Seriation & Ranking: Spectral Approach

Fajwel Fogel, *CNRS & ENS, Paris.*

with Alexandre d'Aspremont, Francis Bach, Rodolphe Jenatton, & Milan Vojnovic

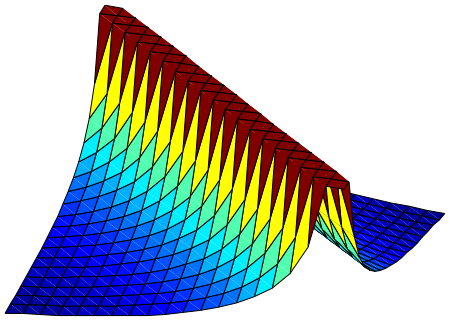
CNRS, INRIA, ENS Paris & MSR Cambridge

The seriation problem

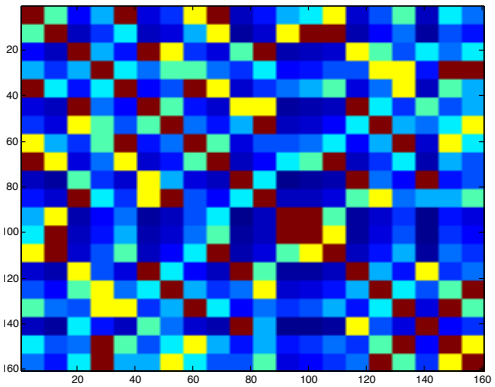
- Pairwise **similarity information** S_{ij} on n variables.
- Suppose the data has a **serial structure**, i.e. there is an order π such that

$$S_{\pi(i)\pi(j)} \text{ decreases with } |i-j| \quad \textbf{(R-matrix)}$$

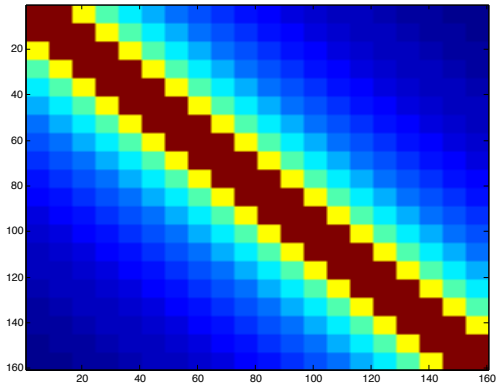
Recover π?



