more recognition
- "Current standing" option unfair for teams who only played stronger teams

What if Ligue 1 has to stop now?

Who is champion? *What is the ranking?*

→ who goes to L2, to European league etc.

- Nice should get more recognition
- "Current standing" option unfair for teams who only played stronger teams

Inherent problem when *games are not all-to-all*

- Tennis ranking
- Chess
- (…)

→ *How to* ***build ranking / # points*** *from* ***results of "arbitrary" comparisons***

How to evaluate pain-killer efficiency

Asking patients number between 1 and 10 ?

- Good but not very objective + patient dependent
- Can’t test all on all patient
- Preference for giving “good ones”

Practical data collection: try 2 and ask which is best
+ learn quality

Online review

less than 5* often an insult

➔ Not very informative

Alternative: did you prefer this place or this place

Comparison can be all you have

Preference expressed by action

Multiple items, not everyone compares all

How to rank / recover value based on (non-exhaustive) comparisons?

Bradley-Terry-Luce model

- Items have intrinsic quality (weight): w_i
- When comparing i - j , i wins with probability

$$p_{ij} = \frac{w_i}{w_i + w_j}$$

Example

pick coffee with 80% probability, tea with 20%

4 ***1***

XXX football team: ***3*** YYY football team: ***2***

→ XXX should win with probability 60%

Idea: recover weights w_i from the comparison results

Ranking from pairwise comparisons

- Motivation and Problem
- Weighted Least-Square Estimator
- Algorithm and Complexity
- Error Analysis
  - Error Bound
  - Lower Bound – Minimax Optimality
  - Other criteria
- Experimental Results
- A Surprising Observation
- Generalizations
- Conclusions

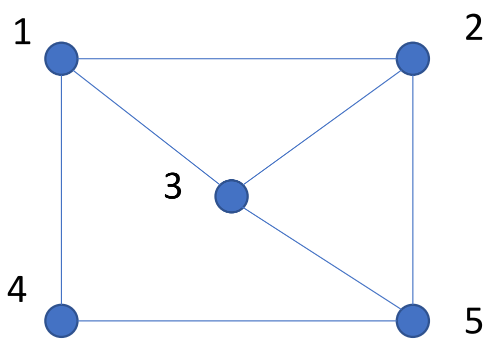
Ranking from pairwise comparisons

- Motivation and Problem
- ***Weighted Least-Square Estimator***
- Algorithm and Complexity
- Error Analysis
  - Error Bound
  - Lower Bound – Minimax Optimality
  - Other criteria
- Experimental Results
- A Surprising Observation
- Generalizations
- Conclusions

Weight recovery

Items $1, \dots, n$ with quality (weights) $w_1, \dots, w_n \in [1, b]$

Comparison graph



k i.i.d. comparisons for each edge

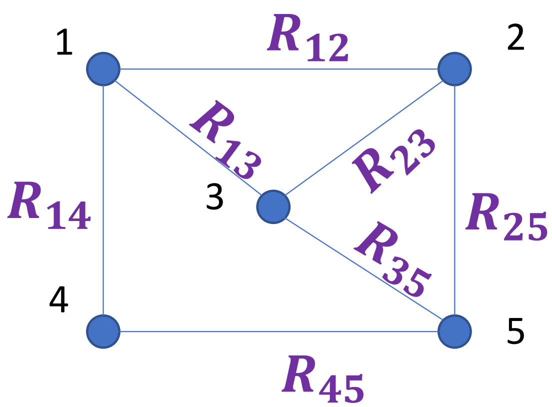
i wins comparison against j with probability

$$p_{ij} = \frac{w_i}{w_i + w_j}$$

Problem: Recover vectors of weights $w = (w_1, \dots, w_n)'$ from results, up to constant multiplicative factor. Range b *exists but is not known*

Sufficient statistics: k and ratio of wins $R_{ij} = \frac{\#\text{ wins i}}{\#\text{ wins j}}$

Data has network structure



Sufficient statistics: k and ratio of wins

$$R_{ij} = \frac{\text{\# wins i}}{\text{\# wins j}}$$

Goal = recover values at nodes

Previous solutions

- Maximum Likelihood
  - Convex optimization problem after reformulation
  - Asymptotically optimal, but only asymptotic guarantees

- Rank centrality [Negahban, Oh, Shah 2016]
  - Based on convergence of Markov Chain built from data

$$\frac{\left|\left|\frac{w}{||w||_1} - \hat{W}\right|\right|_2^2}{\left|\left|\frac{w}{||w||_1}\right|\right|_2^2} \leq O\left(\frac{1}{k}\right) \frac{b^5 \log n}{(1-\rho)^2} \frac{d_{\max}}{d_{\min}^2},$$

$1-\rho$ spectral gap of random walk
d_{max}, d_{min} largest, smallest degree
b maximal weight

Could scale as $n^7 b^5 / k$

Several improvements

Algorithm idea: Least-Square

Probability i wins over j: $\frac{w_i}{w_i+w_j}$

For large number k of comparisons i - j :

win i $\simeq kp_{ij} = k\frac{w_i}{w_i+w_j}$
win j $\simeq kp_{ji} = k\frac{w_j}{w_i+w_j}$

$\Rightarrow$ $R_{ij} = \frac{\text{\# win i}}{\text{\# win j}} \simeq \frac{w_i}{w_j}$

$\Leftrightarrow$ $\log w_i - \log w_j \simeq \log R_{ij}$

(Naïve) Idea 1: Least-square solution of

$$\log \hat{w}_i - \log \hat{w}_j = \log R_{ij} \qquad \forall (i,j) \in E$$

Issue 1: zero wins

Lease square solution of

$$\log\,\widehat{w}_i - \log \widehat{w}_j = \log R_{ij} \qquad \forall (i,j) \in E$$

$$R_{ij} = \frac{\#\text{ wins i}}{\#\text{ wins j}}$$

What if i wins no comparison ? (or all)

$$R_{ij} = 0 \Rightarrow \log R_{ij} = -\infty$$

→ Complete Failure, with positive probability

Solution: Replace 0 victory by ½ victory

- Simple
- provides boundedness properties
- But creates technical complications

Issue 2: Non-uniform Variance

Lease square solution of

$$\log \hat{w}_i - \log \hat{w}_j = \log R_{ij} \qquad \forall (i,j) \in E$$

		5 vs 5	9 vs 1	
Variance # win i	$\frac{k}{v_{ij}}$	$\frac{k}{4}$	$\frac{k}{11.11}$	
"Variance" $\log R_{ij}$	$\simeq \frac{v_{ij}}{k}$	$\frac{4}{k}$	$\frac{11.11}{k}$	$\simeq$ 3×larger

With $v_{ij} := \frac{w_i}{w_j} + 2 + \frac{w_j}{w_i}$

Error in equation (9,1) expected to be larger than for (5,5)

→ ***Corresponding equations should be treated differently.***

Solution: Weighted least square

Least square solution of $\dfrac{\log \hat{w}_i - \log \hat{w}_j}{\sqrt{v_{ij}}} = \dfrac{\log R_{ij}}{\sqrt{v_{ij}}}$

Idea: each equation should have "the same variance" $v_{ij} := \dfrac{w_i}{w_j} + 2 + \dfrac{w_j}{w_i}$

(inspired by Best Linear Unbiased Estimator idea)

→ ***Ideal Estimator***

$$\log \hat{w} = \arg\min_{z} \sum_{(i,j)\in E} \frac{(z_i - z_j - \log R_{ij})^2}{v_{ij}}$$

Weighted least square

→ *Ideal Estimator*

$$\log \widehat{w} = \arg\min_{z} \sum_{(i,j)\in E} \frac{(z_i - z_j - \log R_{ij})^2}{v_{ij}}$$

Issue 3: $v_{ij} := \frac{w_i}{w_j} + 2 + \frac{w_j}{w_i}$ Depends on the values we want to recover

Iterative solution:

Initiate $\hat{v}_{ij} = 4$ for all edges
Repeat
  Compute estimate $\widehat{w}$ with $\hat{v}_{ij}$
  update $\hat{v}_{ij}$ based on $\widehat{w}$

Empirical solution:

$$R_{ij} \simeq \frac{w_i}{w_j} \quad \rightarrow \quad v_{ij} := \frac{w_i}{w_j} + 2 + \frac{w_j}{w_i} \simeq R_{ij} + 2 + R_{ij}^{-1}$$

Weighted least square

→ *Ideal Estimator*

$$\log \hat{w} = \arg\min_{z} \sum_{(i,j)\in E} \frac{(z_i - z_j - \log R_{ij})^2}{v_{ij}}$$

