Seeded Graph Matching via Large Neighborhood Statistics

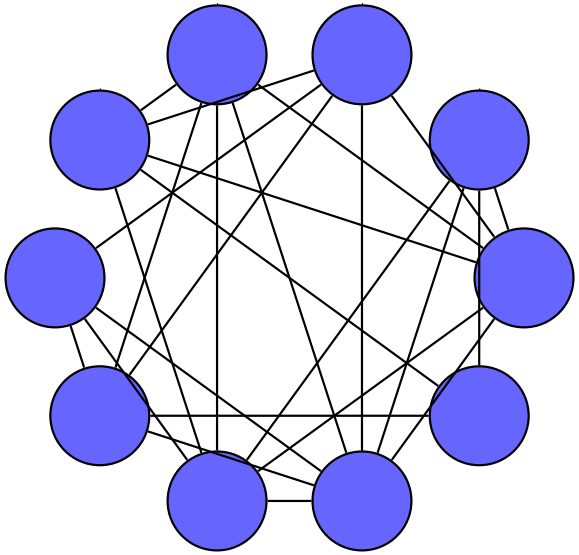
Jiaming Xu

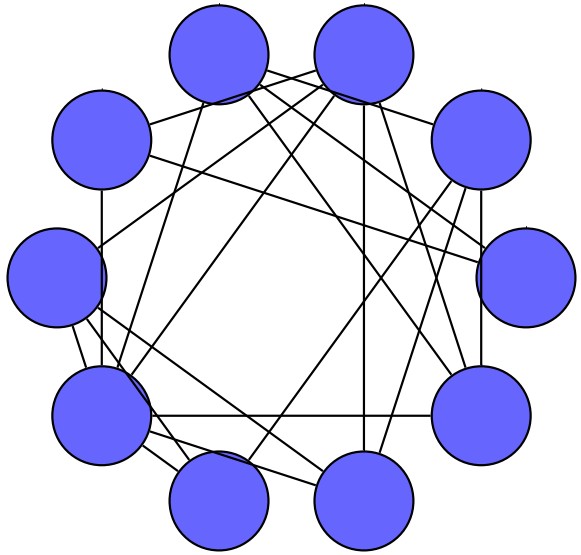
The Fuqua School of Business
Duke University

Joint work with Elchanan Mossel (MIT)

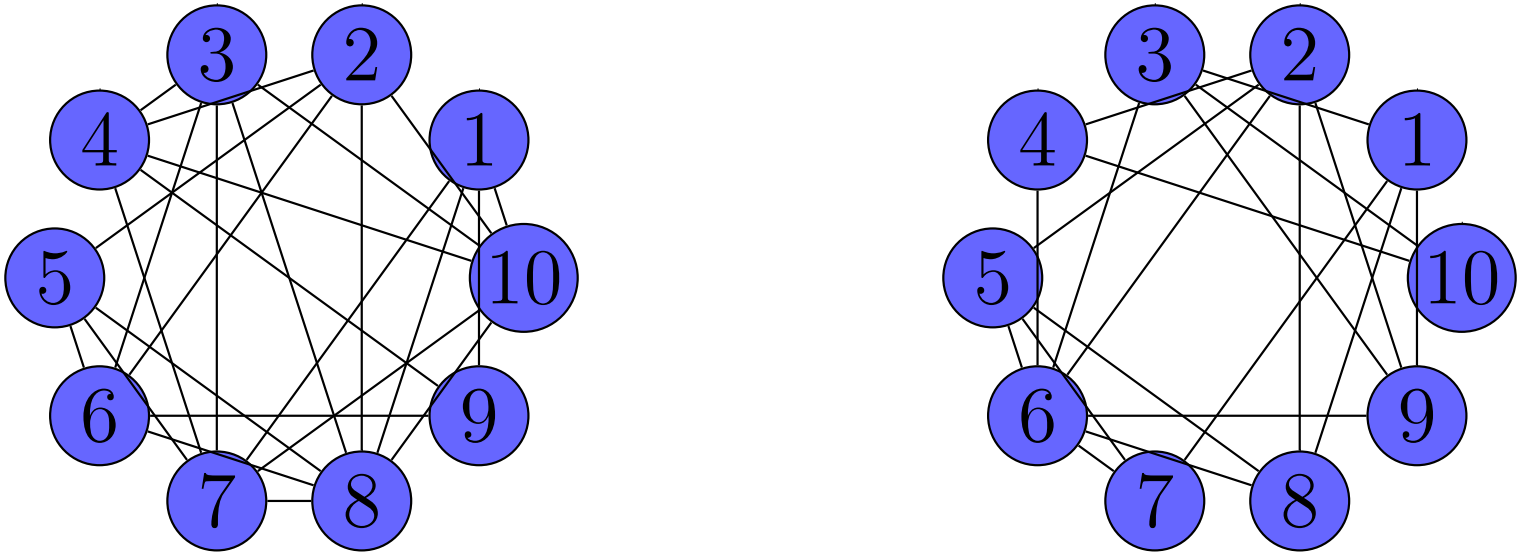
SODA, January 7, 2019

Graph matching (network alignment)