Similarity matrix



Input

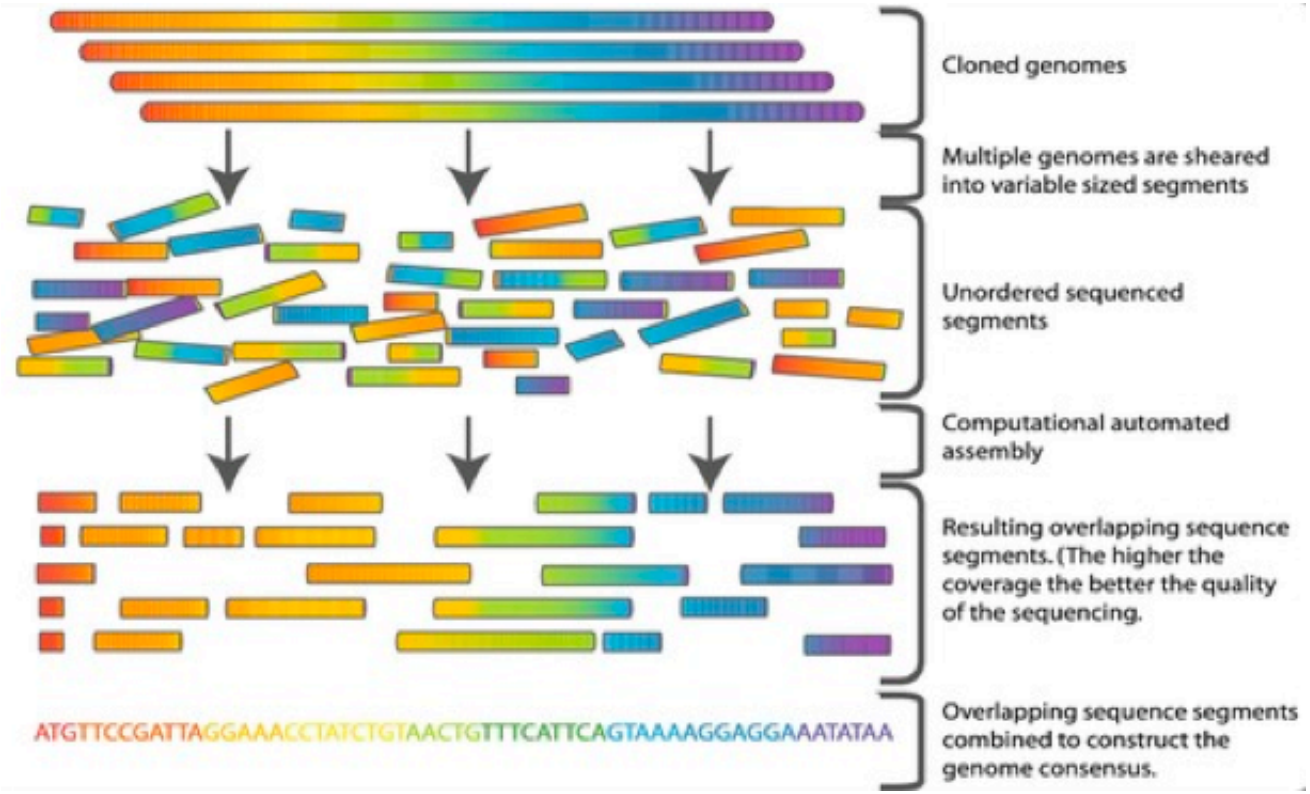


Reconstructed

DNA de novo assembly

Seriation has direct applications in DNA de novo assembly.

- Genomes are cloned multiple times and randomly cut into shorter reads ($\sim$ 400bp), which are fully sequenced.
- Reorder the reads to recover the genome.



(from Wikipedia. . .)

Seriation: a combinatorial problem

- **Combinatorial Solution** [FJBA. 2013, Laurent et Seminaroti 2014]

  For R-matrices, **2-SUM $\Longleftrightarrow$ seriation.**

- **2-SUM**: assign similar items to nearby positions in reordering, i.e. find permutation π of items 1 to n that minimizes

$$\sum_{i,j=1}^{n} S_{i,j}(\pi(i) - \pi(j))^2. \tag{1}$$

- The 2-SUM problem is **NP-Complete** for generic matrices S [George and Pothen 1997].

A spectral solution

Spectral Seriation. Define the Laplacian of S as $L_S = \mathbf{diag}(S\mathbf{1}) - S$, the Fiedler vector of S is written

$$f = \underset{\substack{\mathbf{1}^T x = 0, \\ \|x\|_2 = 1}}{\operatorname{argmin}} \; x^T L_S x.$$

and is the second smallest eigenvector of the Laplacian.

The Fiedler vector reorders a R-matrix in the noiseless case.

Theorem [Atkins, Boman, Hendrickson, et al., 1998]

Spectral seriation. *Suppose $S \in \mathbf{S}_n$ is a pre-R matrix, with a simple Fiedler value whose Fiedler vector f has no repeated values. Suppose that $\Pi \in \mathcal{P}$ is such that the permuted Fielder vector Πv is monotonic, then $\Pi S \Pi^T$ is an R-matrix.*

Spectral solution: advantages

- **Exact** for R-matrices.
- Quite **robust** to noise. Arguments similar to perturbation results in spectral clustering.
- **Scales** very well, especially when similarity matrix is sparse (as in DNA sequencing and ranking).

Ranking with pairwise comparisons

Ranking

Goal: given pairwise comparisons between a set of items, find the most consistent global order of these items.

Applications

- sports competitions (e.g. chess, football. . .)
- crowdsourcing services (e.g. TopCoder. . .)
- online computer games. . .

Ranking

Classical methods

- ranking by **score** (e.g. #wins - #losses) [Huber, 1963; Wauthier et al., 2013]
- ranking by "skills" under a **probabilistic model** [Bradley and Terry, 1952; Luce, 1959; Herbrich et al., 2006]
- ranking according to **principal eigenvector** of a transition matrix [Page et al., 1998; Negahban et al., 2012]
- . . .