Issue 3: $v_{ij} := \frac{w_i}{w_j} + 2 + \frac{w_j}{w_i}$ Depends on the values we want to recover

Iterative solution:

In

Re

- Computationally cheaper
- Simpler to analyze
- More accurate (surprisingly)

Empirical solution:

$$R_{ij} \simeq \frac{w_i}{w_j} \quad \rightarrow \quad v_{ij} := \frac{w_i}{w_j} + 2 + \frac{w_j}{w_i} \simeq R_{ij} + 2 + R_{ij}^{-1}$$

Final Estimator

Would be 0 with "nominal" ratio w_i / w_j and real weights

$$\log \hat{w} = \arg\min_{z} \sum_{(i,j)\in E} \frac{\left(z_i - z_j - \log R_{ij}\right)^2}{\hat{v}_{ij}}$$

With $\hat{v}_{ij} := R_{ij} + 2 + R_{ij}^{-1}$ Empirical "variance"

$R_{ij} =$ # wins i / # wins j

- $\hat{w}$ computed by solving linear least-square problem
- But nonlinear dependence on data and R_{ij}
- No hyper parameter, tuning etc. (can be introduced)
- Can be computed in near linear time

Accuracy ϵ in $O\left(|E| \log^c n \log \frac{1}{\epsilon}\right)$

Ranking from pairwise comparisons

- Motivation and Problem
- Weighted Least-Square Estimator
- ***Algorithm and Complexity***
- Error Analysis
    - Error Bound
    - Lower Bound – Minimax Optimality
    - Other criteria
- Experimental Results
- A Surprising Observation
- Generalizations
- Conclusions

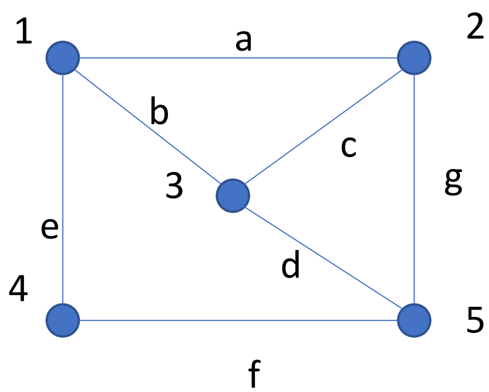
Reminder Incidence matrix B

Relates **nodes to edges**

Column: edge
Row: nodes

If edge e from i to j $\begin{cases} B_{ie} = -1 \\ B_{je} = 1 \end{cases}$

Orientation arbitrary



	a	b	c	…			
1	1	1			1		
2	-1		1				1
3		-1	-1	1			
4				-1	-1	1	
5						-1	-1

Compact reformulation with B

Relates **nodes to edges**

Column: edge
Row: nodes

If edge e from i to j $\begin{cases} B_{ie} = -1 \\ B_{je} = 1 \end{cases}$

Orientation arbitrary

→ System

$$z_i - z_j = \log R_{ij} \qquad \text{for all } (i, j) \in E$$

Can be rewritten compactly

$$B^T z = \log R$$

- One equation / edge
- One variable / node

With $R \in \mathbb{R}^{|E|}$ vector of R_{ij}

Compact reformulation with B

Relates **nodes to edges**

Column: edge
Row: nodes

If edge e from i to j $\begin{cases} B_{ie} = -1 \\ B_{je} = 1 \end{cases}$

Orientation arbitrary

→ System

$$\frac{z_i - z_j}{\sqrt{v_{ij}}} = \frac{\log R_{ij}}{\sqrt{v_{ij}}} \qquad \text{for all } (i, j) \in E$$

Can be rewritten compactly

$$V^{-1/2} B^T z = V^{-1/2} \log R$$

With $R \in \mathbb{R}^{|E|}$ vector of R_{ij}

$V = diag\ (\ldots, v_{ij}, \ldots)$

v_{ij} approximated from data

Least-Square

Estimator: $\log \widehat{w}$ least square solution of

$$V^{-1/2}B^T z = \qquad\qquad V^{-1/2}\log R$$

Normal equations → solution of

$$(V^{-\frac{1}{2}}B^T)^T V^{-1/2}B^T z = (V^{-\frac{1}{2}}B^T)^T V^{-1/2}\log R$$

Least-Square

Estimator: $\log \hat{w}$ least square solution of

$$V^{-1/2} B^T z = \qquad V^{-1/2} \log R$$

Normal equations → solution of

$$(V^{-\frac{1}{2}} B^T)^T V^{-1/2} B^T z = (V^{-\frac{1}{2}} B^T)^T V^{-1/2} \log R$$

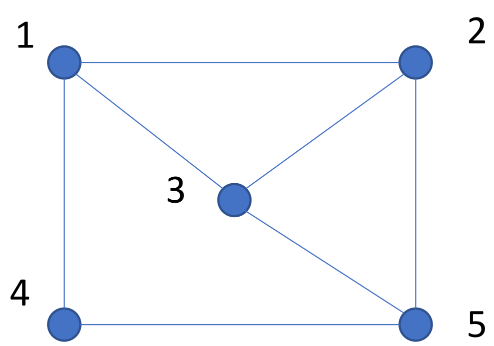
$$B V^{-1} B^T z = B V^{-1} \log R$$

(weighted) Laplacian matrix

Reminder: Laplacian Matrix

Represents
- relations between nodes
- degrees

$$L_{ij} = -1 \text{ if edge } (i, j)$$
$$L_{ii} = degree(i)$$



	1	2	3	4	5
1	3	-1	-1	-1	
2	-1	3	-1		-1
3	-1	-1	3		-1
4	-1			2	-1
5		-1	-1	-1	3

Reminder: Laplacian Matrix

Represents
- relations between nodes
- degrees

$$L_{ij} = -1 \text{ if edge } (i,j)$$
$$L_{ii} = degree(i)$$

Interesting properties

- $L = BB^T$
- $L1 = 0$ (sum line = 0)
- Positive semi-definite
- $\lambda_2 > 0$ if graph connected
  + "algebraic connectivity"

	1	2	3	4	5
1	3	-1	-1	-1	
2	-1	3	-1		-1
3	-1	-1	3		-1
4	-1			2	-1
5		-1	-1	-1	3

Reminder: Weighted Laplacian Matrix

Weights $A_{ij} = A_{ji}$ on edges

$L_{ij} = -A_{ij}$ if edge (i, j)

$$L_{ii} = strength(i) = \sum_{j \neq i} A_{ij}$$

Represents
- Weights of relations between nodes
- Degrees/strengths of nodes

Interesting properties

- $L = B\,diag\left(A_{ij}\right)B^T$
- $L1 = 0$ (sum line = 0)
- Positive semi-definite
- $\lambda_2 > 0$ if graph connected
  + "algebraic connectivity"

$diag\left(A_{ij}\right) \in \mathbb{R}^{|E|\times|E|}$

Final algorithm: Laplacian System

$$BV^{-1}B^T z = BV^{-1}\log R$$

=: L_V ***(weighted) Laplacian matrix***

log $\widehat{w}$ = solutions of $L_V z = BV^{-1}\log R$

$R \in \mathbb{R}^{|E|}$ vector of R_{ij} $\frac{\text{\# wins i}}{\text{\# wins j}}$

$V = diag\ (\dots, v_{ij}, \dots)$
"variance" empirically estimated

Laplacian L_V is ***symmetric*** and ***diagonally*** dominant ($L_{V,ii} = -\sum_{j \neq i} L_{V,ij}$)

[Spielman, Teng 2014], system solved up to accuracy ϵ in $O\left(|E| \log^c n \log\frac{1}{\epsilon}\right)$

→***Near linear time in size*** $|E|$ ***of data.***

For reasonable size systems, easier to use classical solver

Ranking from pairwise comparisons

- Motivation and Problem
- Weighted Least-Square Estimator
- Algorithm and Complexity
- ***Error Analysis***
  - ***Error Bound***
  - Lower Bound – Minimax Optimality
  - Other criteria
- Experimental Results
- A Surprising Observation
- Generalizations
- Conclusions

Error analysis

Disclaimer: Intuitive heuristic analysis

Formal proofs
- Exist
- Were guided by this analysis
- Involve many technical difficulties
- Probably not for a presentation.

