Efficient Random Graph Matching via Degree Profiles

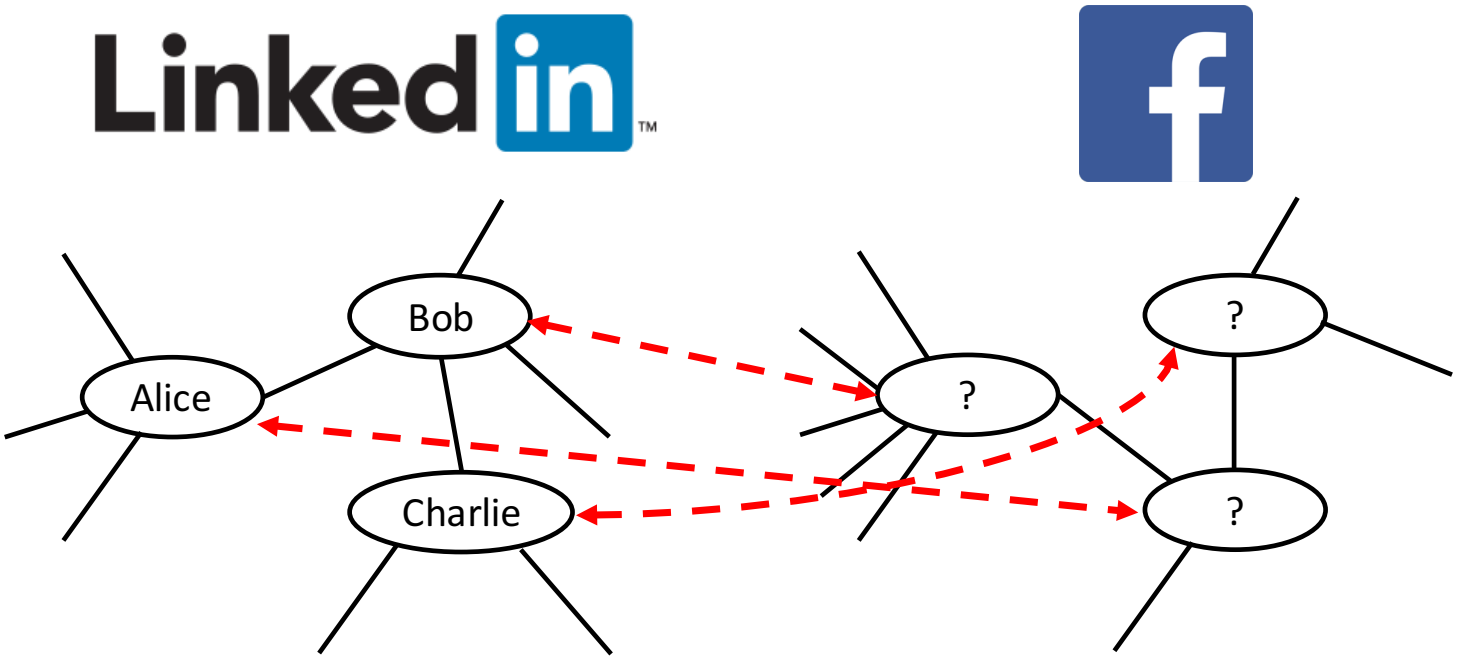
Jiaming Xu

The Fuqua School of Business
Duke University

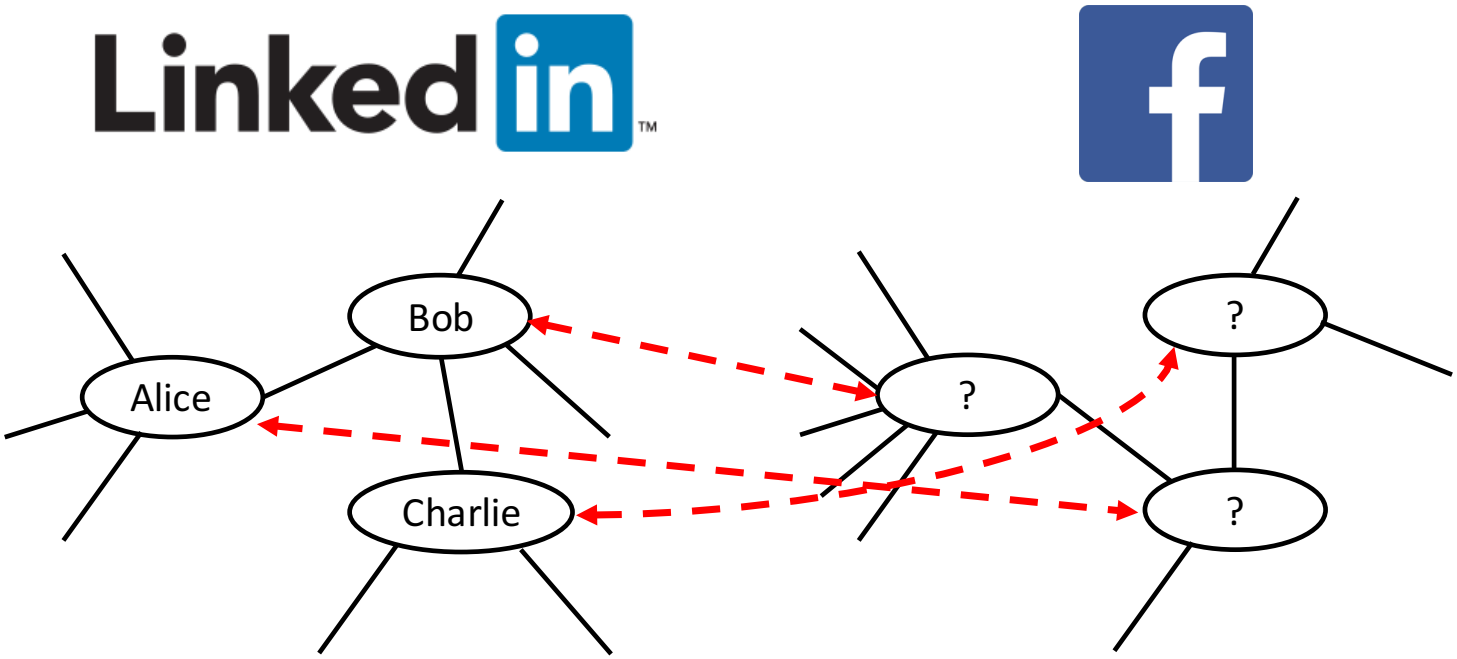
Joint work with
Jian Ding, Statistics department, Wharton School, UPenn
Zongming Ma, Statistics department, Wharton School, UPenn
Yihong Wu, Department of Statistics and Data Science, Yale

Fuqua Summer Seminar Series
July 17, 2019

Motivation 1: Network de-anonymization

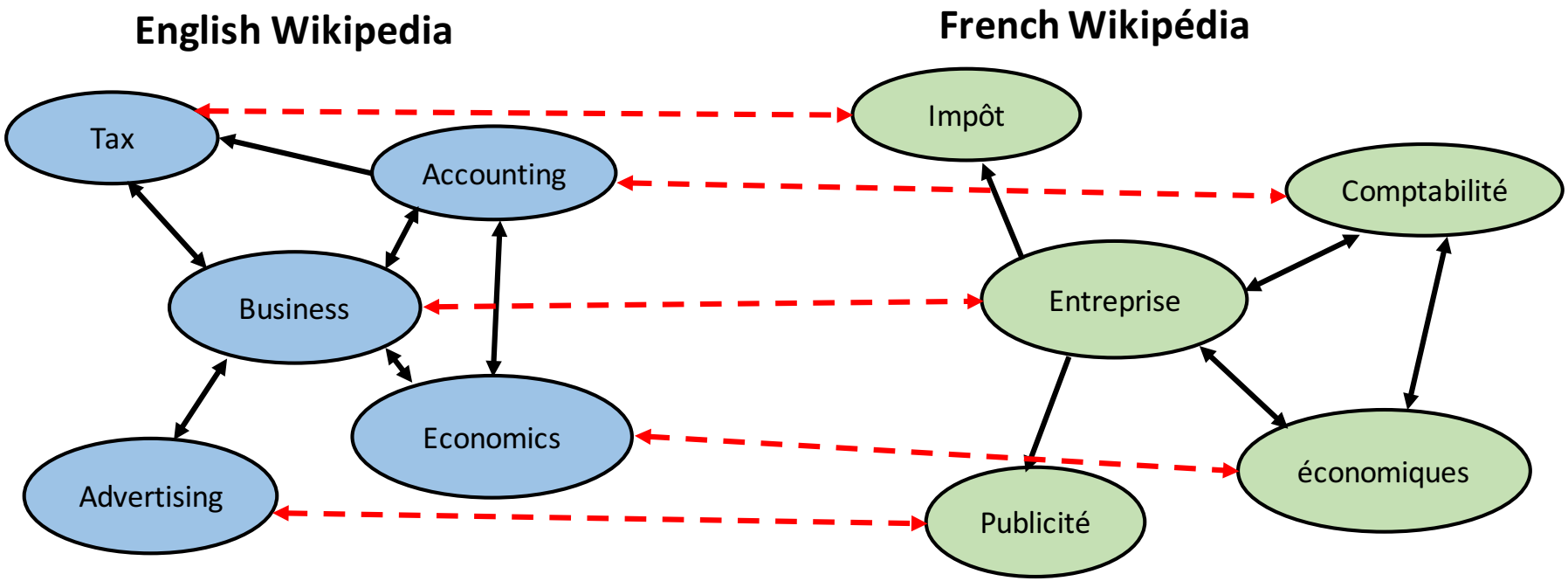


Motivation 1: Network de-anonymization



- Successfully de-anonymize Netflix by matching it to IMDB
  [Narayanan-Shmatikov '08]
- Correctly identify 30.8% of node pairings between Twitter and Flickr
  [Narayanan-Shmatikov '09]

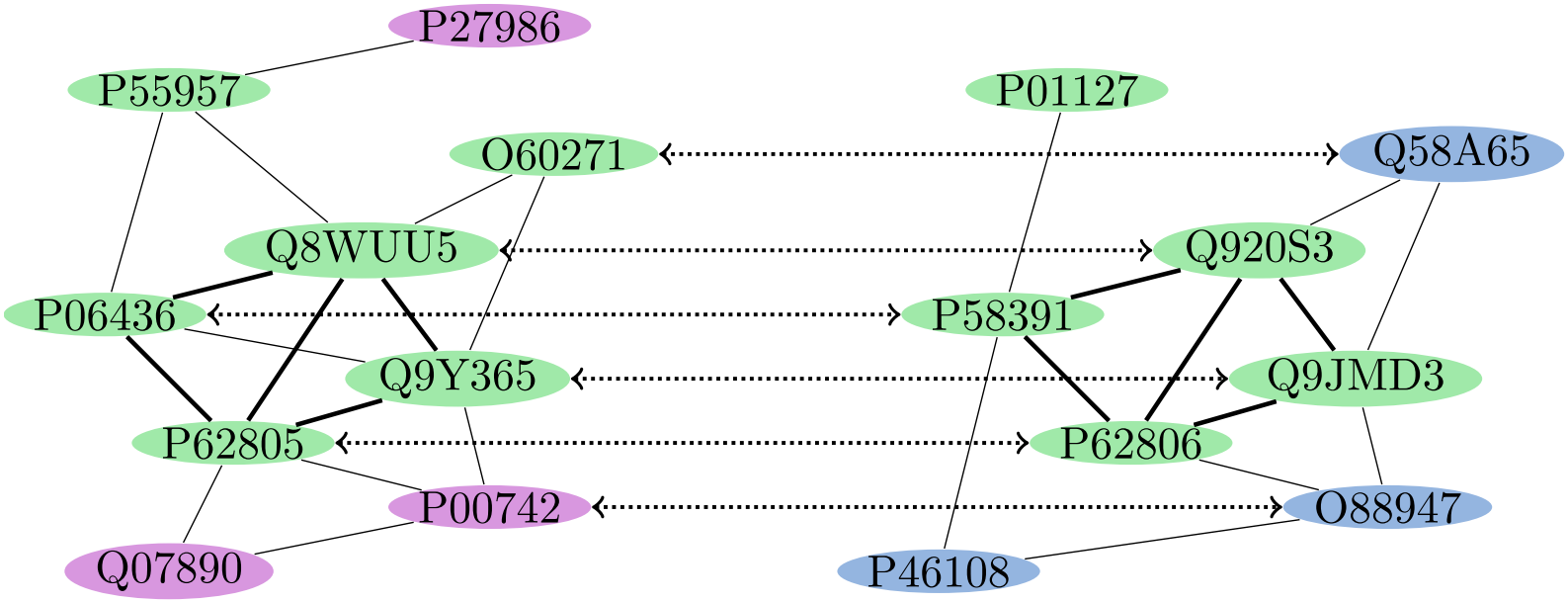
Motivation 2: Machine translation



Automatically find/correct corresp. wiki articles in different languages

[Fishkind-Adali-Patsolic-Meng-Lyzinski-Priebe '12]