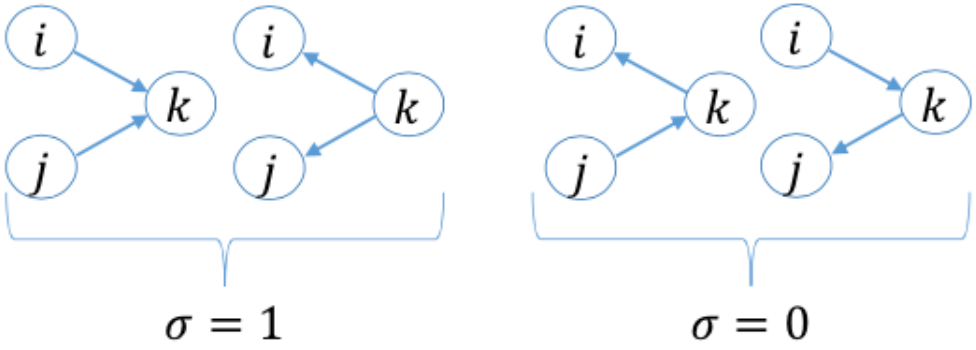
Two main issues

- **missing** comparisons
- **non transitive** comparisons (i.e. $a < b$ and $b < c$ but $a > c$).

Ranking

Casting the ranking problem as a seriation problem

- **Input**: a matrix of pairwise comparisons C where $C_{i,j} \in [-1, 1]$ e.g. for a tournament $C_{i,j} \in \{-1, 0, 1\}$ (loss, tie, win)
- **Idea**: count matching comparisons of i and j against other items k



Example: in a tournament setting, if players i and j had the same outcomes against other opponents k , they should have a similar rank.

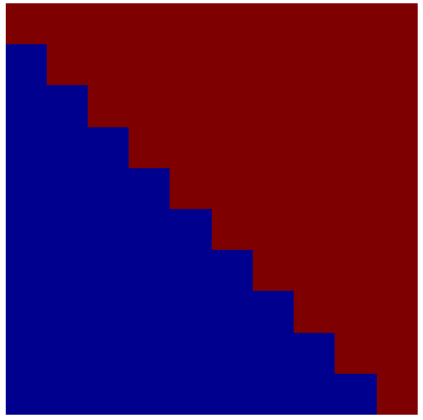
Casting the ranking problem as a seriation problem

- Construct a **similarity** matrix S

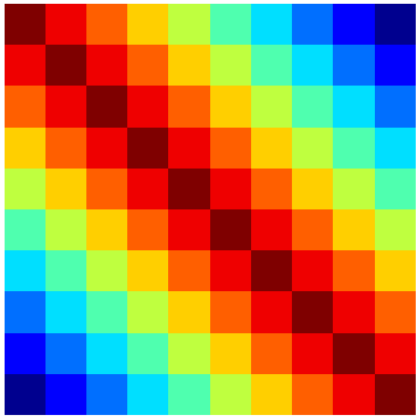
$$S_{i,j} = \sum_{i,j \text{ compared with } k} \sigma(C_{i,k}, C_{j,k}),$$

  where σ is a similarity measure.

- Example: when $\sigma(a, b) = 1 + ab$, $S = n\mathbf{1}\mathbf{1}^T + CC^T$.



Comparison matrix



Similarity matrix

- Is it the right way to solve the ranking problem, in the presence of corrupted and missing comparisons?

SerialRank

New ranking algorithm: SerialRank

- A very simple procedure:
  - compute a similarity matrix from pairwise comparisons ($e.g.$count matching comparisons)
  - solve the corresponding seriation problem ($e.g.$use the spectral solution).
- Might be improved by designing new similarities.

Choice of similarity

- In applications, the design of the similarity can have a **major impact**.
- For ranking, depending on the nature of your data (cardinal or ordinal data, ties etc.), you might adapt your similarity.
- For DNA assembly, you would like to have a similarity robust to sequencing noise.
- Ongoing work...