In particular we assume
- $E \log R_{ij} = \log \rho_{ij}$ $\rho_{ij} := \frac{w_i}{w_j}$
- Variance $\log R_{ij} = \frac{v_{ij}}{k}$
- Exact v_{ij} used in the algorithm

(all this is “asympotically” true)

Error analysis

$\log \widehat{w}$ = solutions of $\quad \mathrm{L}_V z = BV^{-1} \log R$

How accurate is this estimate? → characterize $\Delta \log w = \log \widehat{w} - \log w$

<u>Scale Problem :</u>

- $w, \widehat{w}$ only defined ***up to multiplicative constant***
- $\log w$, $\log \widehat{w}$ defined up to ***additive constant***

$$p_{ij} = \frac{w_i}{w_i + w_j}$$

→ Arbitrary choice: $\log w$, $\log \widehat{w}$ sum to 0, i.e. orthogonal to **1**

→ $\log \widehat{w} = L_V^{\dagger} BV^{-1} \log R$

With $L_A^{\dagger}$ Monroe Penrose Pseudo-inverse (kernel and image orthogonal to **1**)

$\log w = L_V^{\dagger} BV^{-1} \log \rho$

$\rho_{ij} := \frac{w_i}{w_j}$ true ratio

$$\log \hat{w} = L_V^{\dagger} B V^{-1} \log R$$

$$\log w = L_V^{\dagger} B V^{-1} \log \rho$$

→ $$\Delta \log w = L_V^{\dagger} B V^{-1} \Delta \log R$$

$$
\begin{aligned}
E\, \Delta \log w \, \Delta \log w^T &= E \left(L_V^{\dagger} B V^{-1} \Delta \log R\right)\left(L_V^{\dagger} B V^{-1} \Delta \log R\right)^T \\
&= E L_V^{\dagger} B V^{-1} \Delta \log R \, \Delta \log R^T \, V^{-1} B^T L_V^{\dagger} \\
&= L_V^{\dagger} B V^{-1} \left(\mathrm{E}\Delta \log R \, \Delta \log R^T\right) V^{-1} B^T L_V^{\dagger}
\end{aligned}
$$

$\log \hat{w} = L_V^\dagger B V^{-1} \log R$

$\log w = L_V^\dagger B V^{-1} \log \rho$

→ $\Delta \log w = L_V^\dagger B V^{-1} \Delta \log R$

$$E\, \Delta \log w\, \Delta \log w^T = E\, \left(L_V^\dagger B V^{-1} \Delta \log R\right)\left(L_V^\dagger B V^{-1} \Delta \log R\right)^T$$
$$= E L_V^\dagger B V^{-1} \Delta \log R\, \Delta \log R^T\, V^{-1} B^T L_V^\dagger$$
$$= L_V^\dagger B V^{-1} (\mathrm{E}\Delta \log R\, \Delta \log R^T) V^{-1} B^T L_V^\dagger$$

Square "co-variance" matrix, $|E| \times |E|$

- Diagonal because edges independent and we assume $E\, \Delta\log R_{ij} = 0$
- for edge (i, j) value v_{ij}/k

→ $\mathrm{E}\Delta \log R\, \Delta \log R^T = \frac{1}{k} V$

$$\log \hat{w} = L_V^\dagger B V^{-1} \log R$$

$$\log w = L_V^\dagger B V^{-1} \log \rho$$

$\rightarrow$ $$\Delta \log w = L_V^\dagger B V^{-1} \Delta \log R$$

$$E\, \Delta \log w\, \Delta \log w^T = E\left(L_V^\dagger B V^{-1} \Delta \log R\right)\left(L_V^\dagger B V^{-1} \Delta \log R\right)^T$$

$$= E L_V^\dagger B V^{-1} \Delta \log R\, \Delta \log R^T\, V^{-1} B^T L_V^\dagger$$

$$= L_V^\dagger B V^{-1} \left(E \Delta \log R\, \Delta \log R^T\right) V^{-1} B^T L_V^\dagger$$

$$= \frac{1}{k} L_V^\dagger B V^{-1} V V^{-1} B^T L_V^\dagger$$

$$= \frac{1}{k} L_V^\dagger B V^{-1} B^T L_V^\dagger$$

$$= \frac{1}{k} L_V^\dagger L_V L_V^\dagger = \frac{1}{k} L_V^\dagger$$

by property of Monroe-Penrose inverse

Summary: For a given graph and vector of weight, for large enough k (non-asymptotic)

$$E\ \Delta\ \log w\ \ \Delta\ \log w^T \simeq \frac{1}{k} L_V^\dagger$$

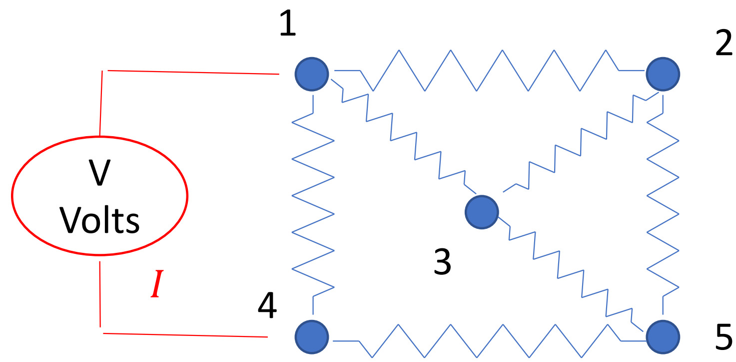
Pseudo-inverse of weighted Laplacian,
Weights = inverse variance v_{ij}^{-1}

Square Error $E\ \| \log \widehat{w} - \log w \|^2 \simeq \frac{1}{k} Tr(L_V^\dagger)$

Reminder: Graph resistance

Weights $A_{ij} = A_{ji}$ represent ***conductance*** of wires

$$\Omega_{14} = V/I$$



Effective Resistance Ω_{ij} = V / current if V volts between i and j

Average resistance: Average over all pairs

$$\Omega_{av} = \frac{1}{n} Tr\ (L_A^{\dagger}) = \frac{1}{n} \sum_{i>1} \frac{1}{\sigma_i(L_A)}$$

With $L_A^{\dagger}$ Monroe Penrose Pseudo-inverse

Alternative measure of connectivity – less centered on "worst-case"

Summary: For a given graph and vector of weight, for large enough k (non-asymptotic)

$$E\ \Delta\ \log w\ \ \Delta\ \log w^T \simeq \frac{1}{k} L_V^\dagger$$

Pseudo-inverse of weighted Laplacian,
Weights = inverse variance v_{ij}^{-1}

Square Error $E\ \| \log \widehat{w} - \log w \|^2 \simeq \frac{1}{k} Tr\left(L_V^\dagger\right) = \frac{n}{k} \Omega_{V,av}$

(→ Mean square error $\frac{1}{k} \Omega_V, av$)

Summary: For a given graph and vector of weight, for large enough k (non-asymptotic)

$$E\ \Delta\ \log w\ \ \Delta\ \log w^T \simeq \frac{1}{k} L_V^\dagger$$

Pseudo-inverse of weighted Laplacian,
Weights = inverse variance v_{ij}^{-1}

Square Error $E\ \|\log \hat{w} - \log w\ \|^2 \simeq \frac{1}{k} Tr\left(L_V^\dagger\right) = \frac{n}{k}\Omega_{V,av}$

$$= O\left(\frac{bn^2}{k}\right) \ = O\left(\frac{bn\Omega_{av}}{k}\right)$$

- Ω_{av} resistance unweighted graph
- b maximal ratio of weights.

- Accuracy determined by ***average resistance***
- $O\left(\frac{bn^2}{k}\right)$ *vs* $O\left(\frac{b^5n^7}{k}\right)$ (But criteria not strictly comparable)

Bound comparison

Graph	Negahban 16	Our result
Line	$b^{5/2}n^2$	$b\sqrt{n}$
Circle	$b^{5/2}n^2$	$b\sqrt{n}$
2D grid	$b^{5/2}n$	b
3D grid	$b^{5/2}n^{2/3}$	b
Star graph	$b^{5/2}\sqrt{n}$	b
2 stars joined at centers	$b^{5/2}n^{1.5}$	b
Barbell graph	$b^{5/2}n^{3.5}$	$b\sqrt{n}$
Geo. random graph	$b^{5/2}n$	b
Erdos-Renyi	$b^{5/2}$	b

Factor 1/k omitted

Ranking from pairwise comparisons

- Motivation and Problem
- Weighted Least-Square Estimator
- Algorithm and Complexity
- Error Analysis
    - Error Bound
    - ***Lower Bound – Minimax Optimality***
    - Other criteria
- Experimental Results
- A Surprising Observation
- Generalizations
- Conclusions

Lower bound

$\frac{1}{k} L_V^\dagger$ = Fisher information matrix,
But, many relevant estimates biased → ***Cramer-Rao not directly applicable***

Nevertheless:

Theorem: For any nominal weights w and any comparison graph,
There is a way of generating w_z randomly in a ball of radius $O_{w,G}\left(\frac{1}{\sqrt{k}}\right)$
(with $\sum_i (w_z)_i = \sum_i w_i$)
such that for any estimator $\hat{w}$ using the outcome Y of k comparisons

$$E \parallel \log \hat{w}(Y) - \log w_z \parallel^2 \geq \Omega\left(\frac{1}{k}\right) Tr\left(L_V^\dagger\right)$$

→ For large enough # comparisons, simple least square algorithm
is minimax optimal (up to constant factor)

Proof technique

1) Generate w_z by combining i.i.d. variations along eigenvectors of L_V

2) Exploit **Lemma 6.1.** *Let μ be any joint probability distribution of a random pair (w, w'), such that the marginal distributions of both w and w' are equal to π. Then*

$$\mathbb{E}_{\pi,\mathbf{Y}}[d(w, \hat{w}(\mathbf{Y}))] \geq \mathbb{E}_{\mu}\left[d(w, w')(1 - \|P_w - P_{w'}\|_{TV}\right]$$

where $||\cdot||_{\mathrm{TV}}$ represents the total-variation distance between distributions and $\mathbf{Y}$ the observations.