Motivation 3: Protein-protein interaction network

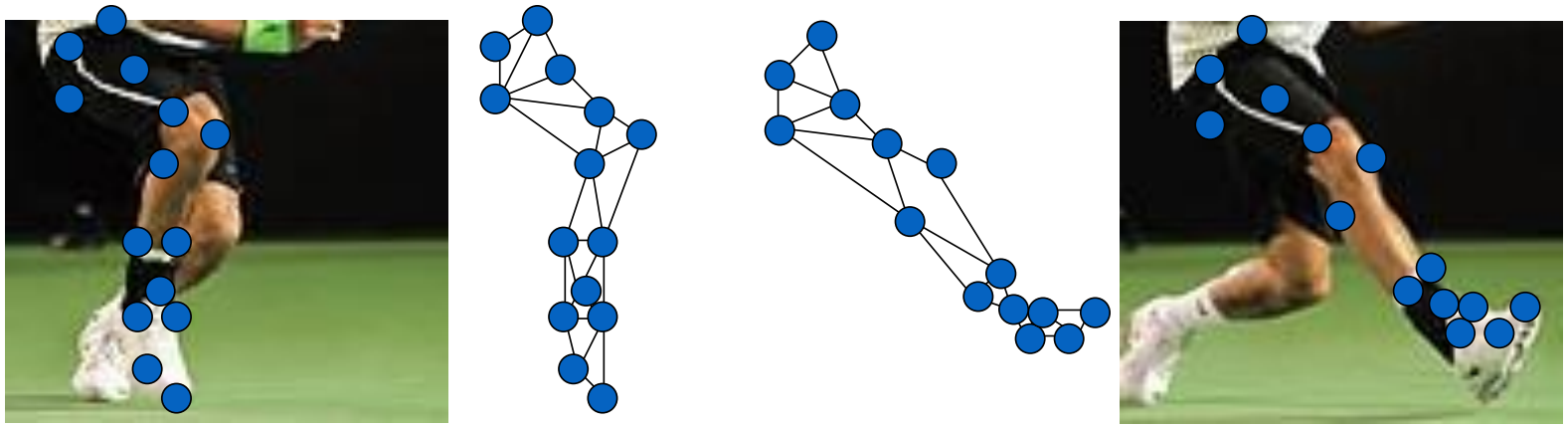


Human network Mouse network

[Kazemi-Hassani-Grossglauser-Modarres '16]

Ontology: Discover proteins with similar functions across different species based on interaction network topology

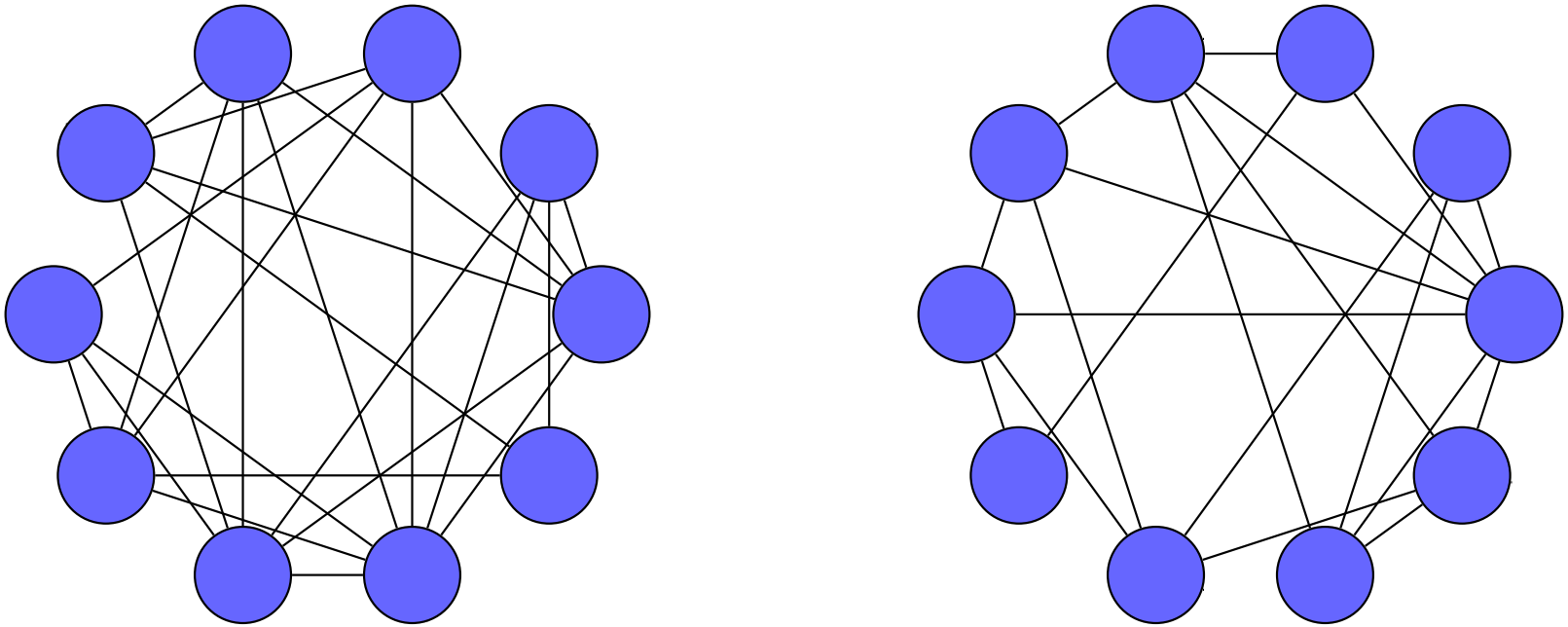
Motivation 4: Computer vision



objects $\rightarrow$ graphs (features $\rightarrow$ nodes, distances $\rightarrow$ edges)
match objects by matching graphs

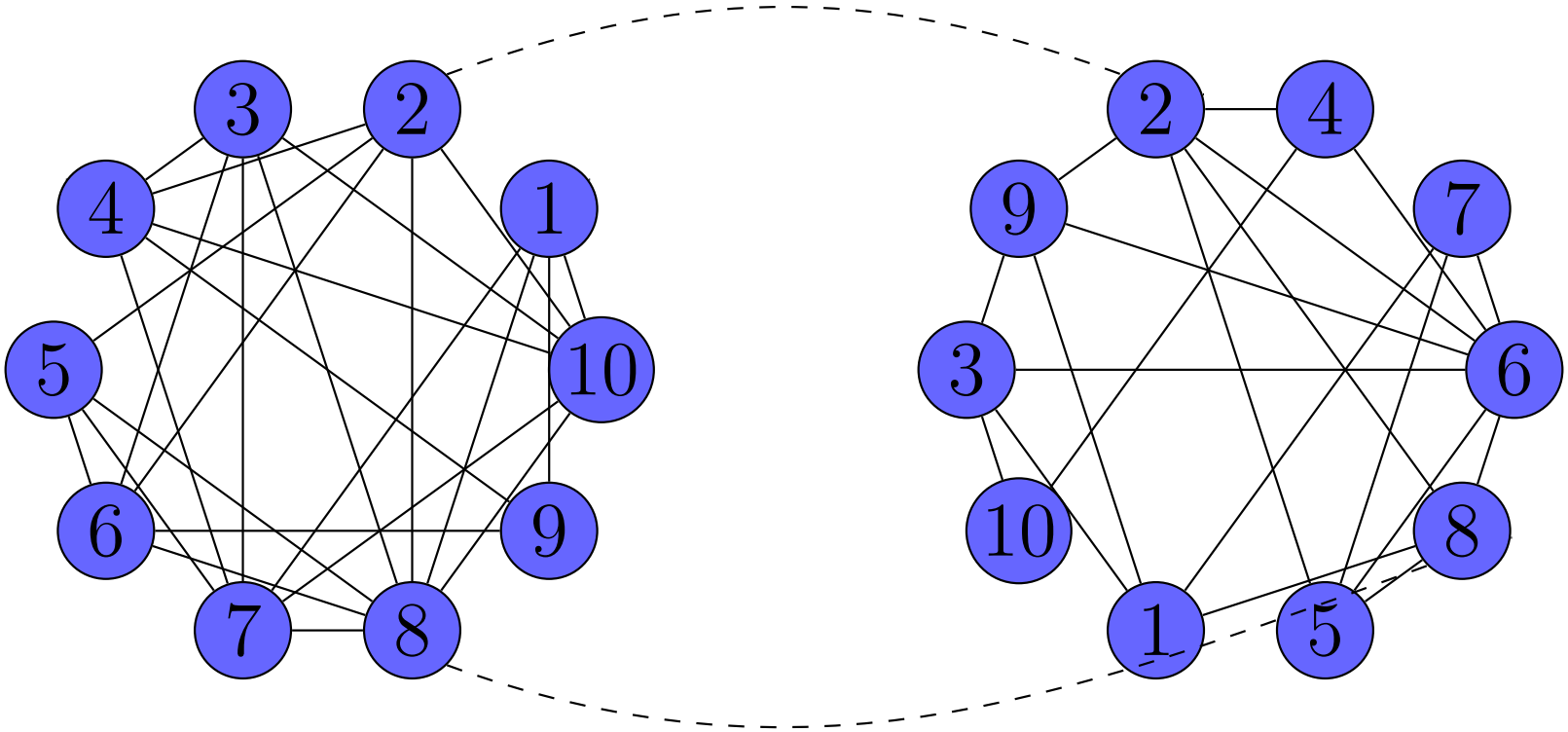
A fundamental problem in computer vision with applications in 3D reconstruction, object tracking, shape matching, image classification, autonomous driving, ...

Graph matching (network alignment)



Goal: find a mapping between two node sets that maximally aligns the edges (i.e. minimizes # of adjacency disagreements)

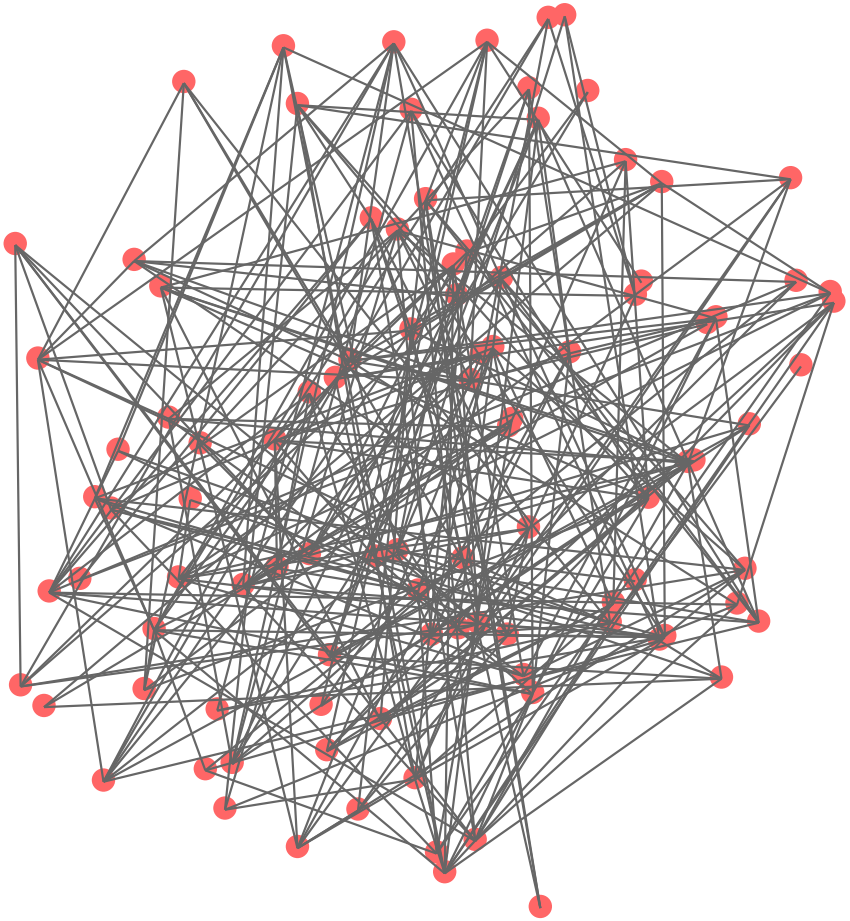
Graph matching (network alignment)