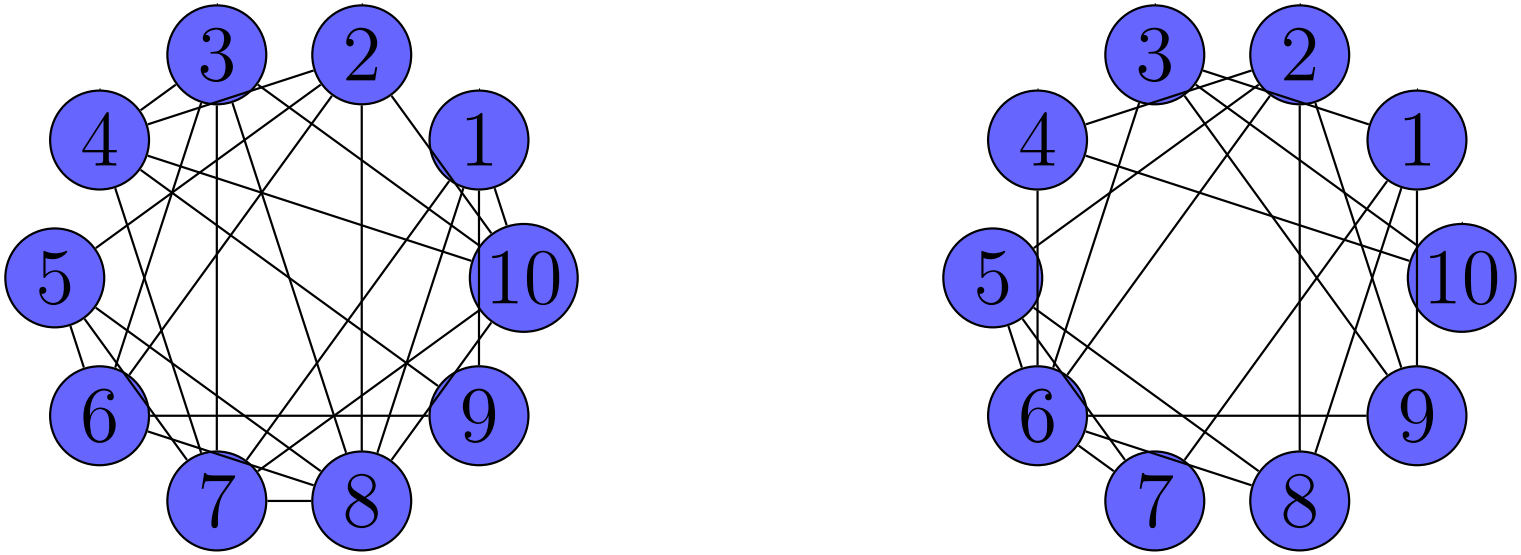


Graph matching (network alignment)



Goal: find a bijection between two vertex sets that minimizes # of adjacency disagreements

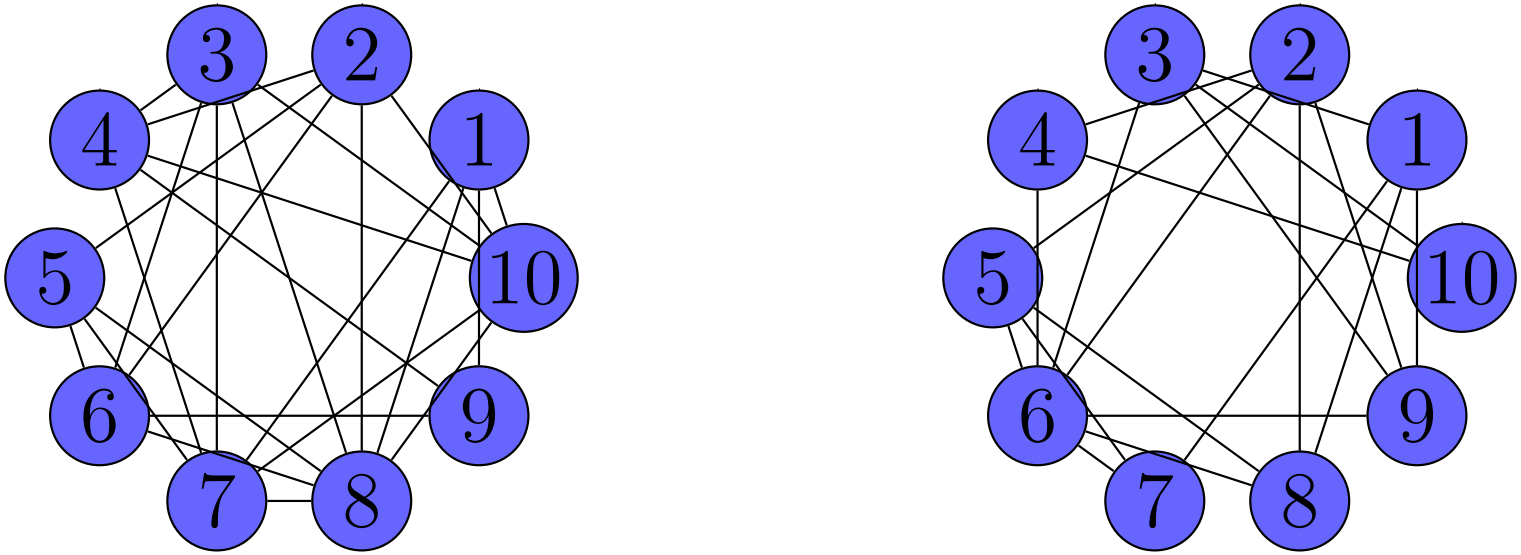
Graph matching (network alignment)