(see e.g. *[Hajek & Raginsky, 2019]*)

3) Use Pinsker's inequality $\|P_w^{\otimes k} - P_{w'}^{\otimes k}\|_{\mathrm{TV}}^2 \quad \leq \quad \frac{1}{2} D_{KL}(P_w^{\otimes k} \| P_w'^{\otimes k})$

+ exploit decomposition properties of KL-divergence

Ranking from pairwise comparisons

- Motivation and Problem
- Weighted Least-Square Estimator
- Algorithm and Complexity
- Error Analysis
    - Error Bound
    - Lower Bound – Minimax Optimality
    - ***Other criteria***
- Experimental Results
- A Surprising Observation
- Generalizations
- Conclusions

Other performance criteria?

How about $\mathrm{E}\ \| A\Delta\log w\ \|^2$

Ex: $\Delta\log w_i - \Delta\log w_j$ = error on $(\log w_i - \log w_j)$

$\sim$ relative error on of $\frac{w_i}{w_j}$

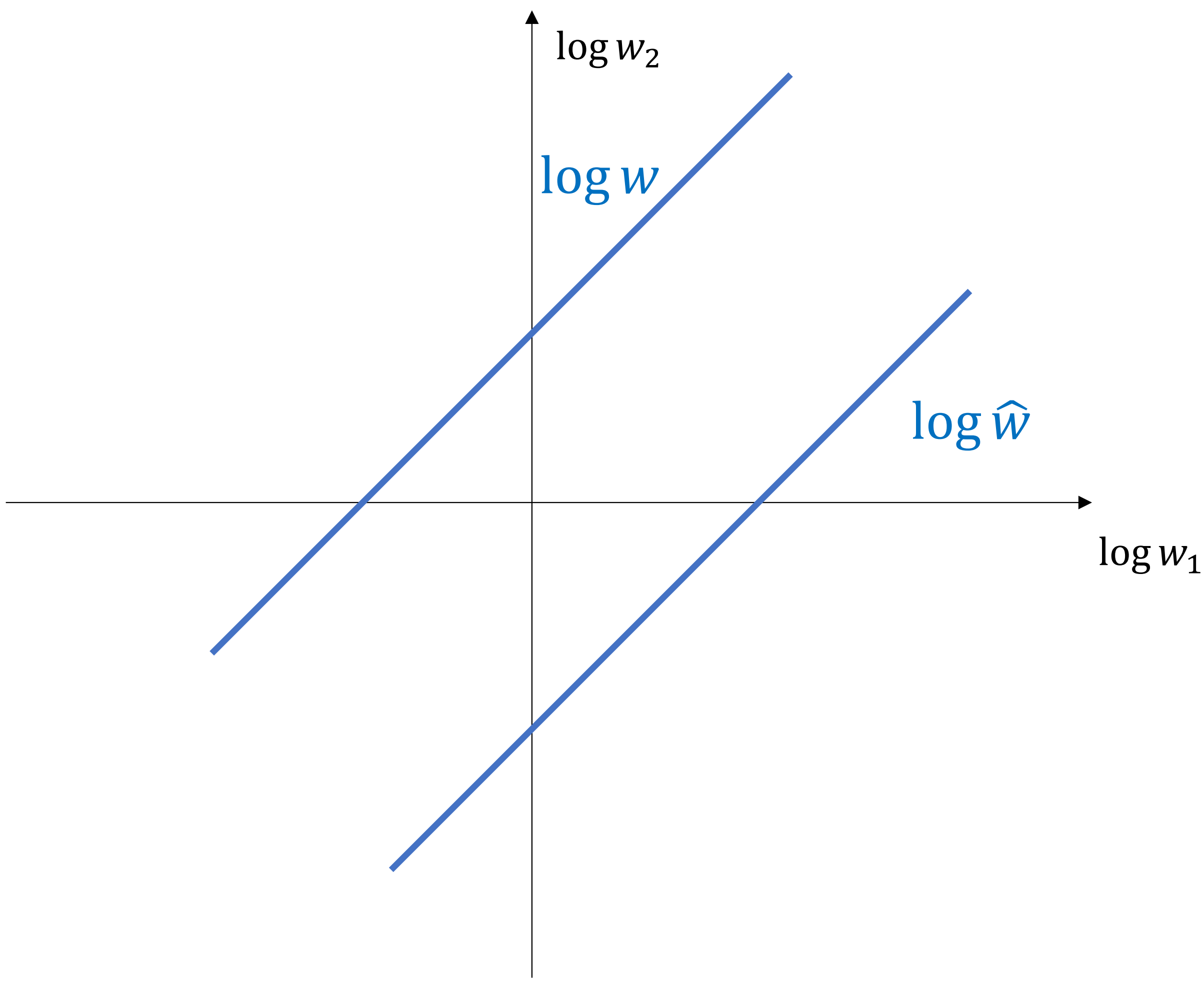
Direct (naïve) approach:

$$E\ \Delta\ \log w\ \Delta\ \log w^T \simeq \frac{1}{k} L_V^\dagger$$

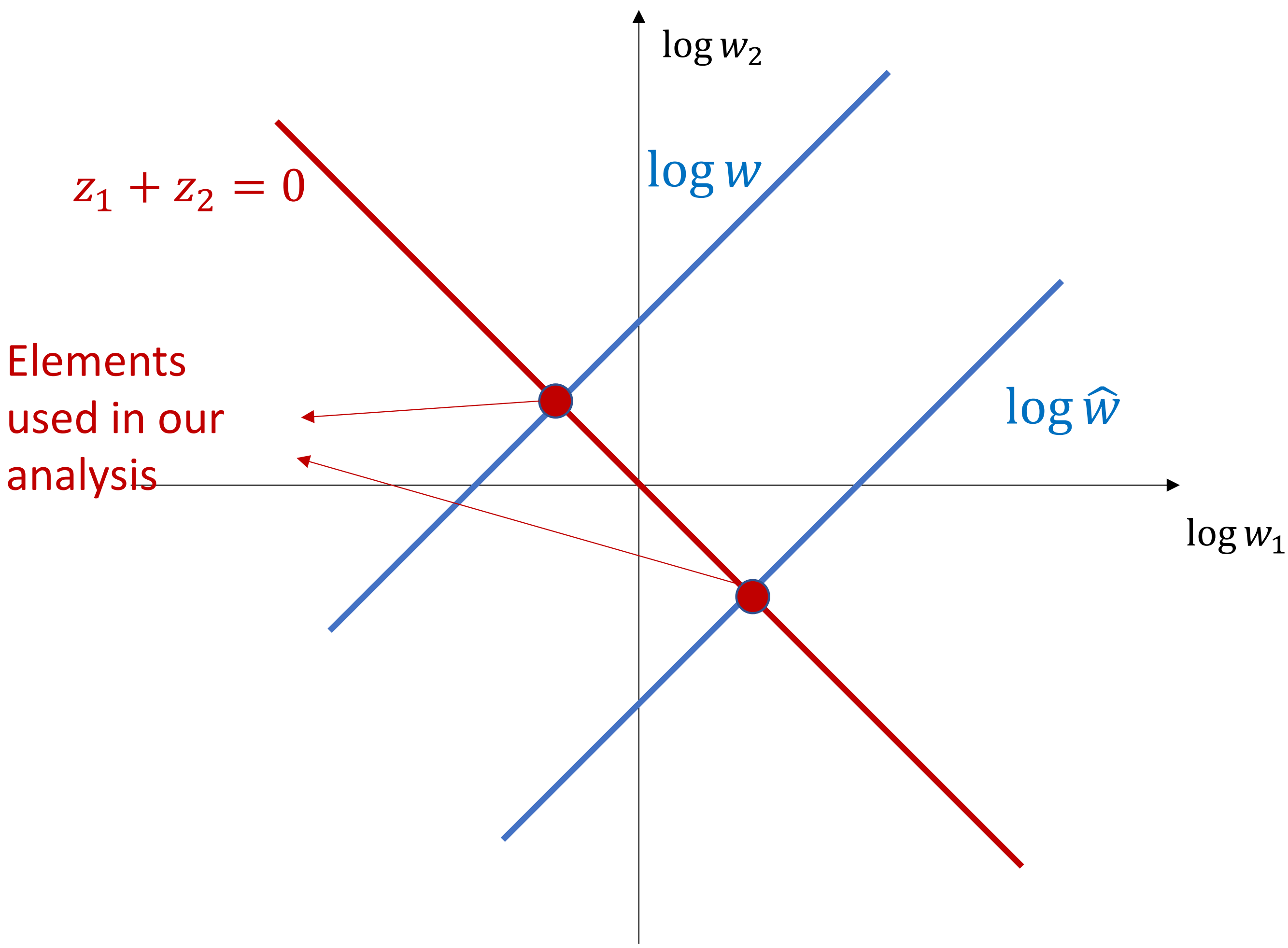
$$\mathrm{E}\ \| A\Delta\log w\ \|^2 = Tr(A\ E\ \Delta\ \log w\ \Delta\ \log w^T\ A^T) \simeq \frac{1}{k} Tr(A L_V^\dagger A^T)$$