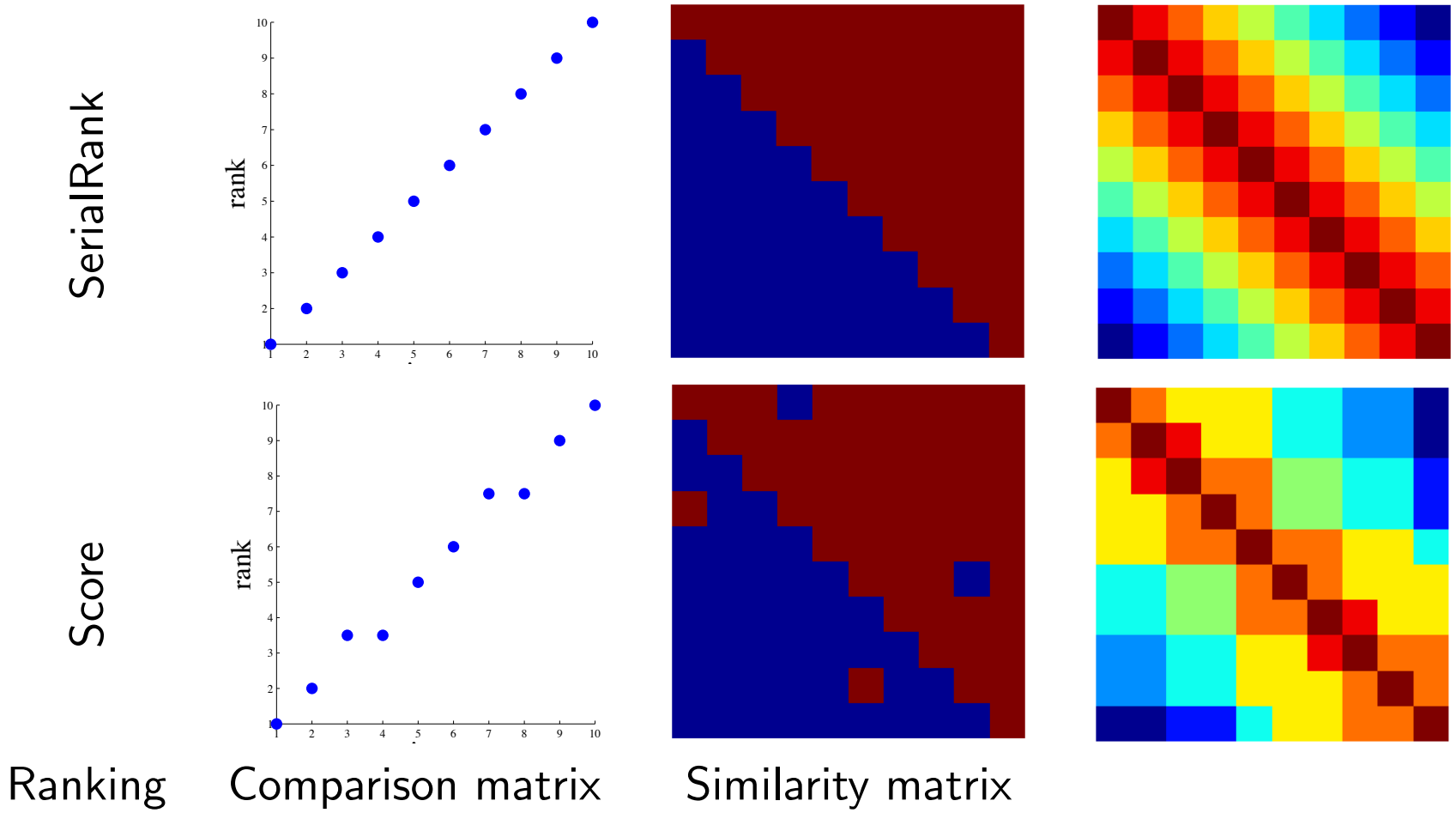
Performance guarantees for SerialRank

- **Robustness to missing/corrupted comparisons**
  Similarity based ranking is more robust than typical score based rankings (i.e. #wins - #losses).

- **Exact recovery regime**
  Exact recovery of underlying ranking with probability $1 - o(1)$ for $o(\sqrt{n})$ random missing/corrupted comparisons.

- **Approximate recovery regime** Competitive to other approaches for partial observations and corrupted comparisons (cf. numerical experiments).

Performance guarantees for SerialRank



All comparisons given, corrupted entries induce ties in score based ranking but not in similarity based ranking.

Perturbation analysis

- Derive asymptotic **analytical expression of Fiedler vector** in noise free setting.
- Use **perturbation results** (i.e. Davis-Kahan) in order to bound the perturbation of the Fiedler vector with missing/corrupted comparisons.
- Get **theoretical guarantees** for SerialRank in settings with only few comparisons available.

Perturbation analysis

Analytical expression of Fiedler vector

- Use results on the **convergence of Laplacian operators** to provide a description of the spectrum of the unperturbed Laplacian.