Goal: find a bijection between two vertex sets that minimizes # of adjacency disagreements

Quadratic assignment problem : $\min_{\Pi \in \mathcal{S}_n} \|A - \Pi B \Pi^\top\|_F^2$

Graph matching (network alignment)



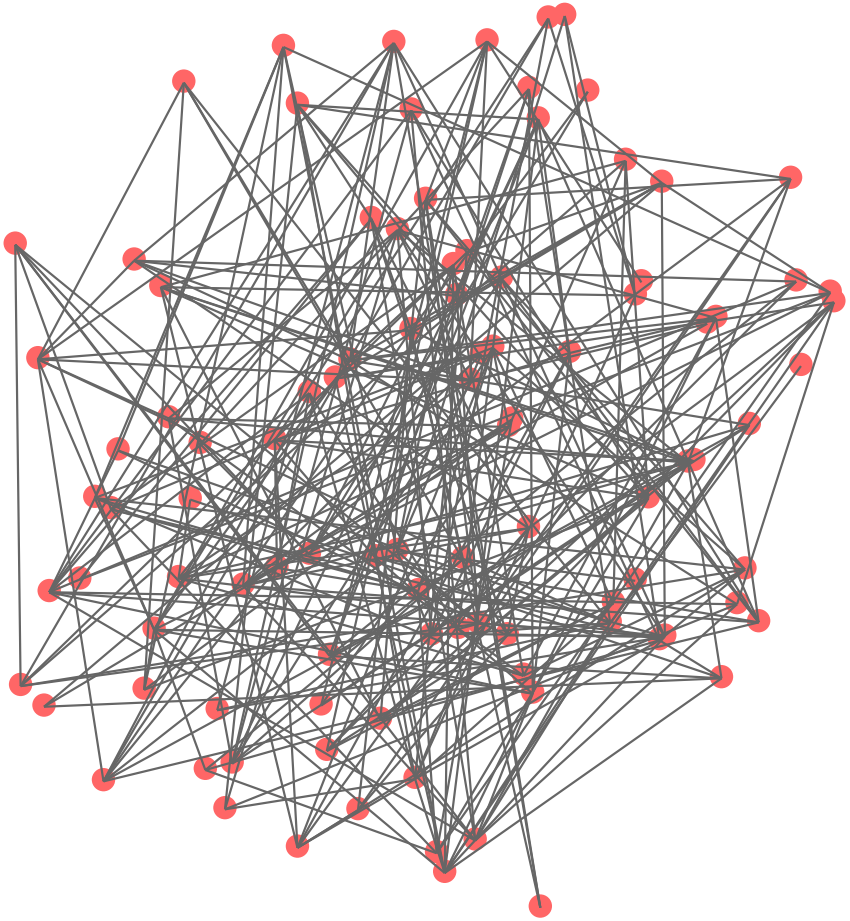
Goal: find a bijection between two vertex sets that minimizes # of adjacency disagreements

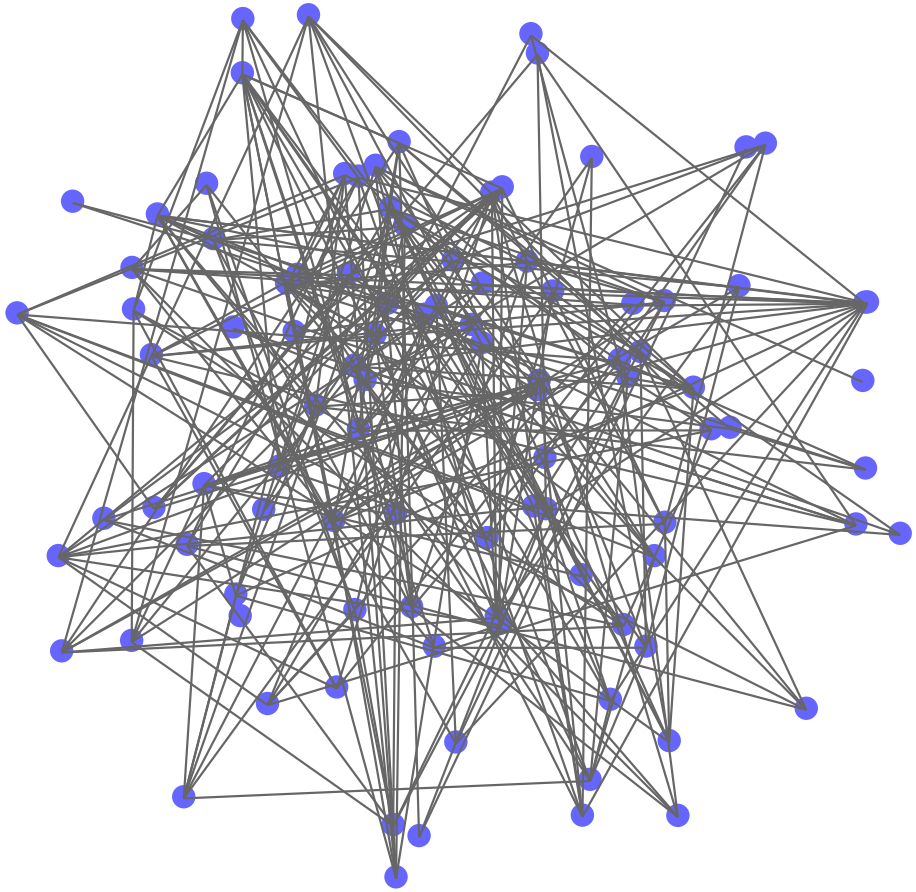
Quadratic assignment problem : $\min_{\Pi \in \mathcal{S}_n} \|A - \Pi B \Pi^\top\|_F^2$