Problem: assumption $\sum_i \log w_i = 0$ not necessarily "fair"/ relevant

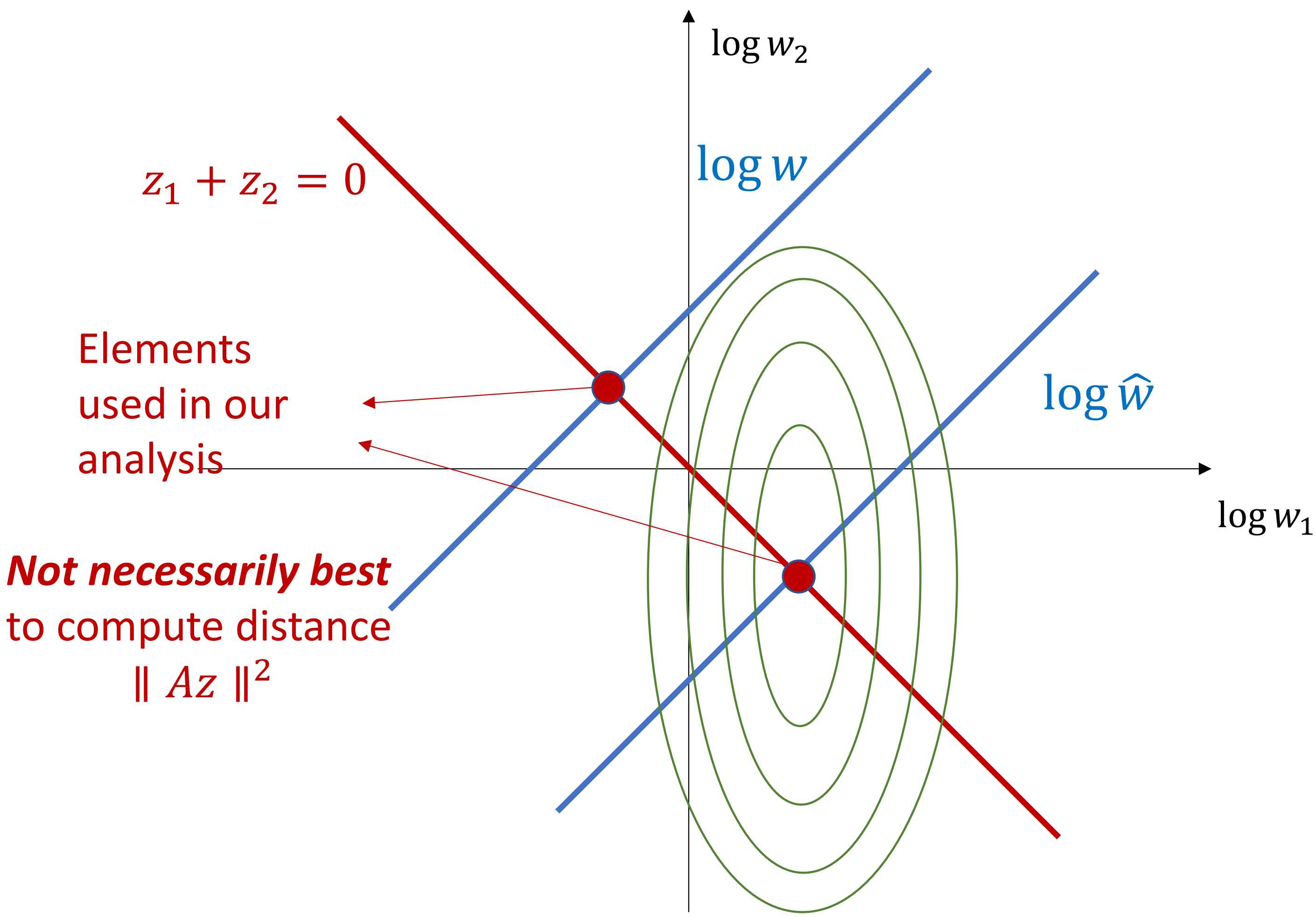
Invariance under addition of constant
→ need to analyze distance between equivalence classes



Invariance under addition of constant
→ need to analyze distance between equivalence classes



Invariance under addition of constant
→ need to analyze distance between equivalence classes



Other performance Criteria: Summary

- ***Quadratic*** $\mathrm{E}\ \| A\Delta \log w\ \|^2$

  - Result and minimax optimality extend
  - Direct approach $\frac{1}{k}Tr(AL_V^{\dagger}A^T)$ valid if $A1 = 0$
  - Also simple expression for full rank A.

In particular error on $(\log w_i - \log w_j)$

$$\mathrm{E}\ \| \Delta \log w_i - \Delta \log w_j\ \|^2 = \frac{1}{\mathrm{k}} Tr\left((\mathrm{e_i} - \mathrm{e_j})^T L_V^{\dagger} (\mathrm{e_i} - \mathrm{e_j}) \right) = \Omega_{V,ij}$$

Resistance between i and j

- ***Nonlinear criteria:*** ex: $\sin(w, \widehat{w})$
  - Also extends under assumptions
  - Based on $\| \nabla V \Delta \log w\ \|^2$

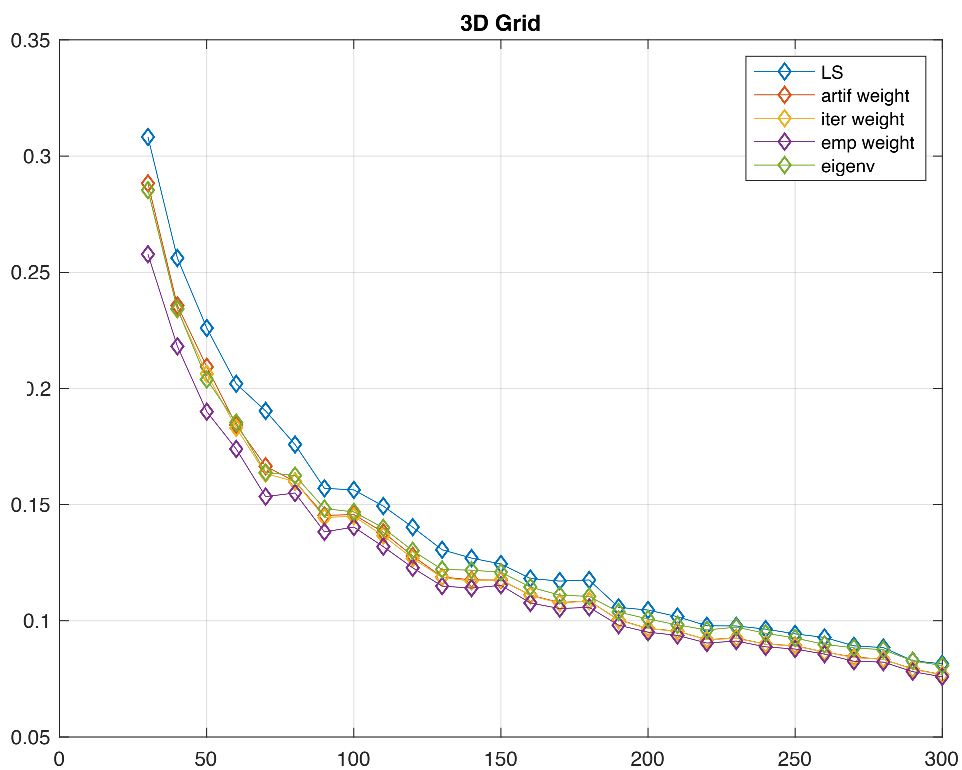
Ranking from pairwise comparisons

- Motivation and Problem
- Weighted Least-Square Estimator
- Algorithm and Complexity
- Error Analysis
  - Error Bound
  - Lower Bound – Minimax Optimality
  - Other criteria
- ***Experimental Results***
- A Surprising Observation
- Generalizations
- Conclusions