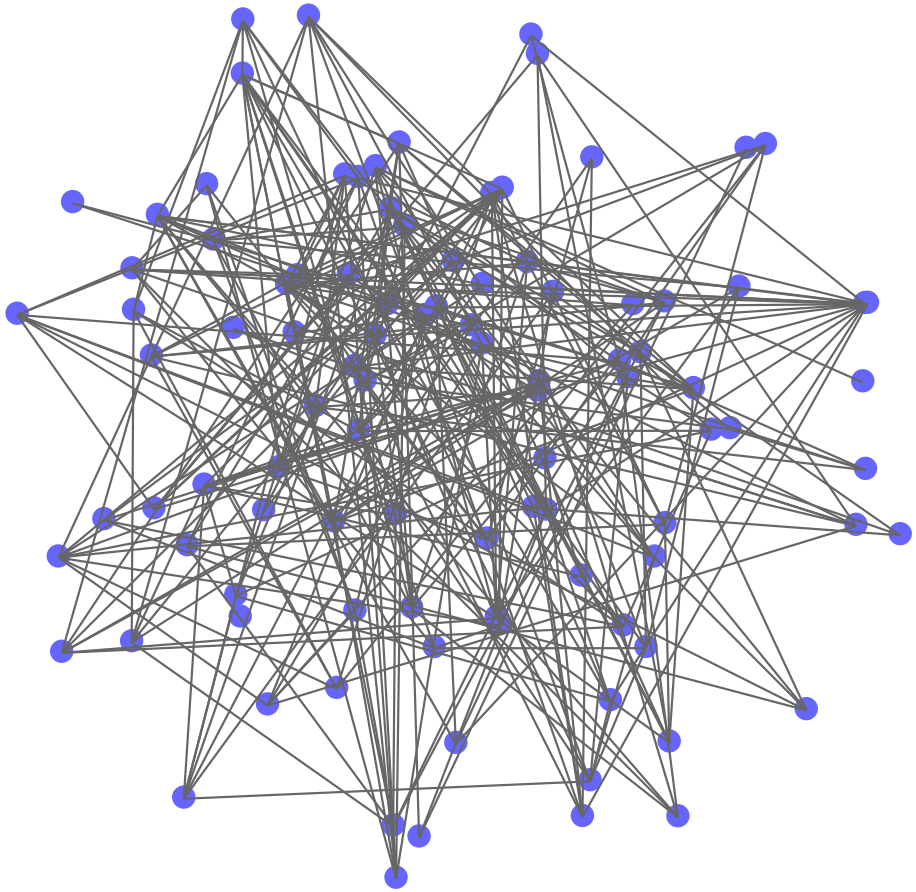


Goal: find a mapping between two node sets that maximally aligns the edges (i.e. minimizes # of adjacency disagreements)

Two key challenges

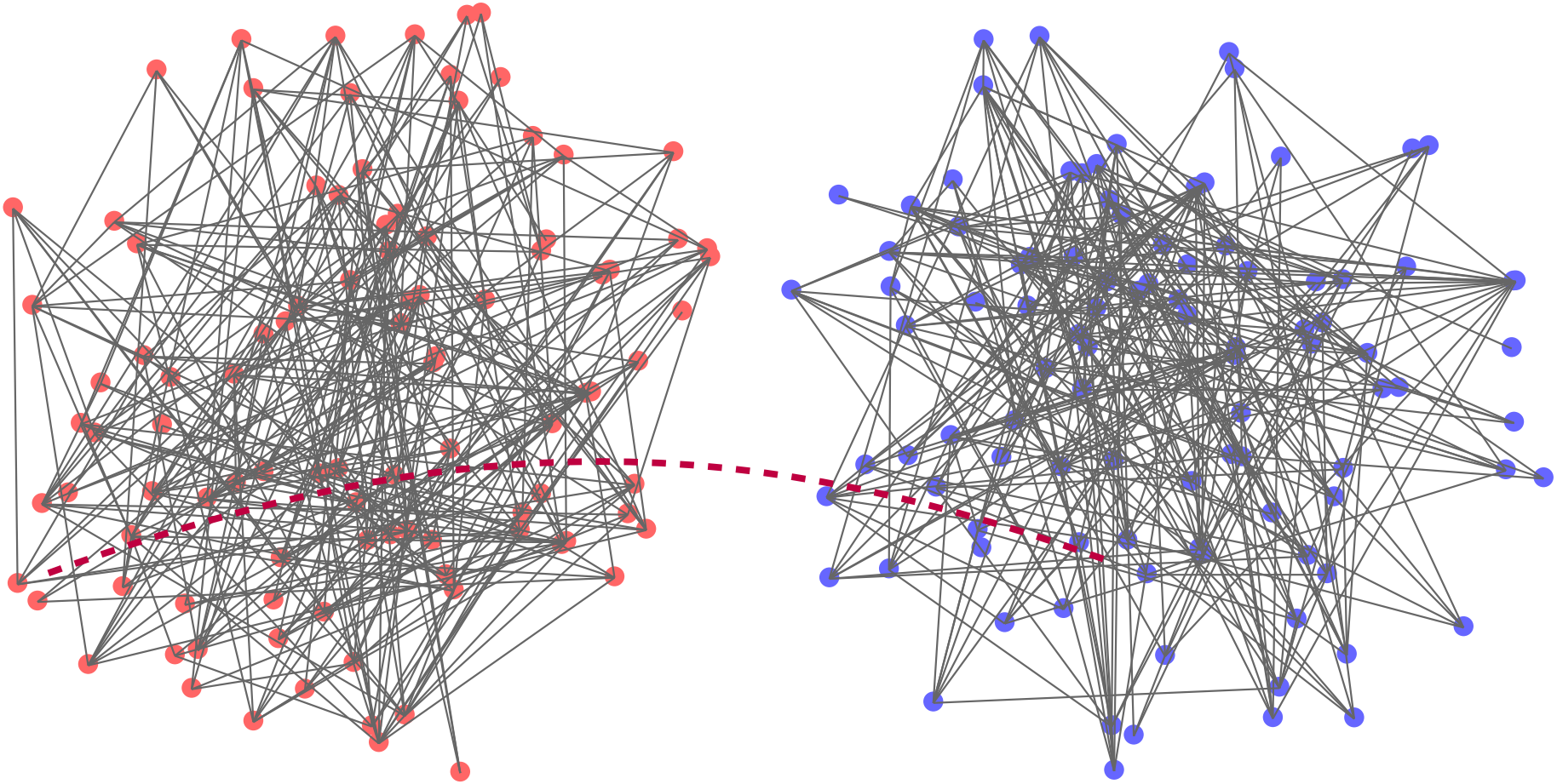
- Statistical: two graphs may not be the same
- Computational: # of possible node mappings is $n!$ ($100! \approx 10^{158}$)





Two key challenges

- Statistical: two graphs may not be the same
- Computational: # of possible node mappings is $n!$ ($100! \approx 10^{158}$)



Beyond computational intractability

- NP-hard for matching two general graphs
- However, real networks are not arbitrary and have latent structures

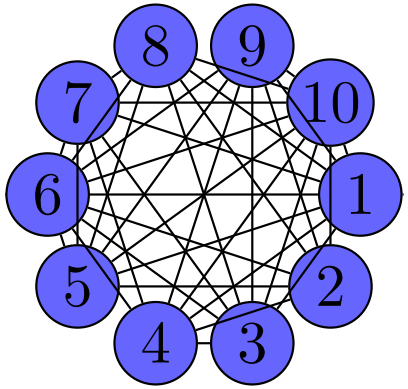
Beyond computational intractability

- NP-hard for matching two general graphs
- However, real networks are not arbitrary and have latent structures

Focus of this talk

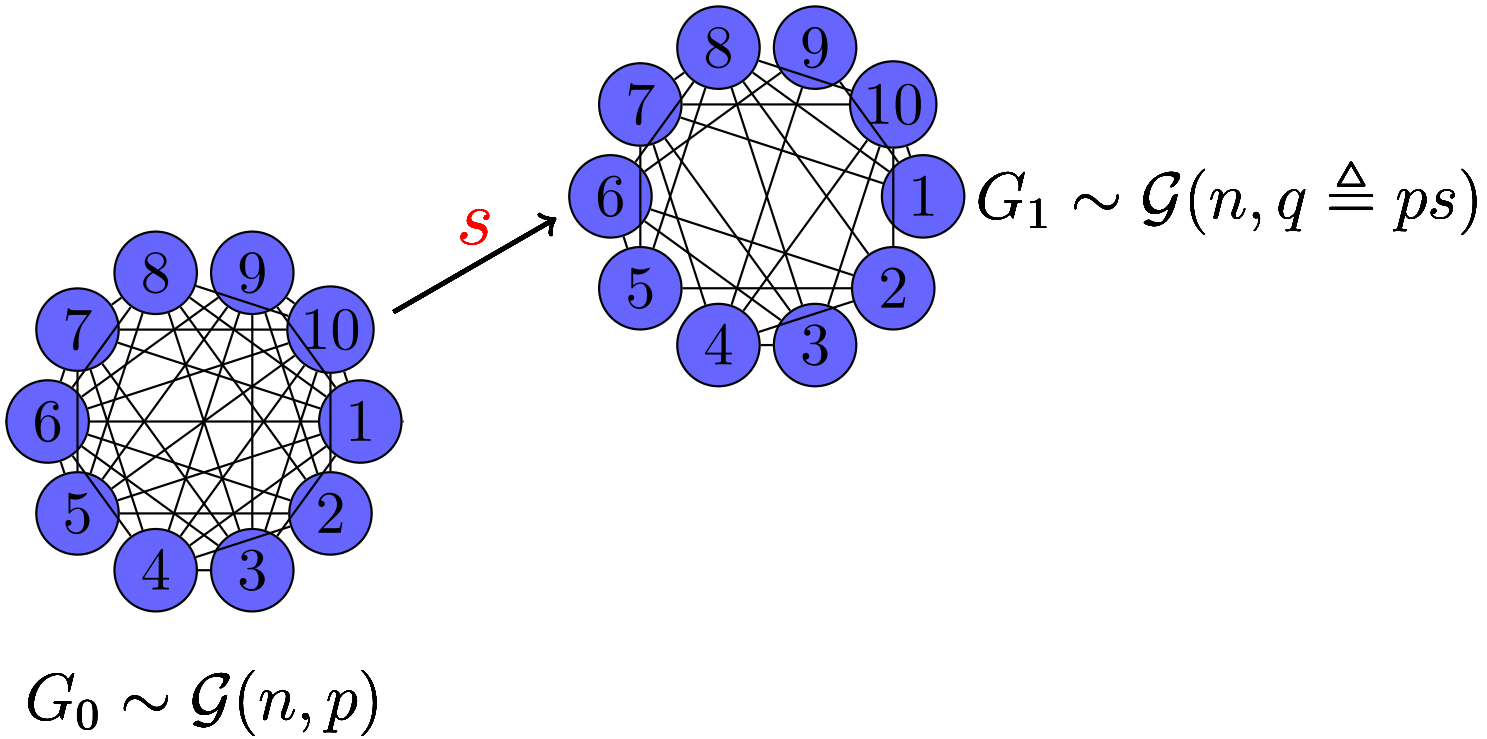
Statistical models for graph matching: graphs are randomly generated

Correlated Erdős-Rényi graphs model [Pedarsani-Grossglauser '11]