Noiseless case: reduce to graph isomoprhism

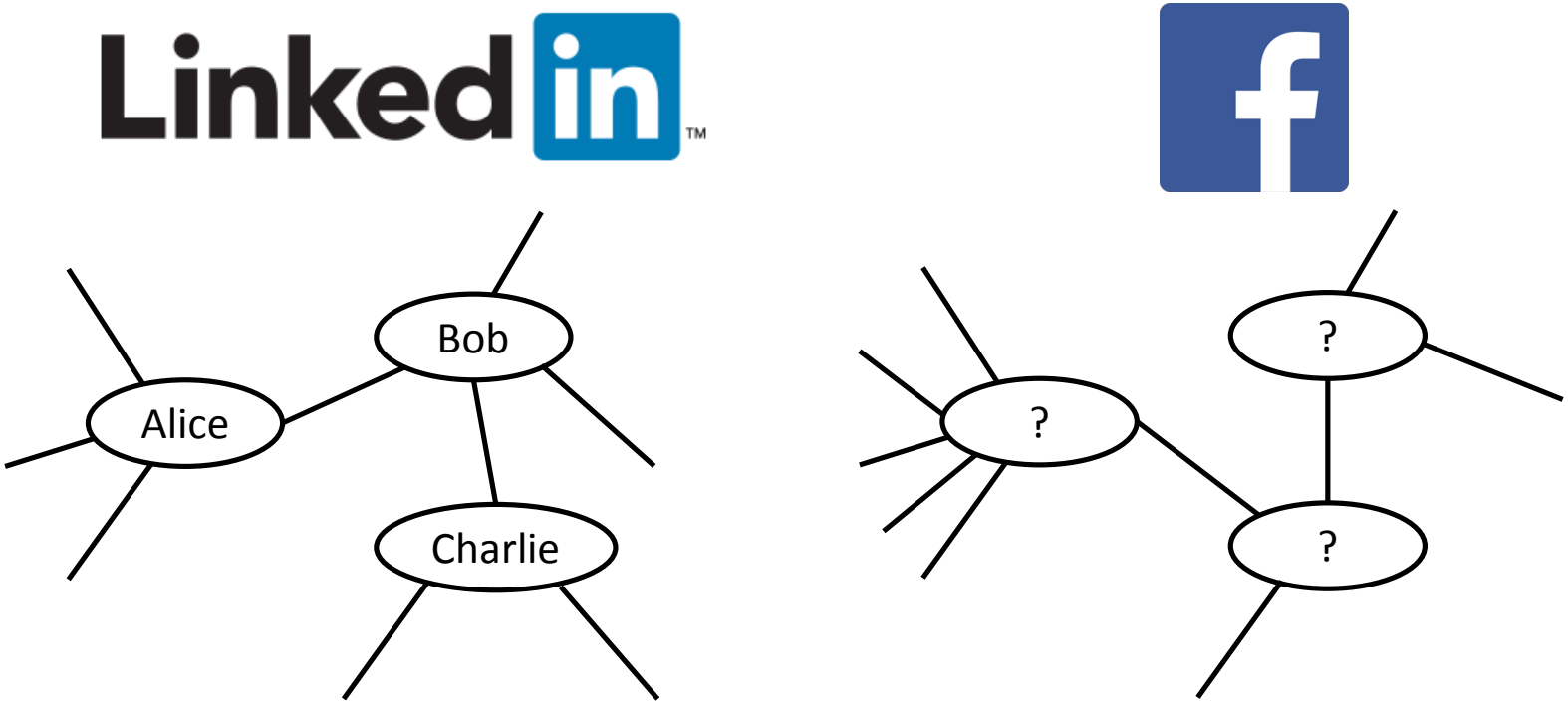
Two key challenges

- Statistical: two graphs are not exactly isomorphic
- Computational: # of possible node mappings is $n!$