3D grid

125 nodes

w_i i.i.d. geometric distribution in [1, 20]

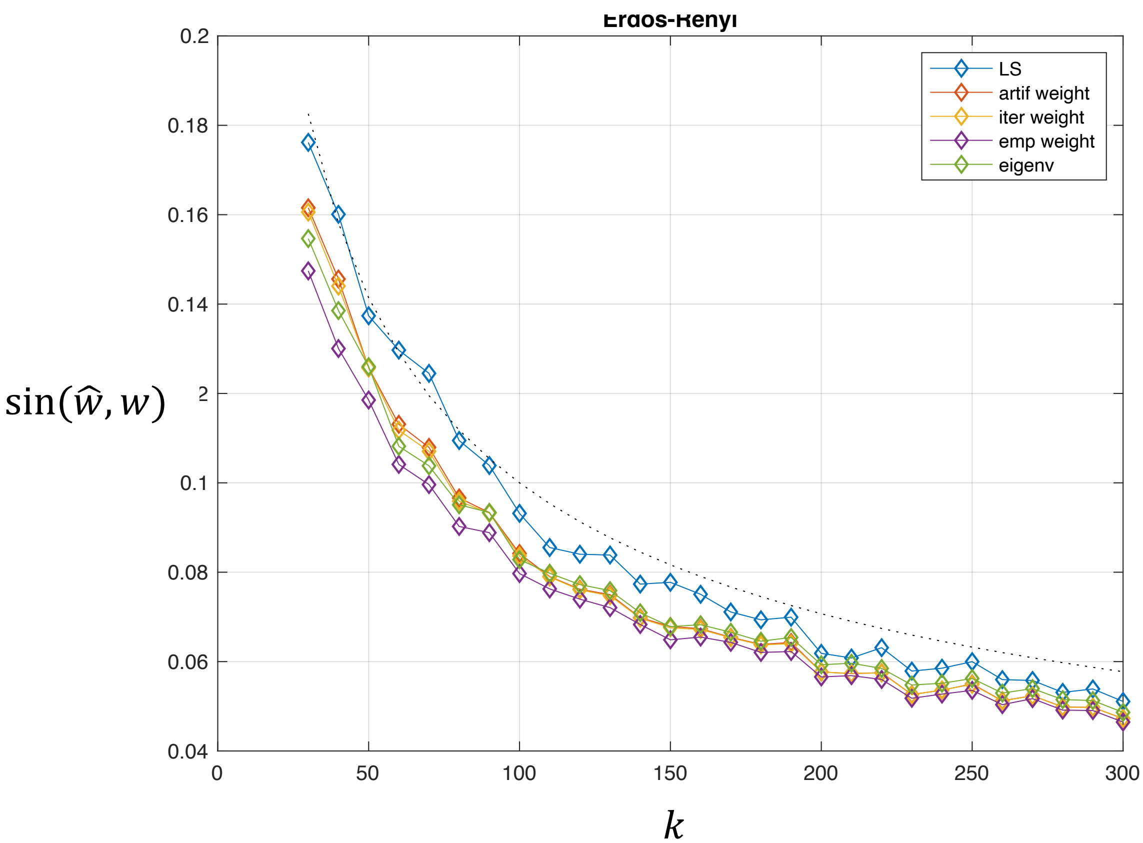


$\sin(\hat{w}, w)$

k

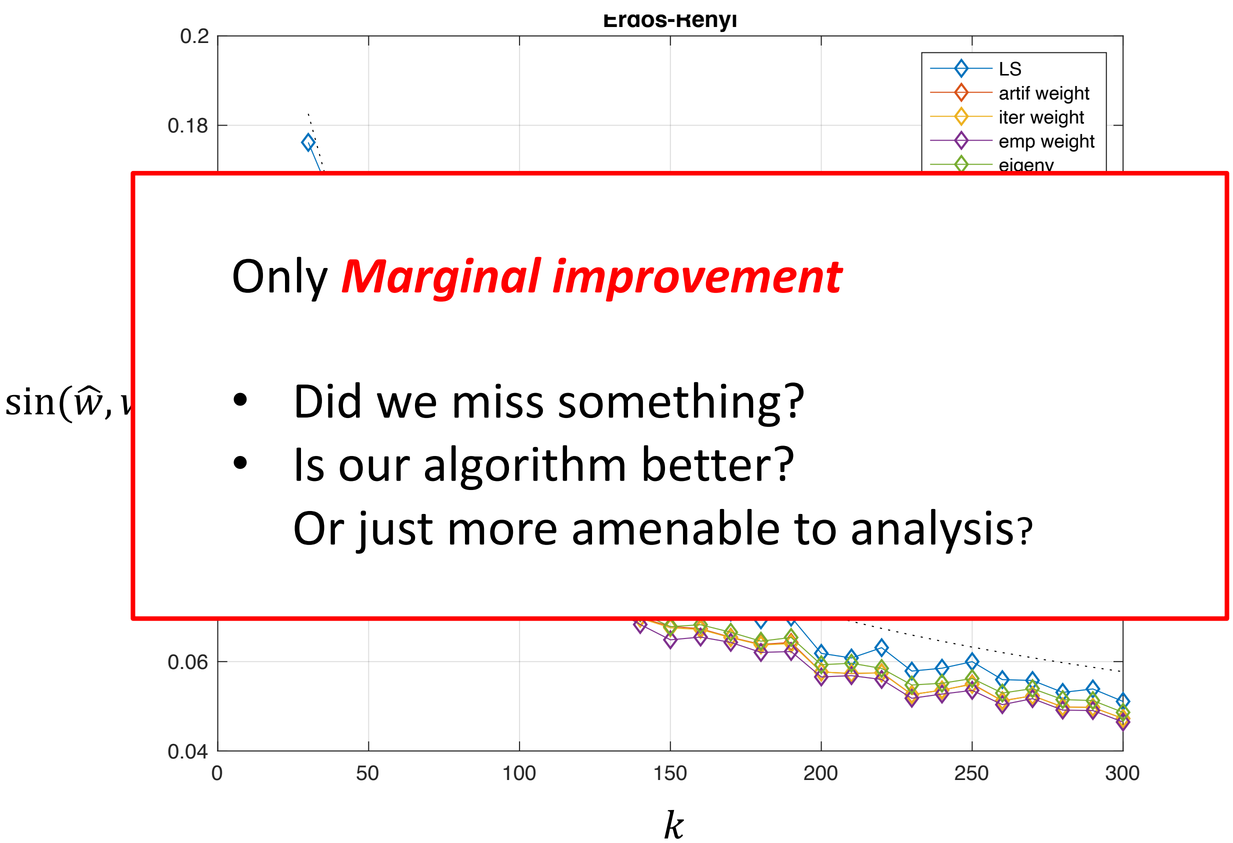
Erdos-Renyi

100 nodes, avg degree 10
w_i i.i.d. geometric distribution in [1, 20]



Erdos-Renyi

100 nodes, avg degree 10
w_i i.i.d. geometric distribution in [1, 20]



Only ***Marginal improvement***

- Did we miss something?
- Is our algorithm better?
  Or just more amenable to analysis?