- Following the same analysis as in [Von Luxburg '08] we can prove that asymptotically, once normalized by n^2, apart from the first and second eigenvalue, the spectrum of the Laplacian matrix is contained in the interval $[0.5, 0.75]$.

- Moreover, we can characterize the eigenfunctions of the limit Laplacian operator ($i.e. \lim \frac{L_n}{n}$) by a differential equation, which gives an **asymptotic analytical expression** for the Fiedler vector.

Perturbation analysis

Analytical expression of Fiedler vector

- Taking the same notations as in [Von Luxburg '08] we have here $k(x, y) = 1 - |x - y|$. The degree function is

$$d(x) = \int_0^1 k(x, y) dP(y) = \int_0^1 k(x, y) d(y)$$

(samples are uniformly ranked).

$$d(x) = -x^2 + x + 1/2.$$

- We deduce that **the range of d is** $[0.5, 0.75]$. Interesting eigenvectors ($i.e.$ here the second eigenvector) are not in this range.

Perturbation analysis

Analytical expression of Fiedler vector

- We can also characterize eigenfunctions f by a **differential equation**

$$Uf(x) = \lambda f(x) \quad \forall x \in [0,1]$$

$$\Rightarrow f''(x)(1/2 - \lambda + x - x^2) + 2f'(x)(1-2x) = 0 \quad \forall x \in [0,1]. \tag{2}$$

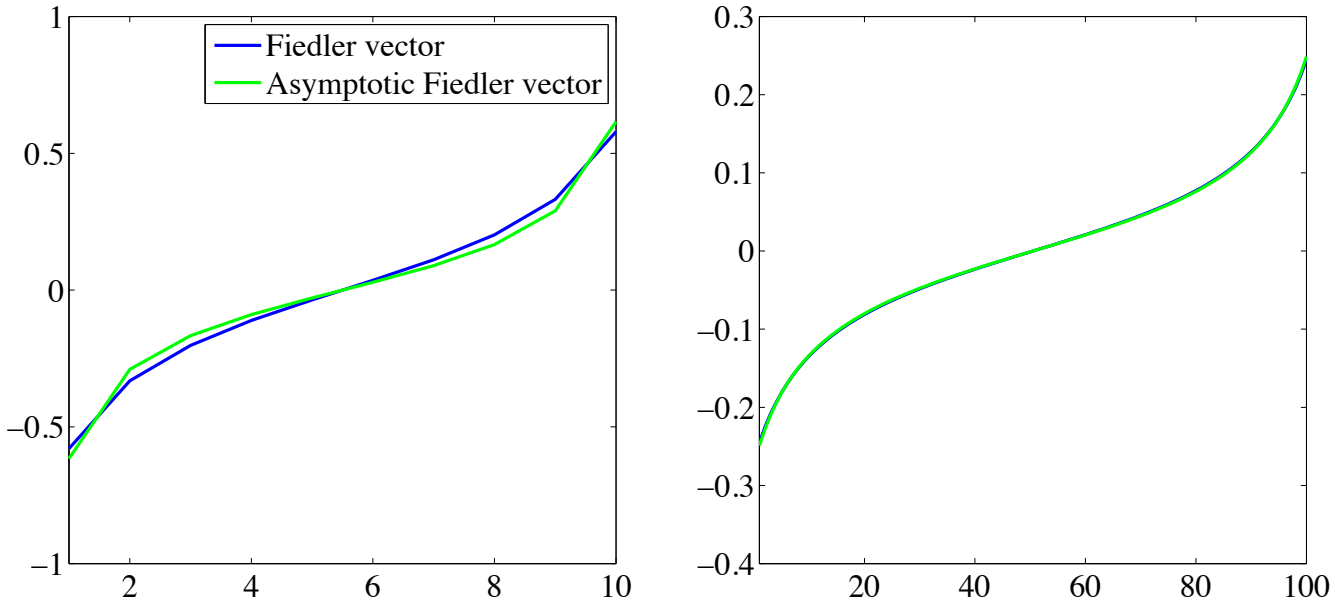
- The asymptotic expression for the Fiedler vector is a solution to this differential equation, with $\lambda < 0.5$.

- **Very accurate numerically**, even for small values of n.

Perturbation analysis

Analytical expression of Fiedler vector

Comparison between the asymptotic analytical expression of the Fiedler vector and the numeric values obtained from eigenvalue decomposition, for $n = 10$ (*left*) and $n = 100$ (*right*).