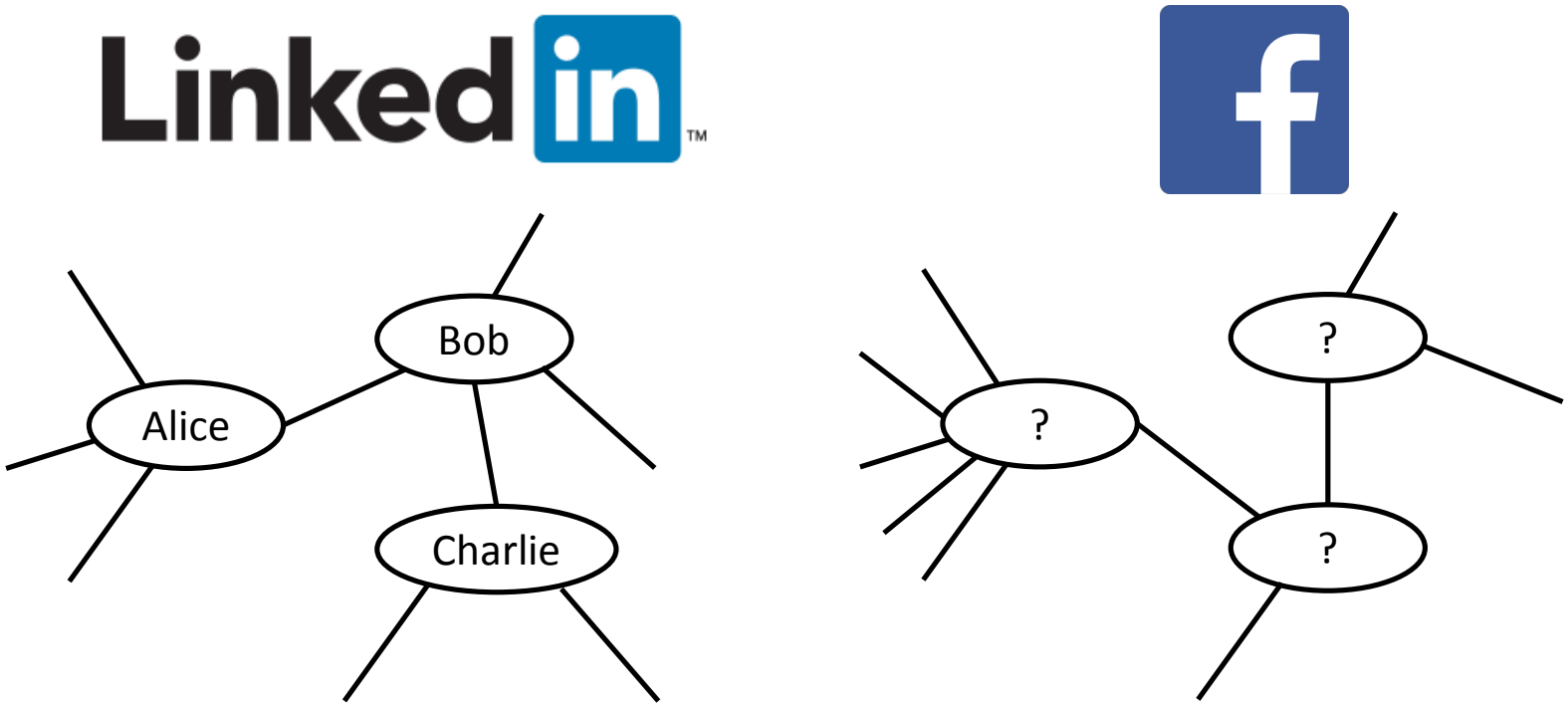


Application 1: Network de-anonymization



Picture courtesy of R. Srikant

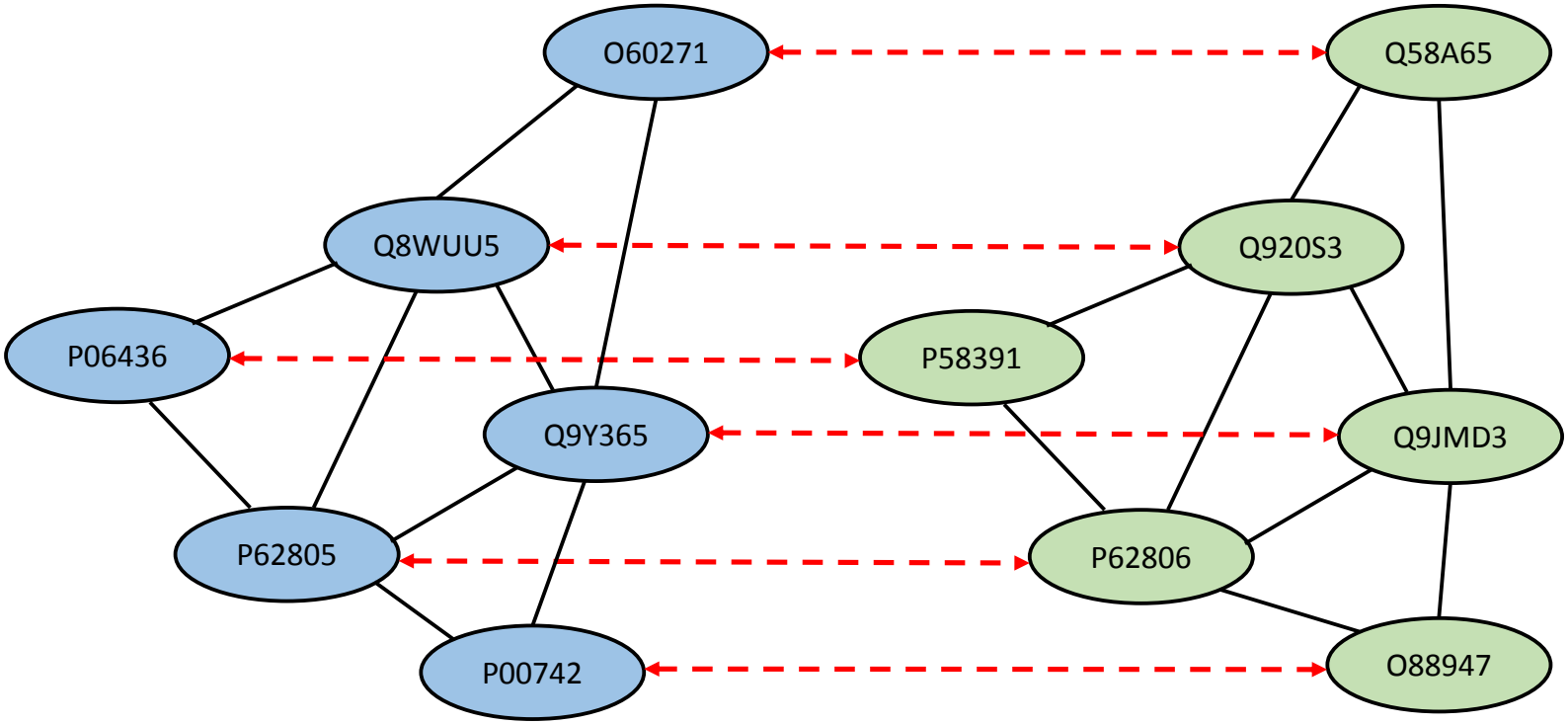
Application 1: Network de-anonymization