Worst-case ≠ Typical case for a distribution

- Eigenvector method [Negahban 16] does indeed appear to perform better than its bound.
- But, $\simeq$ as weighted least-square method with weights

$$\left(\frac{1}{\frac{1}{w_i}+\frac{1}{w_j}}\right)^2 \qquad \text{Vs our} \qquad \frac{1}{\frac{w_i}{w_j}+2+\frac{w_j}{w_i}}$$

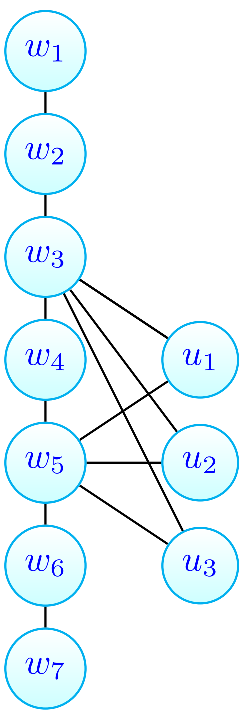
Grows with $\sqrt{w_i w_j}$ Only depends on ratio w_i/w_j

→ ***Neglects information*** combing from edges between “small weights”

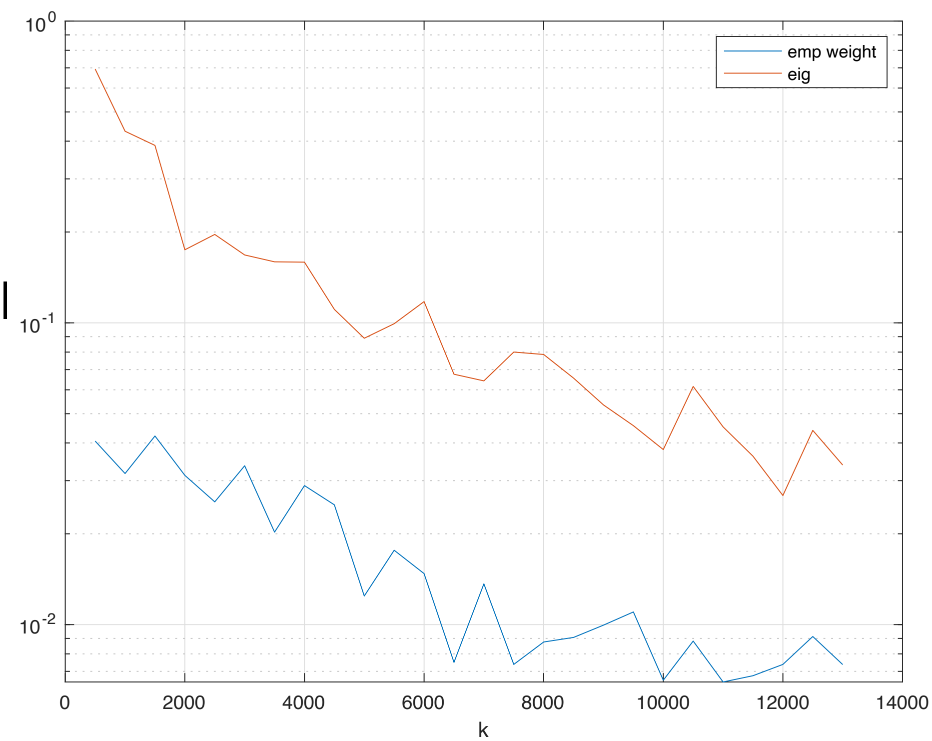
But effect can be ***averaged out*** when weights i.i.d. randomly selected

On a specific graph

(50 nodes u_i)



Error on $|W_3 - W_5|$



Weights selected so that relevant information between small values

Conclusion on simulations

- Outperforms previously existing methods
- Effect marginal on “randomized case”
- Significantly more accurate
  - For local differences
  - When information comes from edges between small w_i

Ranking from pairwise comparisons

- Motivation and Problem
- Weighted Least-Square Estimator
- Algorithm and Complexity
- Error Analysis
  - Error Bound
  - Lower Bound – Minimax Optimality
  - Other criteria
- Experimental Results
- ***A Surprising Observation***
- Generalizations
- Conclusions

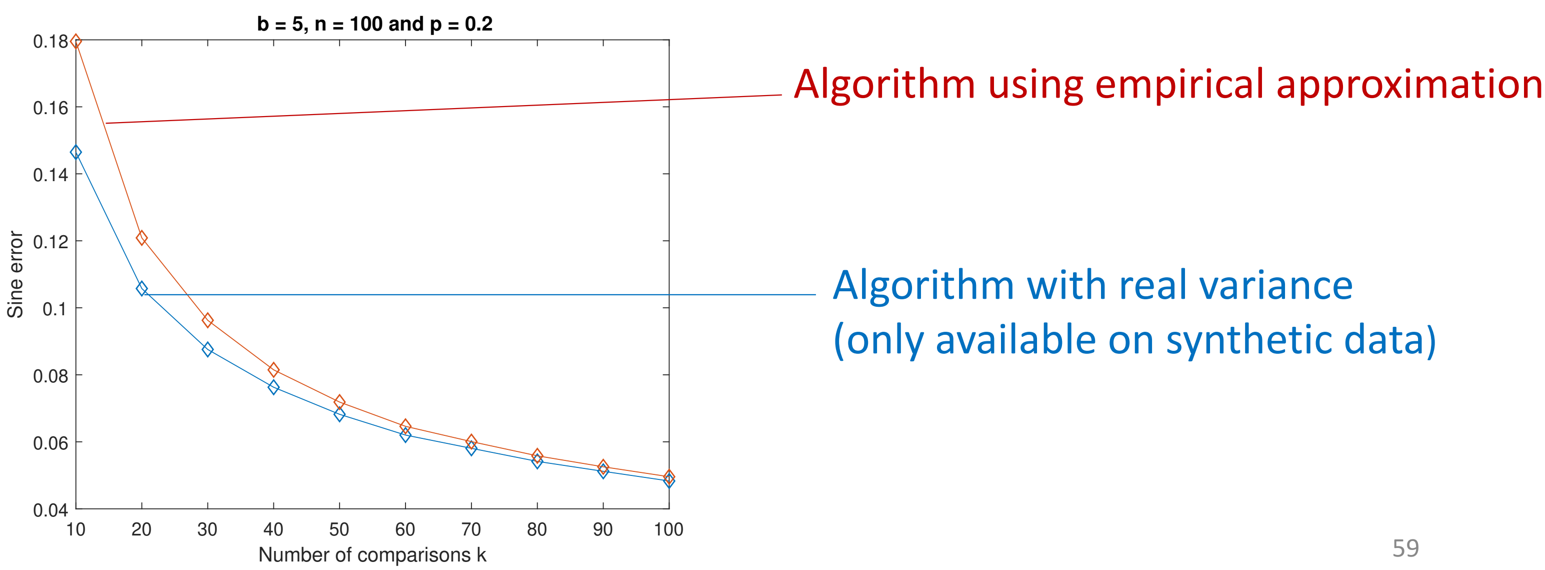
Impact of variance approximation

Idealized algorithm uses $v_{ij} := \frac{w_i}{w_j} + 2 + \frac{w_j}{w_i}$

Not available → approximated by empirical $\hat{v}_{ij} := R_{ij} + 2 + R_{ij}^{-1}$

Theoretical analysis: empirical approx. shown “not to degrade solution too much”

But Experimentally: Empirical variance ***<u>outperforms real</u>*** one



Implicit "regularization"

k=10: $w_1 = 8, w_2 = 2$

	Prob.	$\log R_{ij}$	$\hat{v}_{12}$	Weight in least square
8 wins (expected)	30%	$\log \frac{8}{2} \simeq 1.38$	$\frac{8}{2} + 2 + \frac{2}{8} = 6.25$	0.16
7 wins	20%	$\log \frac{7}{3} \simeq 0.85$ - 38%	$\frac{7}{3} + 2 + \frac{3}{7} = 4.76$	0.21 + 30%
9 wins	26%	$\log \frac{9}{1} \simeq 2.19$ + 58%	$\frac{9}{1} + 2 + \frac{1}{9} = 11.11$	0.09 - 43%

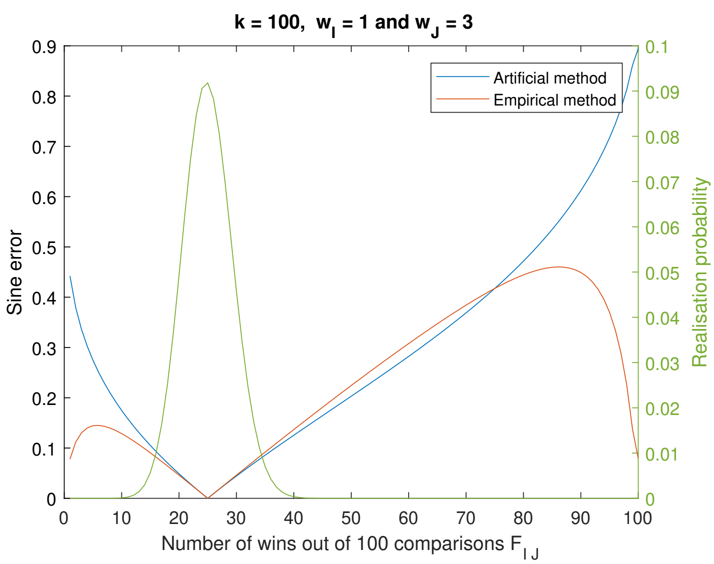
Empirical variance appear to "smoothen outs" dangerous outlyers.