Perturbation analysis

Goal

Get similar result as for point score method (cf [Wauthier et al., 2013]).

Show that for any precision parameter μ, with a proportion of observations

$$p \gtrsim \frac{\log n}{\mu n}$$

$$\max |\tilde{\pi} - \pi| \lesssim \mu n \quad \text{whp}$$

.

... up to constants and $\log(n)$ factors.

Perturbation analysis

Classical perturbation results

Davis-Kahan Theorem

If $|\hat{\lambda}_3 - \lambda_2| > |\lambda_3 - \lambda_2|/2$ *and* $|\hat{\lambda}_1 - \lambda_2| > |\lambda_1 - \lambda_2|/2$, *then*

$$||f - \hat{f}||_2 \leq \sqrt{2}\frac{||\hat{L} - L||_{\text{op}}}{\min(\lambda_2 - \lambda_1, \lambda_3 - \lambda_2)}.$$

Weyl's Inequality

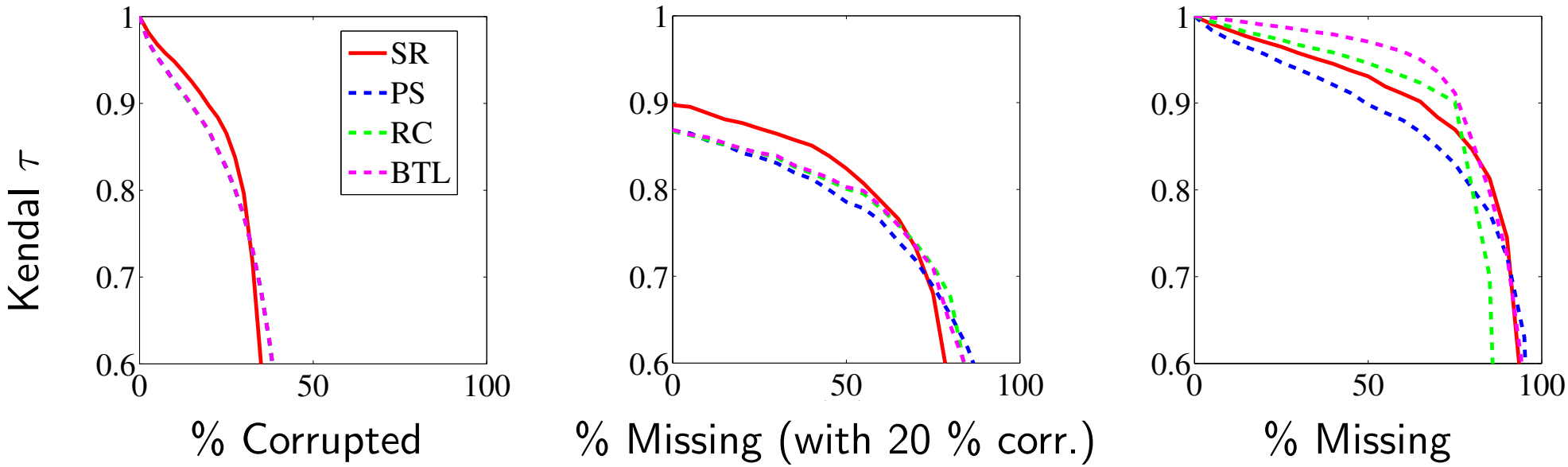
Let L_S *and* $L_{\tilde{S}}$ *be* $n \times n$ *positive definite matrices and let* $L_R = L_{\tilde{S}} - L_S$. *Let* $\lambda_1 \leq \ldots \lambda_n$ *and* $\tilde{\lambda}_1 \leq \ldots \tilde{\lambda}_n$ *be the eigenvalues of* L_S *and* $L_{\tilde{S}}$ *respectively. Then, for all* i, $|\tilde{\lambda}_i - \lambda_i| \leq ||L_R||_2$.

+ concentration inequalities

Numerical results: ranking

Synthetic datasets with random missing/corrupted comp.