Picture courtesy of R. Srikant

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

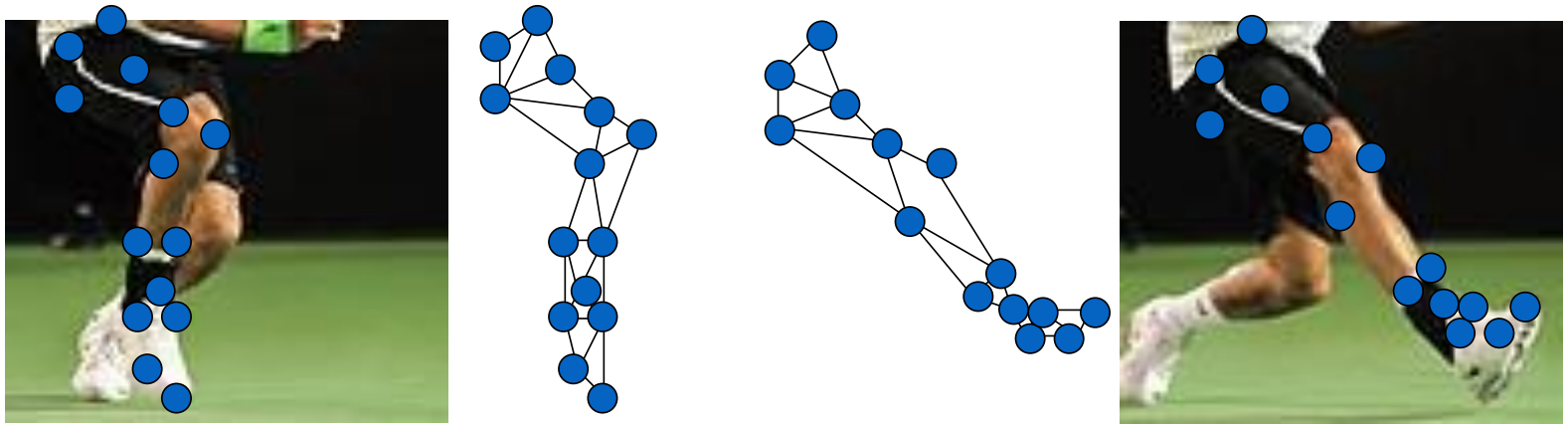
Application 2: Protein interaction network