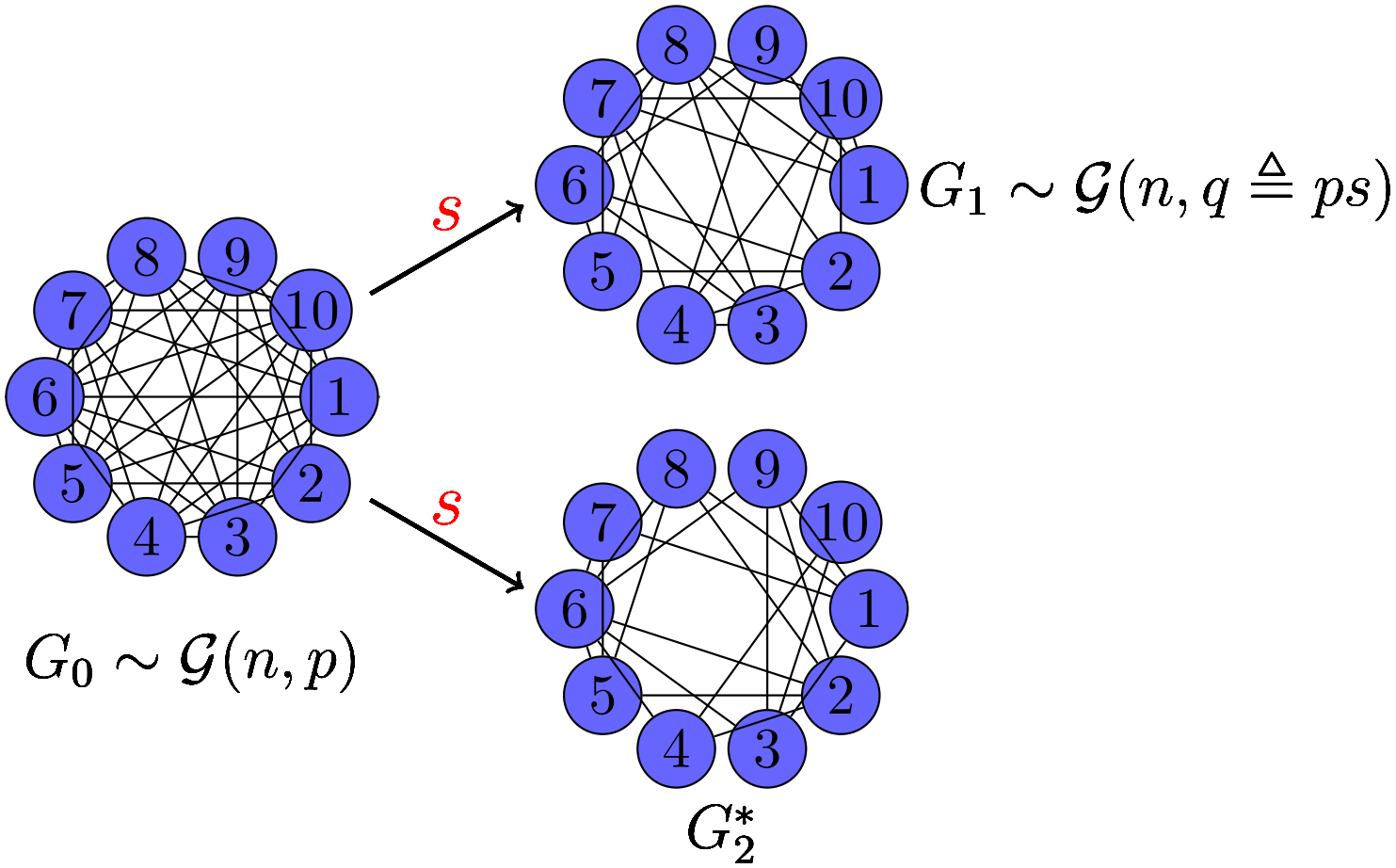


$G_0 \sim \mathcal{G}(n,p)$

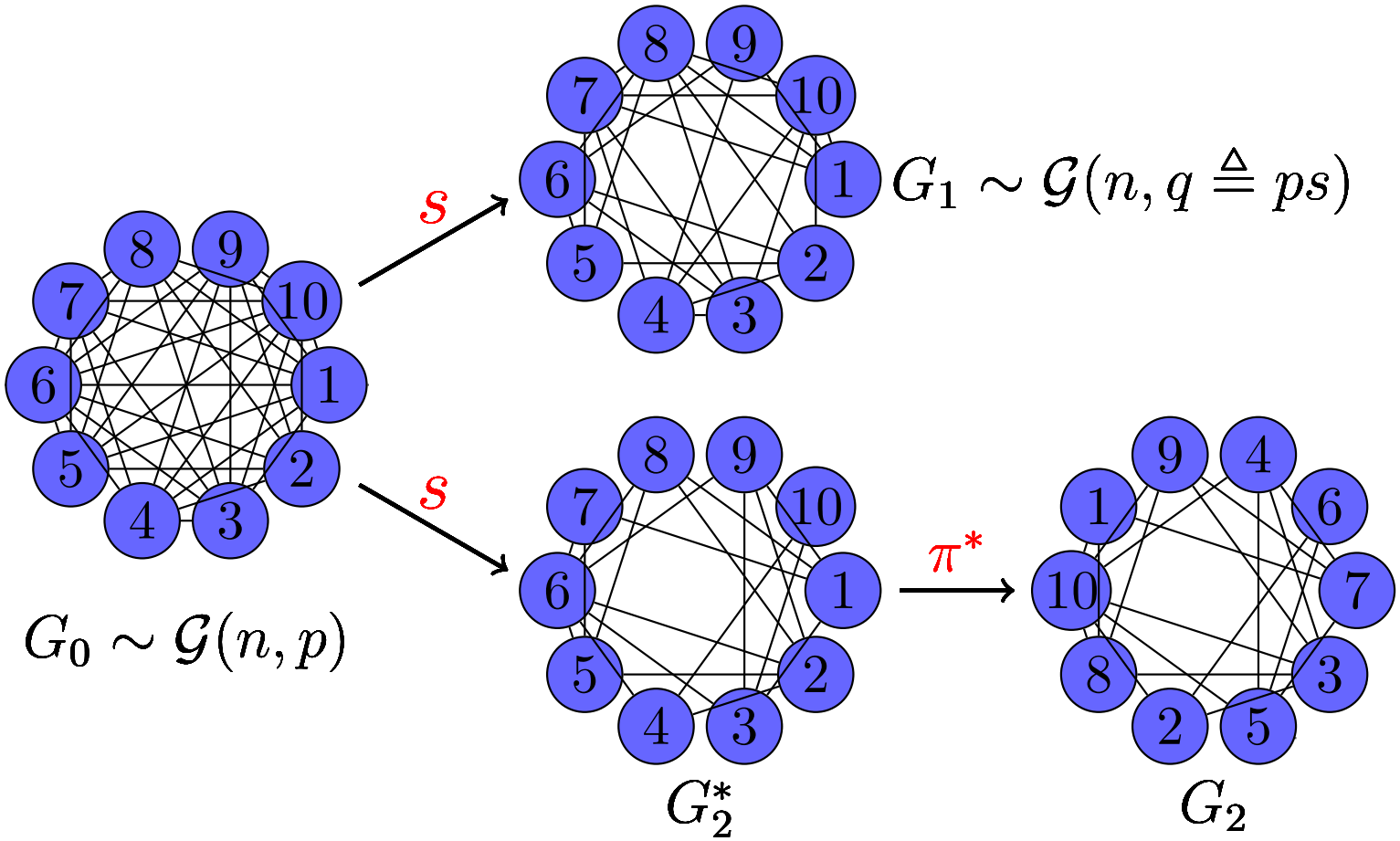
Correlated Erdős-Rényi graphs model [Pedarsani-Grossglauser '11]



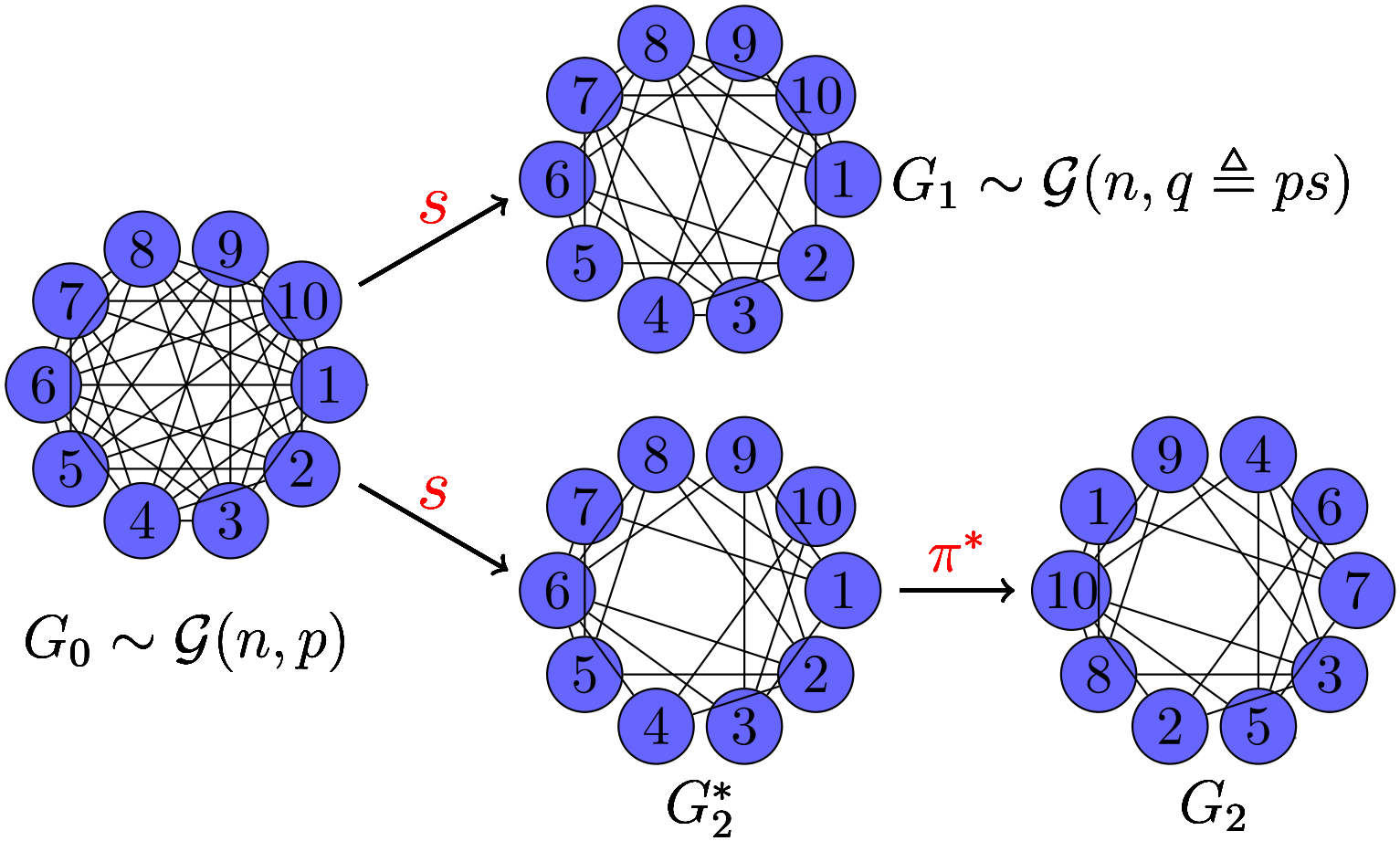
Correlated Erdős-Rényi graphs model [Pedarsani-Grossglauser '11]



Correlated Erdős-Rényi graphs model [Pedarsani-Grossglauser '11]



Correlated Erdős-Rényi graphs model [Pedarsani-Grossglauser '11]



- Erdős-Rényi graphs: canonical model for social/economic networks
- G_1 and G_2 differ by a fraction $\delta \triangleq 1 - s$ of edges, under the correct node mapping

Prior results: Information-theoretic limits

q: edge probability $\qquad$ $\delta = 1 - s$: fraction of errors (differed edges)

Theorem (Cullina-Kiyavash '18)

For $q < 1/2$, exact recovery of π^* is information-theoretically possible if and only if

$$nqs - \log n \to +\infty$$

Interpretation: Intersection graph $G_1 \wedge G_2^* \sim \mathcal{G}(n, qs)$ is connected

Prior results: Information-theoretic limits

q: edge probability $\quad\quad$ $\delta = 1 - s$: fraction of errors (differed edges)

Theorem (Cullina-Kiyavash '18)

For $q < 1/2$, exact recovery of π^* is information-theoretically possible if and only if

$$nqs - \log n \to +\infty$$

Interpretation: Intersection graph $G_1 \wedge G_2^* \sim \mathcal{G}(n, qs)$ is connected

Computationally:

- Noiseless $s = 1(\delta = 0)$: optimal condition is attained in linear-time [Bollobás '82, Czajka-Pandurangan '08]
- Noisy case $s < 1(\delta > 0)$: little is known for efficient algorithms

Prior results: Graph matching in noisy case

- $n^{\log n}$ time algorithm [Barak-Chou-Lei-Schramm-Sheng '18]:

  sparse and dense graphs and large noise tolerance δ

  Not efficient for large-scale networks: $10000^{\log 10000} \geq 10^{36}$

- $O(n^2)$-time algorithm [Dai-Cullina-Kiyavash-Grossglauser '18]

  only dense graphs and small noise tolerance δ

  Far from optimal: real networks are often sparse and noisy

Question

Efficiently match two graphs with large noise tolerance in both sparse and dense regimes?