Experimental validation

3 node graphs, $W_I = 1, W_J = 3$ → 25 wins expected

Edges towards W_K set artificially at expected value

Impact of # wins + probability



Contribution to error

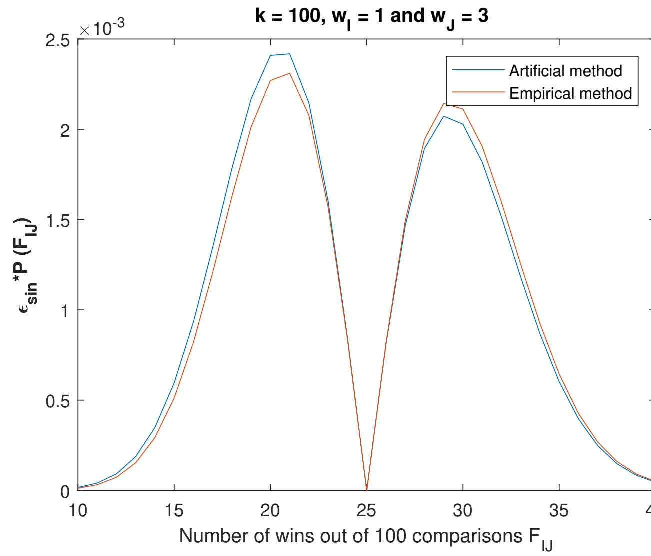


Figure 5.8: $\epsilon(F_{IJ}) * P(F_{IJ})$ for $F_{IJ} \in [10, 40]$

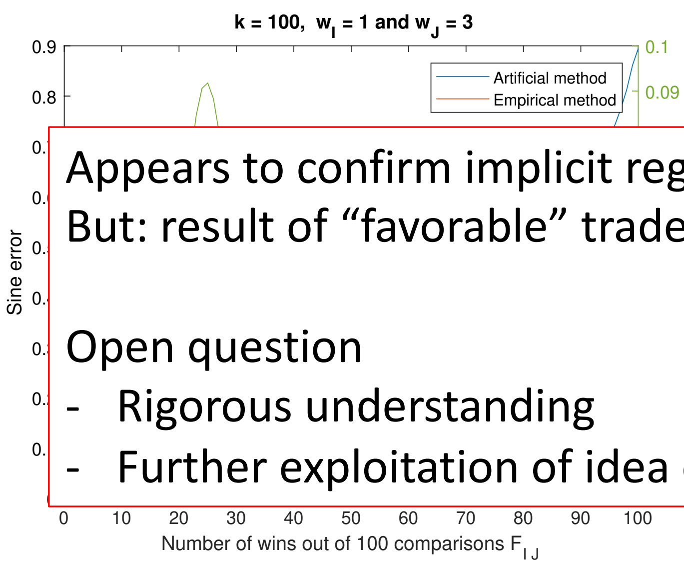
Winand, M., & Hendrickx, J. (2021). Learning from pairwise comparisons: an empirical analysis. *Ecole polytechnique de Louvain, Université catholique de Louvain*.

Experimental validation

3 node graphs, $W_I = 1, W_J = 3$ → 25 wins expected

Edges towards W_K set artificially at expected value

Impact of # wins + probability



Contribution to error

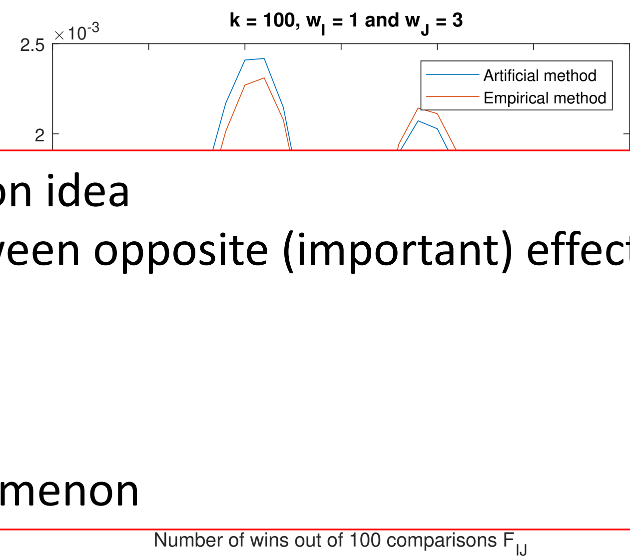


Figure 5.8: $\epsilon(F_{IJ}) * P(F_{IJ})$ for $F_{IJ} \in [10, 40]$

Appears to confirm implicit regularization idea

But: result of “favorable” trade-off between opposite (important) effects

Open question

- Rigorous understanding
- Further exploitation of idea or phenomenon

Winand, M., & Hendrickx, J. (2021). Learning from pairwise comparisons: an empirical analysis. *Ecole polytechnique de Louvain, Université catholique de Louvain*.

Ranking from pairwise comparisons

- Motivation and Problem
- Weighted Least-Square Estimator
- Algorithm and Complexity
- Error Analysis
  - Error Bound
  - Lower Bound – Minimax Optimality
  - Other criteria
- Experimental Results
- A Surprising Observation
- ***Generalizations***
- Conclusions

Relaxing Assumptions

- Same number k of comparisons on every edge
  - Can be relaxed,
  - Some technical aspects
  - Ratio min/max # comparison for some results

- i.i.d. comparisons
  - Bounded dependence between comparison (most likely) OK
  - Persistent dependence between edges → adapting variance

Extending the notion of comparison

- Pick best out of three
- Rank three
- Comparison with ties…

- Many extensions possible (only approximative analysis so far) but depends on model specifics

  Branders, M., Vekemans, A., & Hendrickx, J. *Recovering weights from comparison results in extensions of BTL model*

- Multi-comparisons: sometimes non-diagonal Variance Matrix (expression of least square in terms of non-independent events)

- Game : find relation of the type

  $w_i^{q_i} w_j^{q_j} w_k^{q_k} \simeq$ some function of the outcome (for large k)

Other models - criteria

Bradley-Terry-Luce $p_{ij} = \frac{w_i}{w_i + w_j}$ Other models?

- Results extend to large class of ordinal models:

$$p_{ij} = f(\phi(\beta_i) - \phi(\beta_j))$$

BTL:
- $\phi = \log$
- $f(z) = \frac{1}{1+e^z}$

- Technical assumption needed (e.g. f log-concave)
- Not 100% clear yet which ones are actually necessary

- Extension to (asymptotically) any continuous quality criterion

Conclusions

- Quality of items recovered from results of comparisions on netork → ranking
- Near-linear time algorithm.
- Linear least-square, ***coefficients*** **nonlinear** in data.
- No hyperparameters, tuning etc.
- Outperforms past methods, Minimax optimal
- Performances Driven by $L_V^{\dagger}$ and ***Resistance of comparison graph***
- Many possible generalizations
- Implicit regularization, not fully understood

Some further research directions

- Online version
  - Comparison arriving one by one
  - Choosing Comparison based on past data
  - Explore and Exploit

- Regime of small # comparisons (large n)
- Prior Incorporation?
- Exploitation of implicit regularization

Thank you for your attention

Alex Olshevsky (BU), Venkatesh Saligrama (BU) Balint Daroczy

Maxime Winand Marine Branders Astrid Vekemans

+ Open position to be filled ASAP

julien.hendrickx@uclouvain.be

References

- Hendrickx, J., Olshevsky, A., & Saligrama, V.. ***Minimax rate for learning from pairwise comparisons in the BTL model***. ICML 2020
- Hendrickx, J., Olshevsky, A., & Saligrama, V. ***Graph resistance and learning from pairwise comparisons***. ICML 2019
- Daroczy B., Hendrickx, J., Olshevsky, A., & Saligrama, V. ***Minimax rate for learning ordinal models from pairwise comparisons,*** *coming soon*

- Branders, M., Vekemans, A., & Hendrickx, J. ***Recovering weights from comparison results in extensions of BTL model***, Ms Thesis EPL UCLouvain 2022
- Winand, M., & Hendrickx, J. ***Learning from pairwise comparisons: an empirical analysis***. MS Thesis EPL UCLouvain 2021