[Kazemi-Hassani-Grossglauser-Modarres '16]. Picture courtesy of R. Srikant

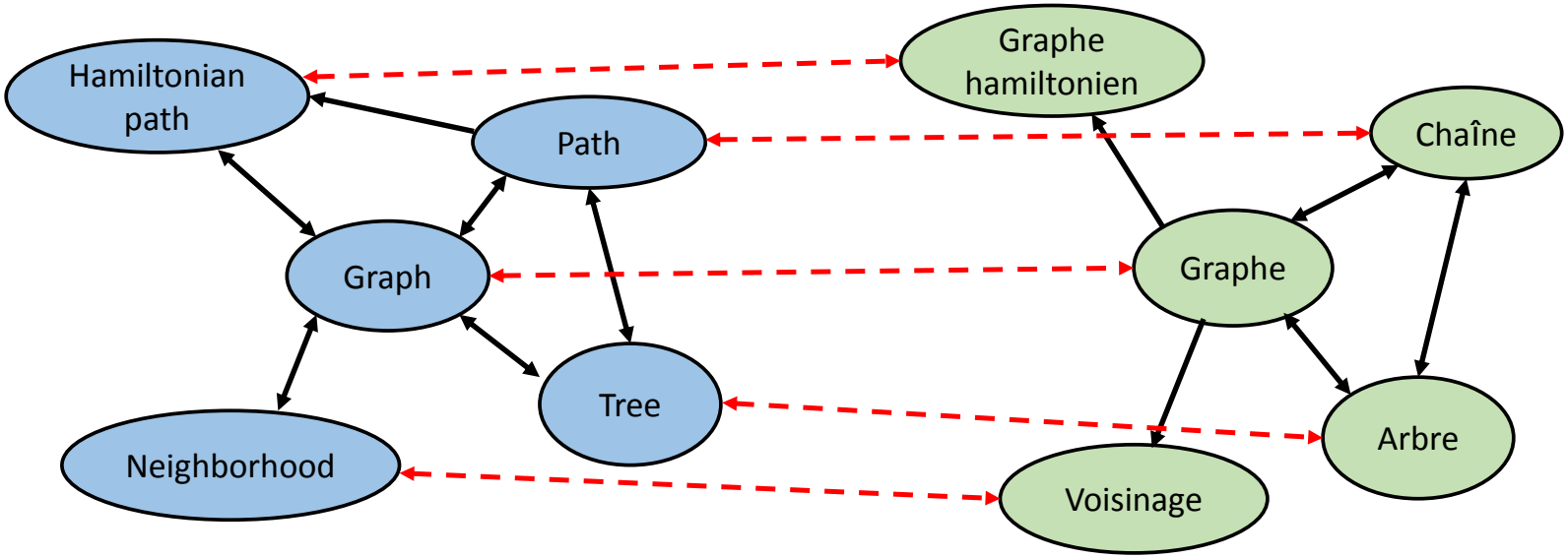
Ontology: Discover proteins with similar functions across different species based on interation network topology

Application 3: Computer Vision



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

Application 4: Machine Translation



Picture courtesy of R. Srikant

Automatically find/correct corresp. wiki articles in different languages

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

Beyond worst-case intractability of QAP

$$\text{QAP}: \quad \min_{\Pi \in \mathcal{S}_n} \|A - \Pi B \Pi^\top\|_F^2$$

- NP-hard in the worst case
- Even approximation within a factor $2^{\log^{1-\epsilon}(n)}$ for $\epsilon > 0$ is NP-hard [Makarychev-Manokaran-Sviridenko '10]
- However, real networks are not designed by adversary!

Beyond worst-case intractability of QAP

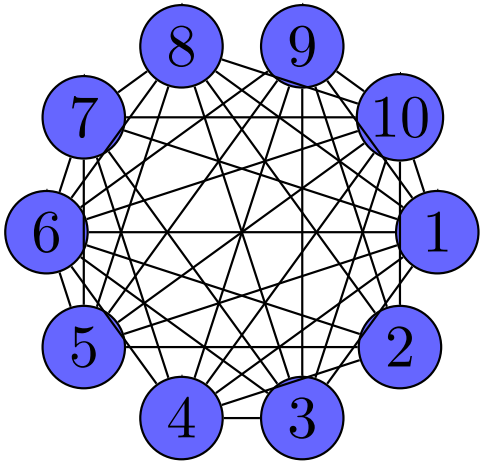
$$\text{QAP}: \quad \min_{\Pi \in \mathcal{S}_n} \|A - \Pi B \Pi^\top\|_F^2$$

- NP-hard in the worst case
- Even approximation within a factor $2^{\log^{1-\epsilon}(n)}$ for $\epsilon > 0$ is NP-hard [Makarychev-Manokaran-Sviridenko '10]
- However, real networks are not designed by adversary!

Focus of this talk

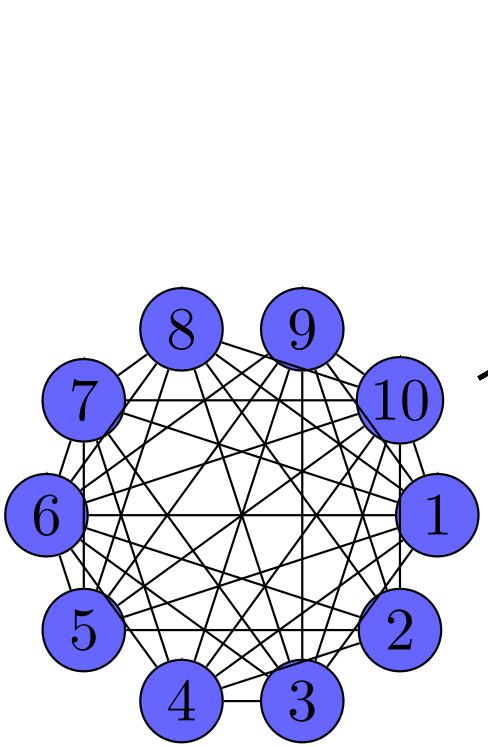
Statistical models for graph matching: (A, B) are random graphs