Main results

q: edge probability $\qquad$ $\delta = 1 - s$: fraction of errors (differed edges)

Theorem (Ding-Ma-Wu-X. '18)

Exact recovery is attainable in $O(n(nq)^2 + n^2)$ time if

$$nq \gtrsim (\log n)^2 \quad \textit{and} \quad \delta \lesssim (\log n)^{-2}$$

Main results

q: edge probability $\qquad$ $\delta = 1 - s$: fraction of errors (differed edges)

Theorem (Ding-Ma-Wu-X. '18)

Exact recovery is attainable in $O(n(nq)^2 + n^2)$ time if

$$nq \gtrsim (\log n)^2 \quad \text{and} \quad \delta \lesssim (\log n)^{-2}$$

Further improvement:

- $\delta \lesssim (\log n)^{-2/3}$ *for dense graphs*
- $\delta \lesssim (\log(nq))^{-2}$ *for sparse graphs*

Main results

q: edge probability $\qquad$ $\delta = 1 - s$: fraction of errors (differed edges)

Theorem (Ding-Ma-Wu-X. '18)

Exact recovery is attainable in $O(n(nq)^2 + n^2)$ time if

$$nq \gtrsim (\log n)^2 \quad \text{and} \quad \delta \lesssim (\log n)^{-2}$$

Further improvement:

- $\delta \lesssim (\log n)^{-2/3}$ *for dense graphs*
- $\delta \lesssim (\log(nq))^{-2}$ *for sparse graphs*

Significantly improves over [Dai-Cullina-Kiyavash-Grossglauser '18]:

$$nq \gg n^{4/5} \quad \text{and} \quad \delta \ll q^4$$

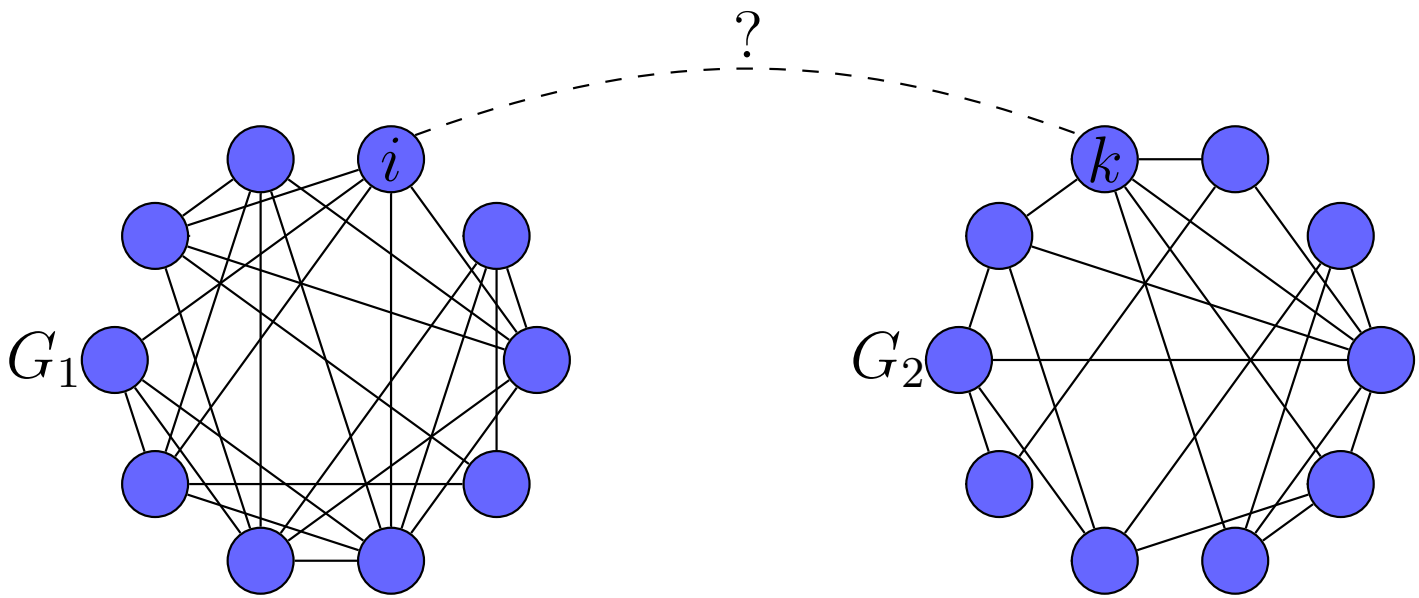
Example: $n = 10^4$

- Our result: $q \approx 0.008$ and $\delta \approx 0.11$
- [Dai-Cullina-Kiyavash-Grossglauser '18]: $q \approx 0.16$ and $\delta \approx 6 \times 10^{-4}$

Outline

1. Degree profile matching algorithm
2. Intuition behind our algorithm
3. Experimental results
4. Concluding remarks

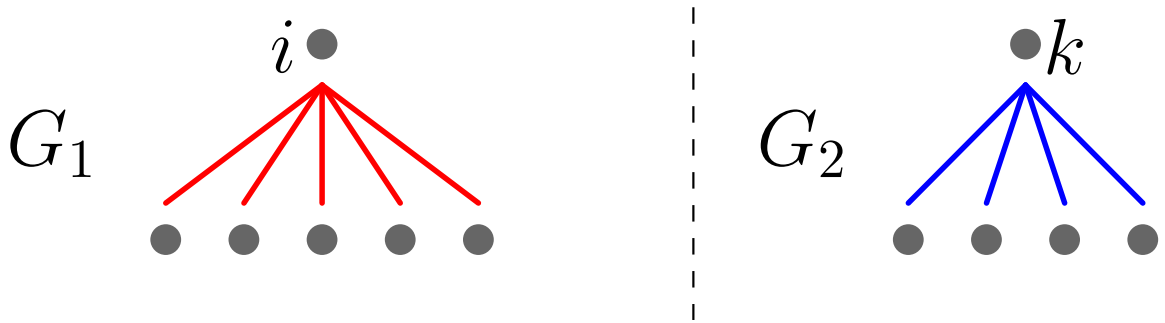
Many matching algorithms are signature-based



Abstractly:

1. attach some signature μ_i to node i in G_1
2. attach some signature ν_k to node k in G_2
3. match pairs based on distance between their signatures

A natural idea: Degree matching



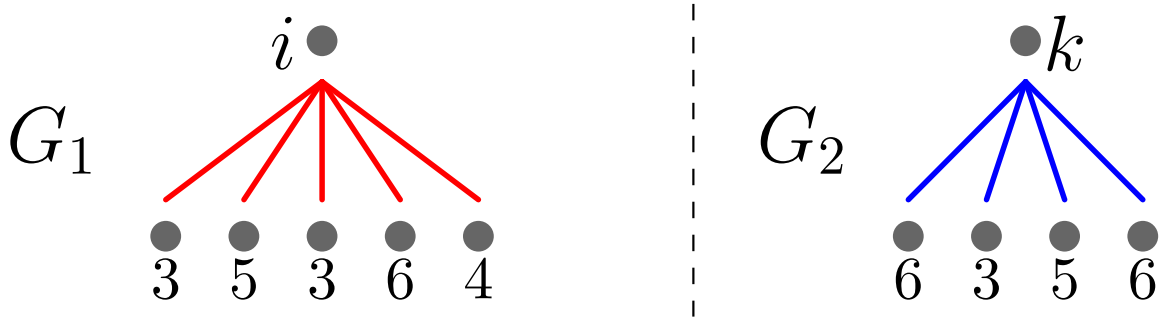
Sort the degrees (# of neighbors):

$\mu_i =$ degree of node i in G_1, $\quad \nu_k =$ degree of node k in G_2

- Highly sensitive to noise
- Graphs need to be dense
- Reason: small or no spacing between ordered degrees

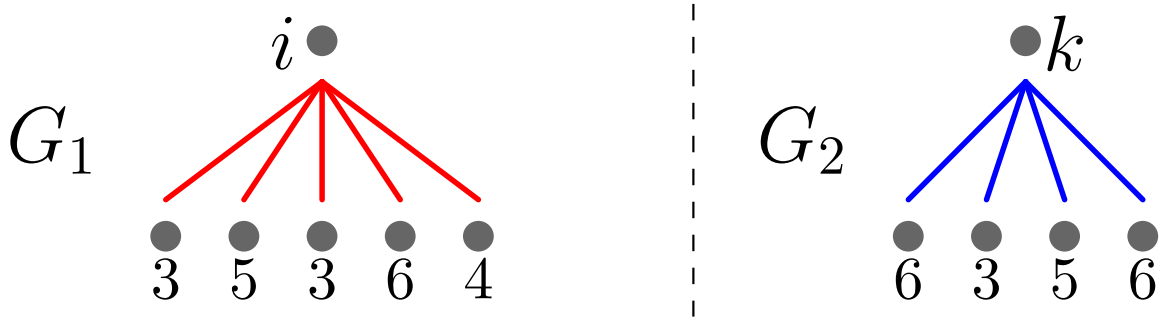
Our idea: Degree profile matching

- Need more robust signature
- Absent labels, can only compute permutation-invariant signature



Our idea: Degree profile matching

- Need more robust signature
- Absent labels, can only compute permutation-invariant signature



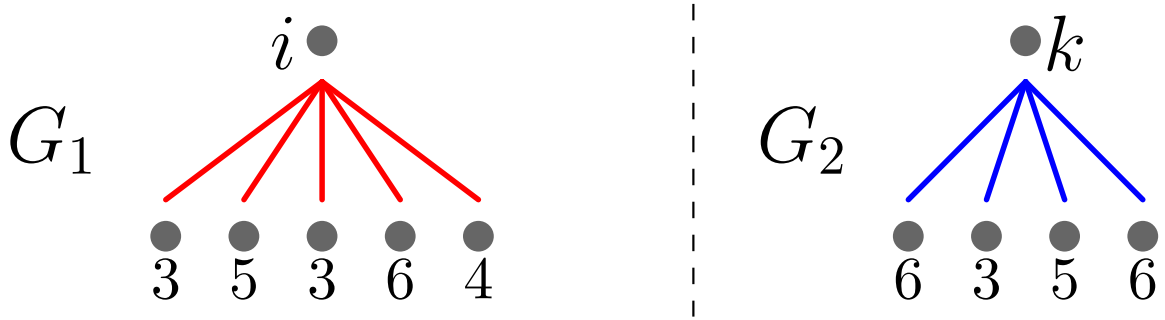
Degree profile: empirical distribution of the degrees of neighbors

$\mu_i \triangleq$ empirical distr. of the degrees of neighbors of i in G_1

$\nu_k \triangleq$ empirical distr. of the degrees of neighbors of k in G_2

Our idea: Degree profile matching

- Need more robust signature
- Absent labels, can only compute permutation-invariant signature



Degree profile: empirical distribution of the degrees of neighbors

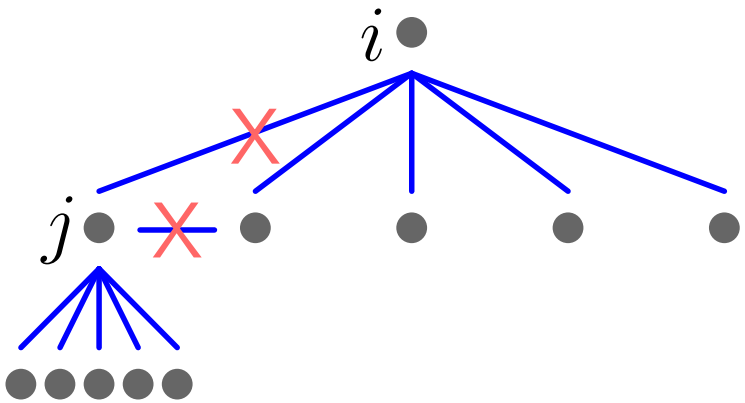
$\mu_i \triangleq$ empirical distr. of the degrees of neighbors of i in G_1

$\nu_k \triangleq$ empirical distr. of the degrees of neighbors of k in G_2

- True pairs: μ_i is "close" to ν_k
- Fake pairs: μ_i is "far" from ν_k
- Difficult to analyze because of dependency