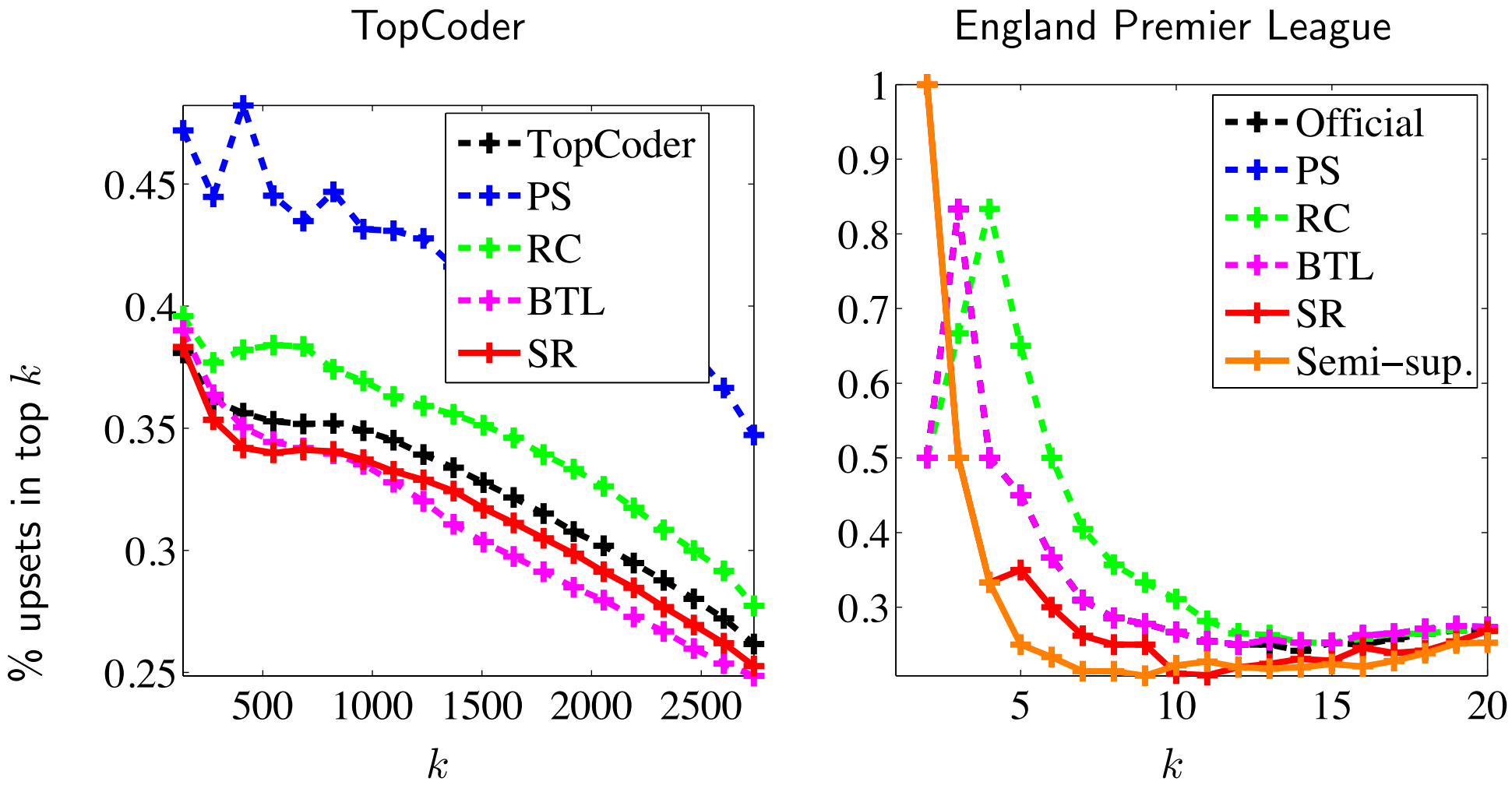
Evaluate Kendall rank correlation coefficient τ between recovered ranking and "true" ranking ($\tau \in [-1, 1]$, $\tau = 1$ means identical rankings).



100 items, SR: SerialRank, PS: point-score, RC: rank centrality, BTL: Bradley-Terry

Numerical results: ranking

Real datasets



SR: SerialRank, PS: point-score, RC: rank centrality, BTL: Bradley-Terry

Conclusion

Results

- **Ranking** as a seriation problem, with perturbation results.
- Good performance on some applications, without specific tuning.

Open problems

- Impact of **similarity** measures.
- Predictive power of SerialRank.

Merci!

- Links to papers & SerialRank tutorial: **www.di.ens.fr/∼fogel**.
- Support from a European Research Council starting grant (project SIPA) and MSR-INRIA Joint Center.

Thanks to the organizers and all the participants!

Bon voyage!