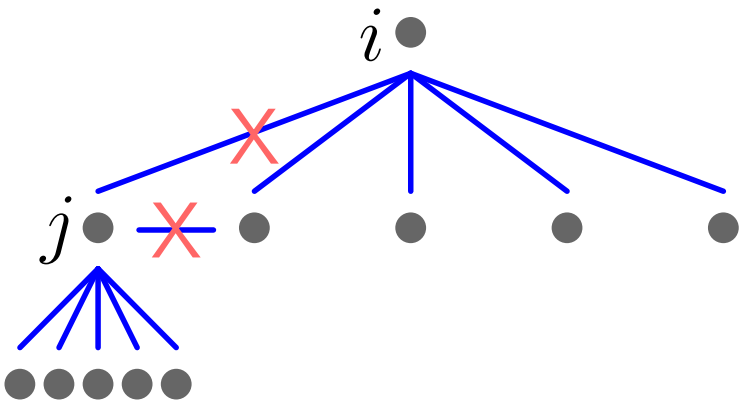
Standardized outdegree profile

- Standardized outdegree: exclude i's closed neighborhood in G_1



Standardized outdegree profile

- Standardized outdegree: exclude i's closed neighborhood in G_1



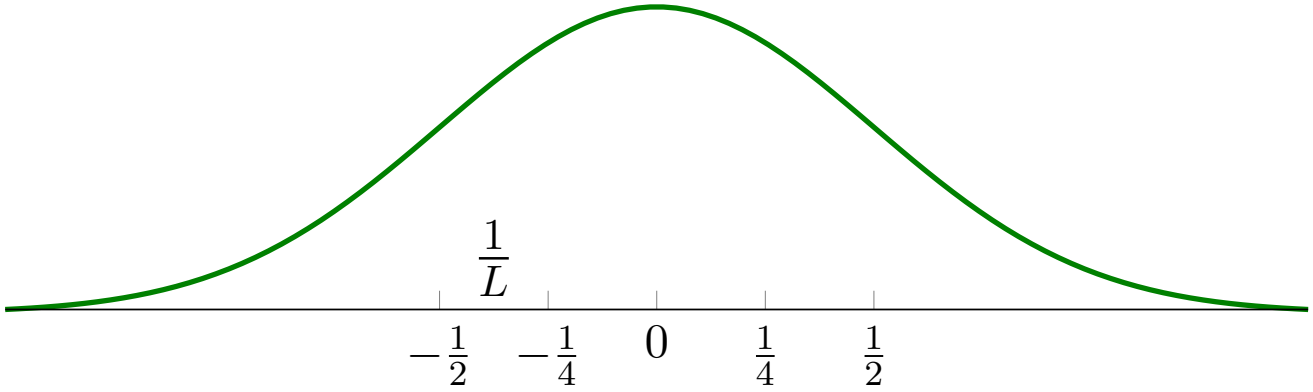
- Empirical distribution of standardized outdegrees:

$$\bar{\mu}_i \triangleq \text{empirical distr. of outdegrees of neighbors of } i \text{ in } G_1$$

$$\bar{\nu}_k \triangleq \text{empirical distr. of outdegrees of neighbors of } k \text{ in } G_2$$

Binning and L_1 distance

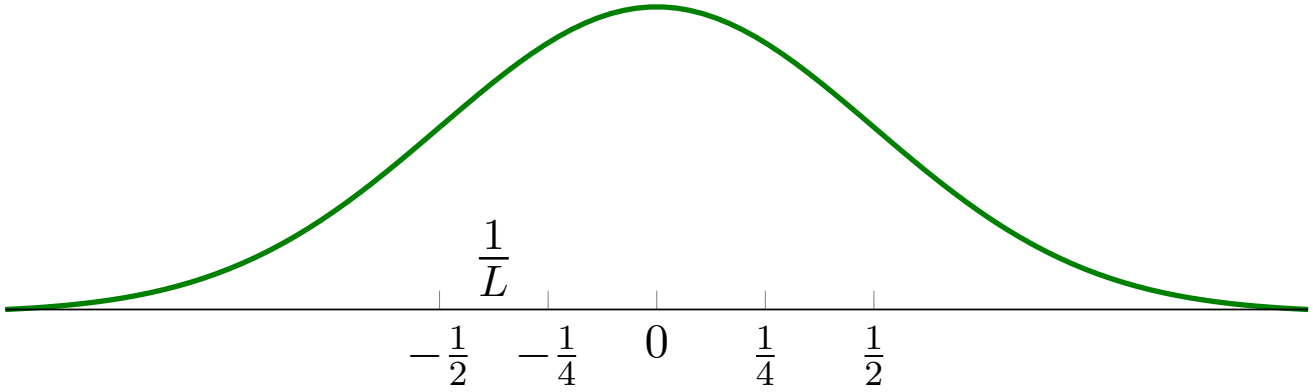
- Empirical distr. of standardized outdegrees of neighbors $\approx N(0,1)$



- Divide $[-1/2, 1/2]$ into L sub-intervals (bins) I_ℓ of length $1/L$

Binning and L_1 distance

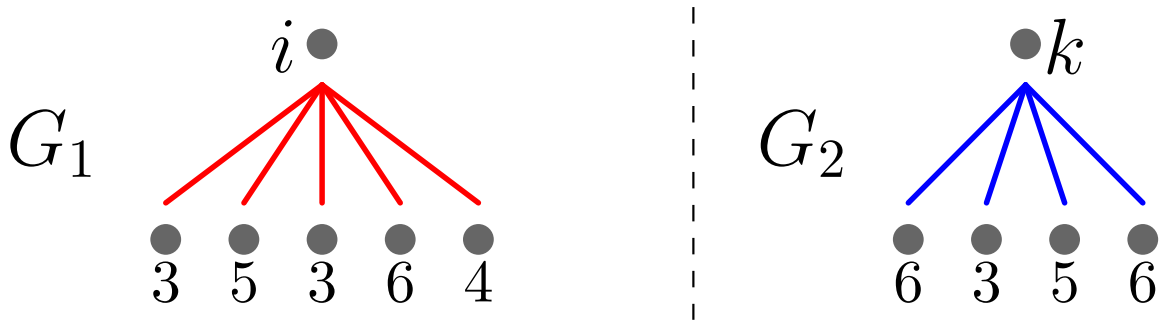
- Empirical distr. of standardized outdegrees of neighbors $\approx N(0,1)$



- Divide $[-1/2, 1/2]$ into L sub-intervals (bins) I_ℓ of length $1/L$
- L_1 distance between discretized degree profiles:

$$Z_{ik} \triangleq \|\bar{\mu}_i - \bar{\nu}_k\| \triangleq \sum_{\ell=1}^{L} |\bar{\mu}_i(I_\ell) - \bar{\nu}_k(I_\ell)|$$

Greedy matching algorithm with degree profiles



- Match i to k whose degree profile is the closest, i.e., minimizing Z_{ik}
- Proof of correctness (assuming $\pi^* = \mathrm{id}$):
  - "True pair": $i = k$
  - "Fake pair": $i \neq k$
  - Show separation:

$$\underbrace{\max_i Z_{ii}}_{\text{distance for true pairs}} < \underbrace{\min_{i \neq k} Z_{ik}}_{\text{distance for fake pairs}}$$

holds with high probability

Outline

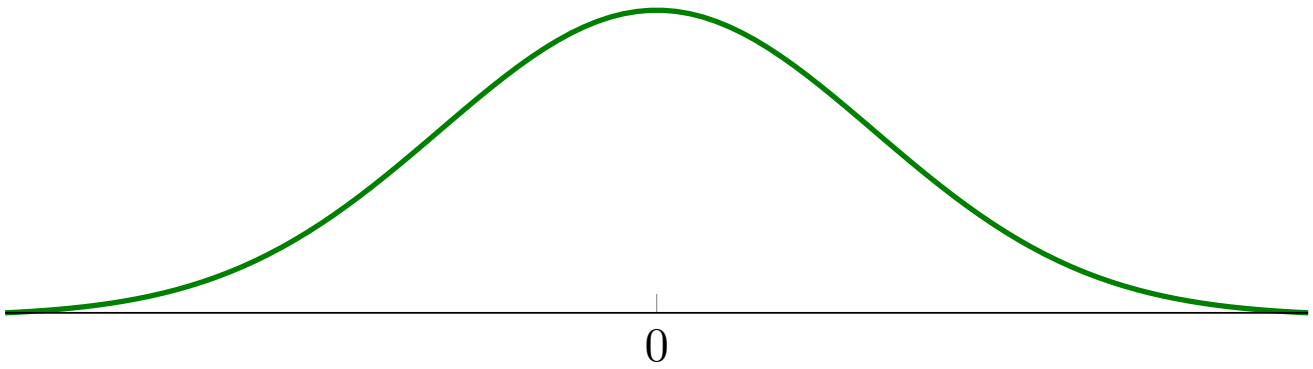
1. Degree profile matching algorithm
2. Intuition behind our algorithm
3. Experimental results
4. Concluding remarks

Why can we distinguish true pairs from fake pairs?

For any pair i, k:

- By Central Limit Theorem, marginally

$$\bar{\pi}_i \approx \bar{\mu}_k \approx N(0,1)$$



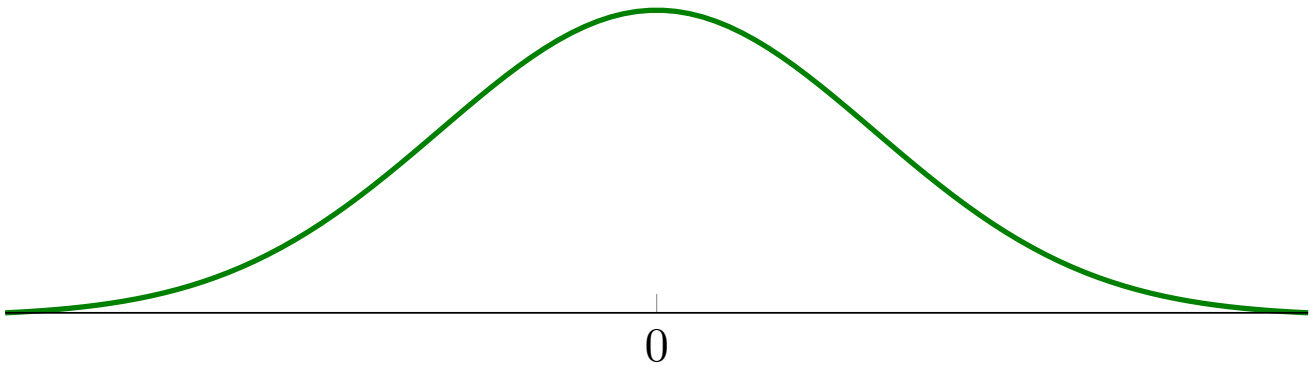
- Their corresponding empirical CDFs $\approx \Phi$ (CDF of $N(0,1)$)

Why can we distinguish true pairs from fake pairs?

For any pair i, k:

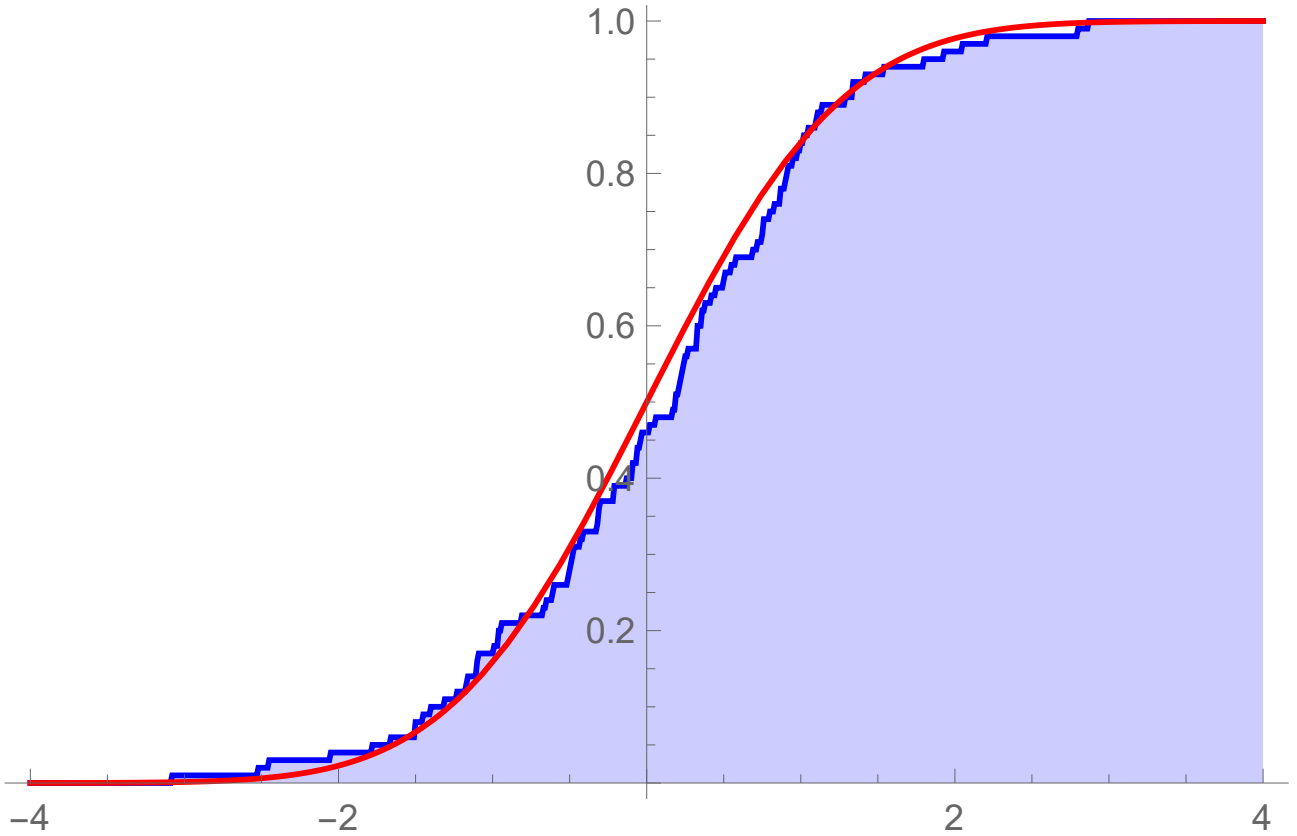
- By Central Limit Theorem, marginally

$$\bar{\pi}_i \approx \bar{\mu}_k \approx N(0,1)$$



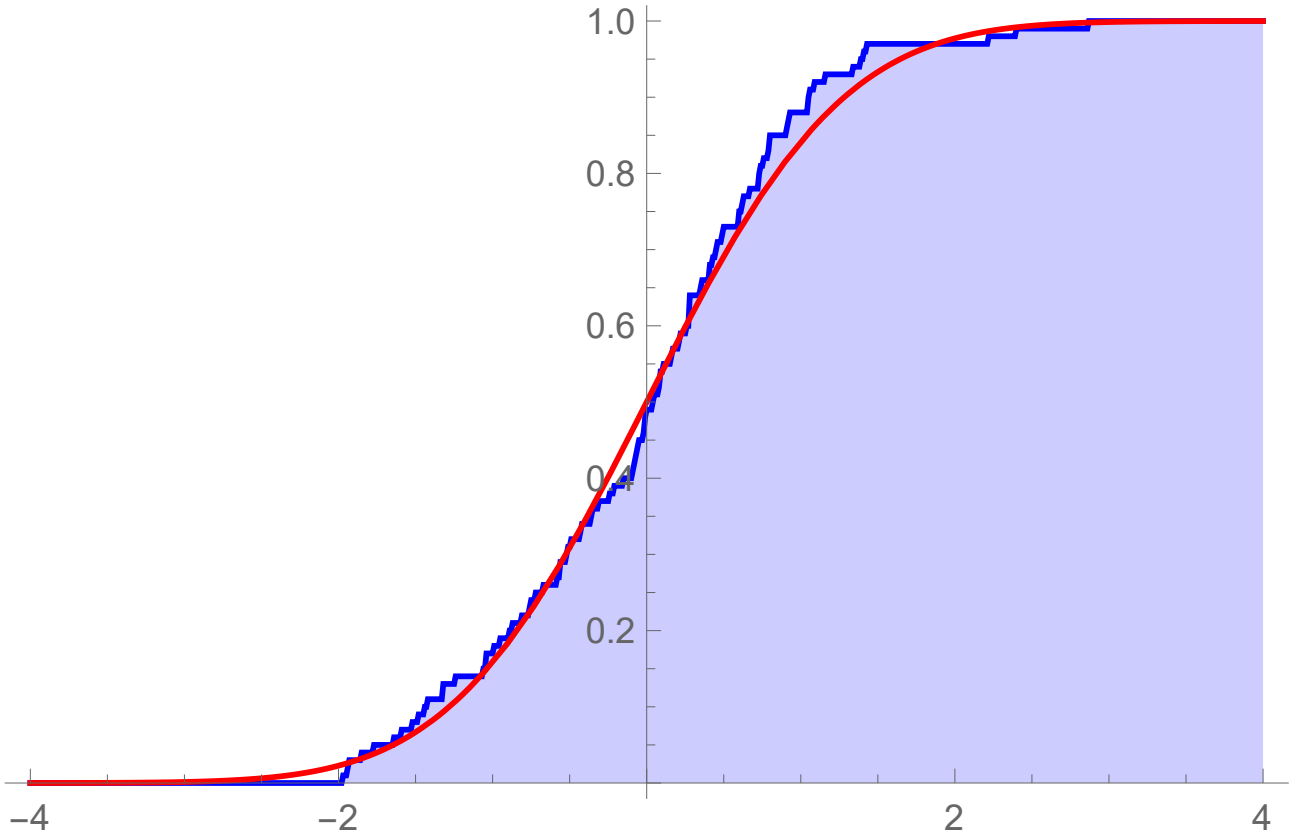
- Their corresponding empirical CDFs $\approx \Phi$ (CDF of $N(0,1)$)
- The signal is in the fluctuation

Natural fluctuation of empirical CDF



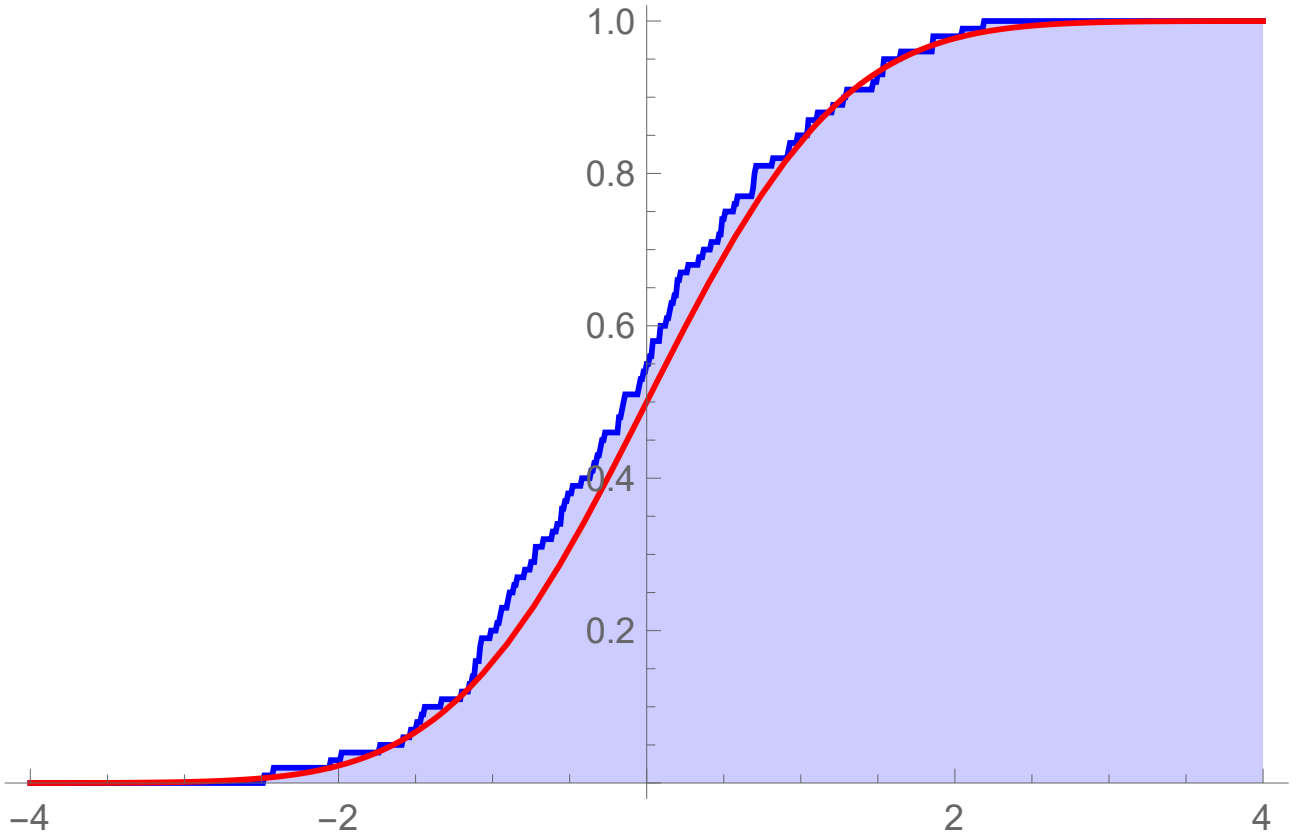
Blue: 100 standard normal samples; Red: normal CDF Φ

Natural fluctuation of empirical CDF



Blue: 100 standard normal samples; Red: normal CDF Φ

Natural fluctuation of empirical CDF



Blue: 100 standard normal samples; Red: normal CDF Φ

Separating true pairs and fake pairs

Informally (assuming $\pi^* = \mathrm{id}$):

- For "fake pair" $i \neq k$:

  outdegrees of their neigbhors are mostly independent

  $\Longrightarrow$ fluctuations of their degree profiles are independent

  $\Longrightarrow$ distance $Z_{ik} \triangleq \|\bar{\mu}_i - \bar{\nu}_k\|$ is large

Separating true pairs and fake pairs

Informally (assuming $\pi^* = \mathrm{id}$):

- For "fake pair" $i \neq k$:

  outdegrees of their neigbhors are mostly independent

  $\Longrightarrow$ fluctuations of their degree profiles are independent

  $\Longrightarrow$ distance $Z_{ik} \triangleq \|\bar{\mu}_i - \bar{\nu}_k\|$ is large

- For "true pair" $i = k$:

  outdegrees of their neigbhors are correlated

  $\Longrightarrow$ fluctuations of their degree profiles are positively correlated and cancel each other

  $\Longrightarrow$ distance $Z_{ii} \triangleq \|\bar{\mu}_i - \bar{\nu}_i\|$ is small

Separation of distances

δ: fraction of errors L: # of bins q: edge probability

Lemma (Ding-Ma-Wu-X. '18)

If $\delta \ll 1/L^2$ *and* $nq \gg \max\{\log n, L^2\}$, *then with probability at least* $1 - e^{-\Omega(L)}$:

$$Z_{ik} \begin{cases} \leq c\sqrt{\frac{L}{nq}} & \text{if } i \text{ and } k \text{ are true pairs} \\ \geq C\sqrt{\frac{L}{nq}} & \text{if } i \text{ and } k \text{ are fake pairs} \end{cases}$$

Separation of distances

δ: fraction of errors L: # of bins q: edge probability

Lemma (Ding-Ma-Wu-X. '18)

If $\delta \ll 1/L^2$ and $nq \gg \max\{\log n, L^2\}$, then with probability at least $1 - e^{-\Omega(L)}$:

$$Z_{ik} \begin{cases} \le c\sqrt{\frac{L}{nq}} & \text{if } i \text{ and } k \text{ are true pairs} \\ \ge C\sqrt{\frac{L}{nq}} & \text{if } i \text{ and } k \text{ are fake pairs} \end{cases}$$

Remark:

- If $\delta \ll 1/\log^2 n$: choose $L = \log n$ and union bound $\Rightarrow$

$$\max_{\text{true pairs } i,k} Z_{ik} < \min_{\text{fake pairs } i,k} Z_{ik}$$

- Technical difficulty: lots of dependency, e.g., degree profiles of fake pairs are also correlated through their common neighbors

Outline

1. Degree profile matching algorithm
2. Intuition behind our algorithm
3. Experimental results
4. Concluding remarks

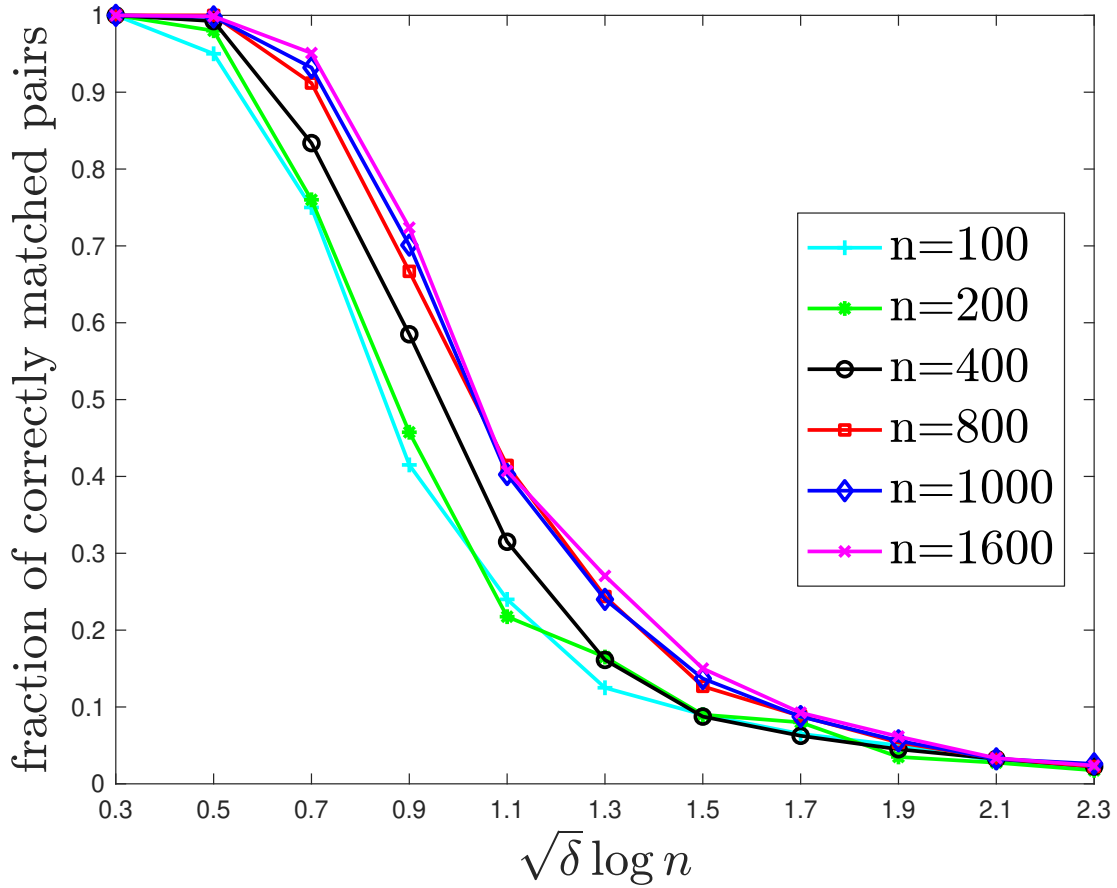
Algorithms

Three algorithms:

- Degree profile (DP): ~~out~~degree, Wasserstein W_1 distance
- Spectral method (SP): align leading eigenvector u of G_1 and v of G_2
- Quadratic programming relaxation (QP):
  permutation matrix $\rightarrow$ doubly stochastic matrix

Erdős-Rényi graphs: Degree profile matching

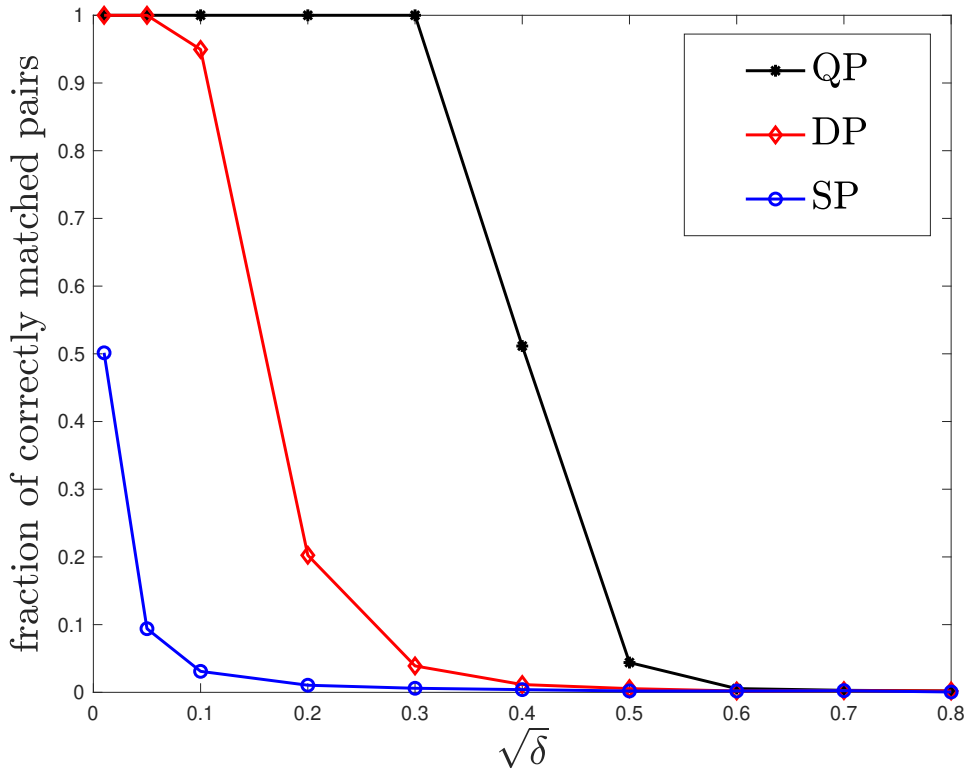
Average degree $nq = \log^2 n$



Consistent with the theory: $\delta = 1 - s = \Theta(\frac{1}{\log^2 n})$

Erdős-Rényi graphs: DP vs SP vs QP

Erdős-Rényi graph model $n = 1000$, $p = \log^2(n)/n$, $s = 1 - \delta$



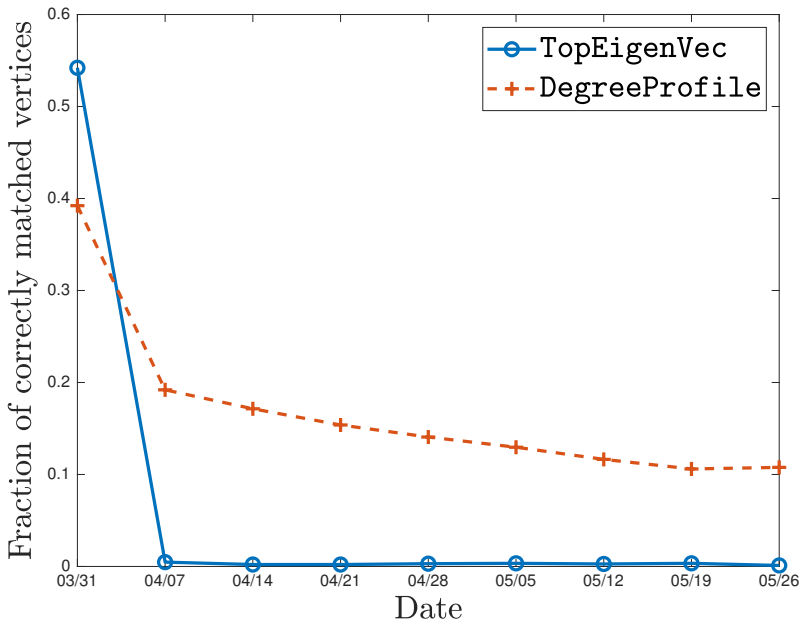
- Matching accuracy: QP $>$ DP $\gg$ SP
- Running time: QP $\gg$ DP $>$ SP

Real network: Oregon Internet router network

- An Internet router network observed on 9 days between March 2001 and May 2001 (10K nodes, 22K-23K edges)
- The network exhibits the addition and deletion of edges over time
- Goal: match 9 networks on 9 days to the network on day 1

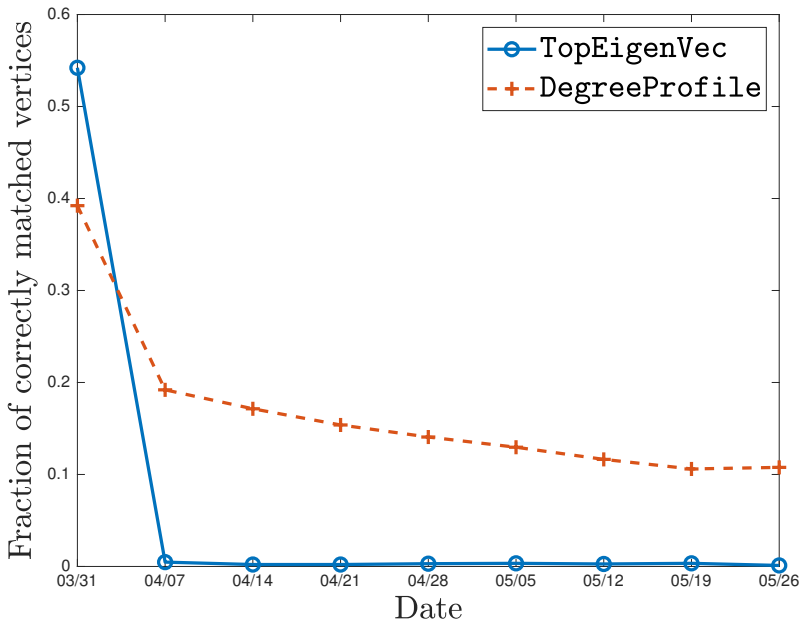
Real network: Oregon Internet router network

- An Internet router network observed on 9 days between March 2001 and May 2001 (10K nodes, 22K-23K edges)
- The network exhibits the addition and deletion of edges over time
- Goal: match 9 networks on 9 days to the network on day 1



Real network: Oregon Internet router network

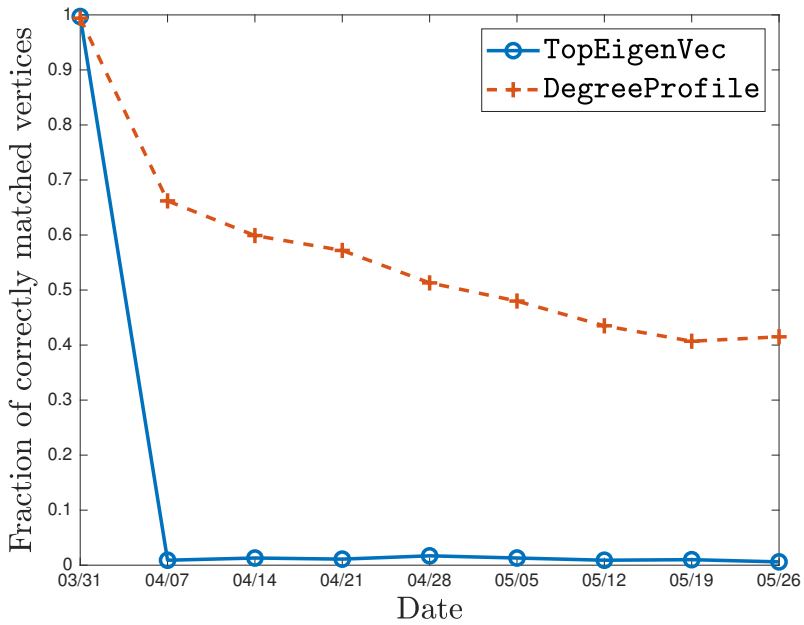
- An Internet router network observed on 9 days between March 2001 and May 2001 (10K nodes, 22K-23K edges)
- The network exhibits the addition and deletion of edges over time
- Goal: match 9 networks on 9 days to the network on day 1



- Performance degrades over time as network becomes less correlated
- DP outperforms SP except for matching the graph on day 1 to itself
- Matching is inexact due to low-degree nodes (3K nodes of degree 1)

Real network: Oregon Internet router network

Re-evaluate performance on subgraphs induced by 1K nodes of the largest degrees



- Both methods exactly match the subgraph from the first day to itself
- DP $\gg$ SP in matching high-degree nodes in general

Outline

1. Degree profile matching algorithm
2. Intuition behind our algorithm
3. Experimental results
4. Concluding remarks

Concluding remarks

- Develop an efficient graph matching algorithm via degree profiles
- Prove that it efficiently matches two graphs with large noise tolerance in both sparse and dense regimes
- Significantly improve the state of the art of efficient graph matching algorithms in terms of run time, noise tolerance, and graph sparsity
- The superiority is also demonstrated on synthetic and real datasets in terms of both statistical accuracy and computational efficency

Concluding remarks

- Develop an efficient graph matching algorithm via degree profiles
- Prove that it efficiently matches two graphs with large noise tolerance in both sparse and dense regimes
- Significantly improve the state of the art of efficient graph matching algorithms in terms of run time, noise tolerance, and graph sparsity
- The superiority is also demonstrated on synthetic and real datasets in terms of both statistical accuracy and computational efficency

References

- J. Ding, Z. Ma, Y. Wu, & J. X. *Efficient random graph matching via degree profiles*. arXiv:1811.07821.