Correlated Erdős-Rényi random graphs model

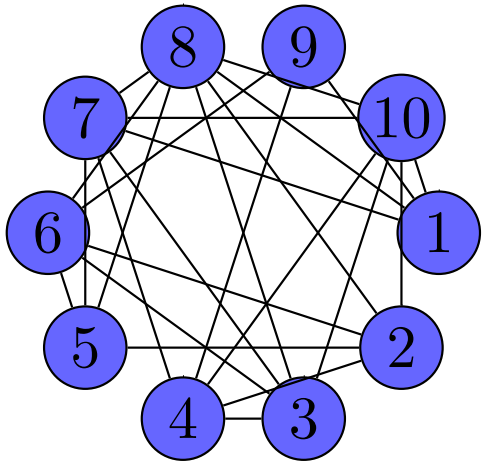


$$G_0 \sim \mathcal{G}(n, p)$$

Correlated Erdős-Rényi random graphs model



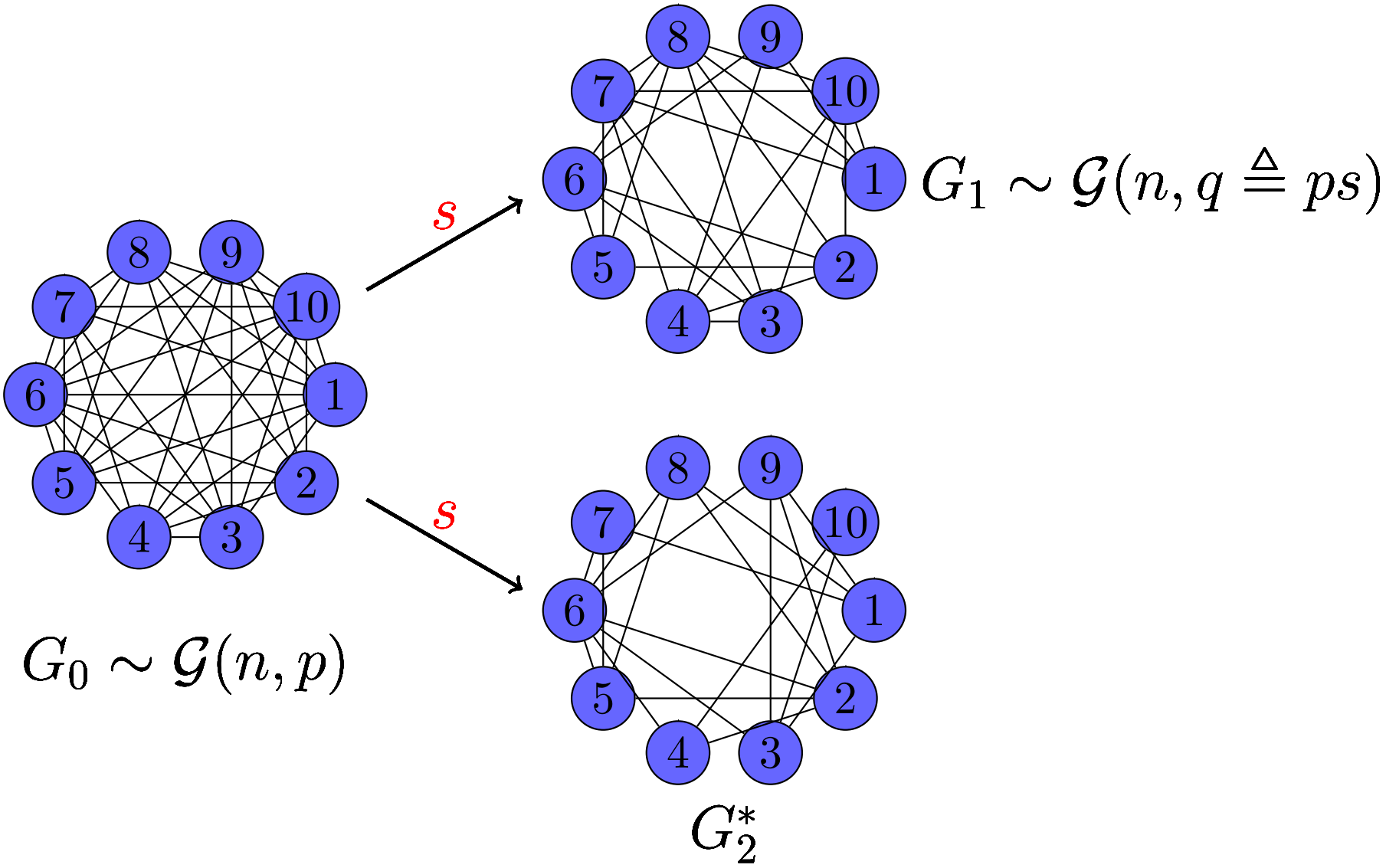
s



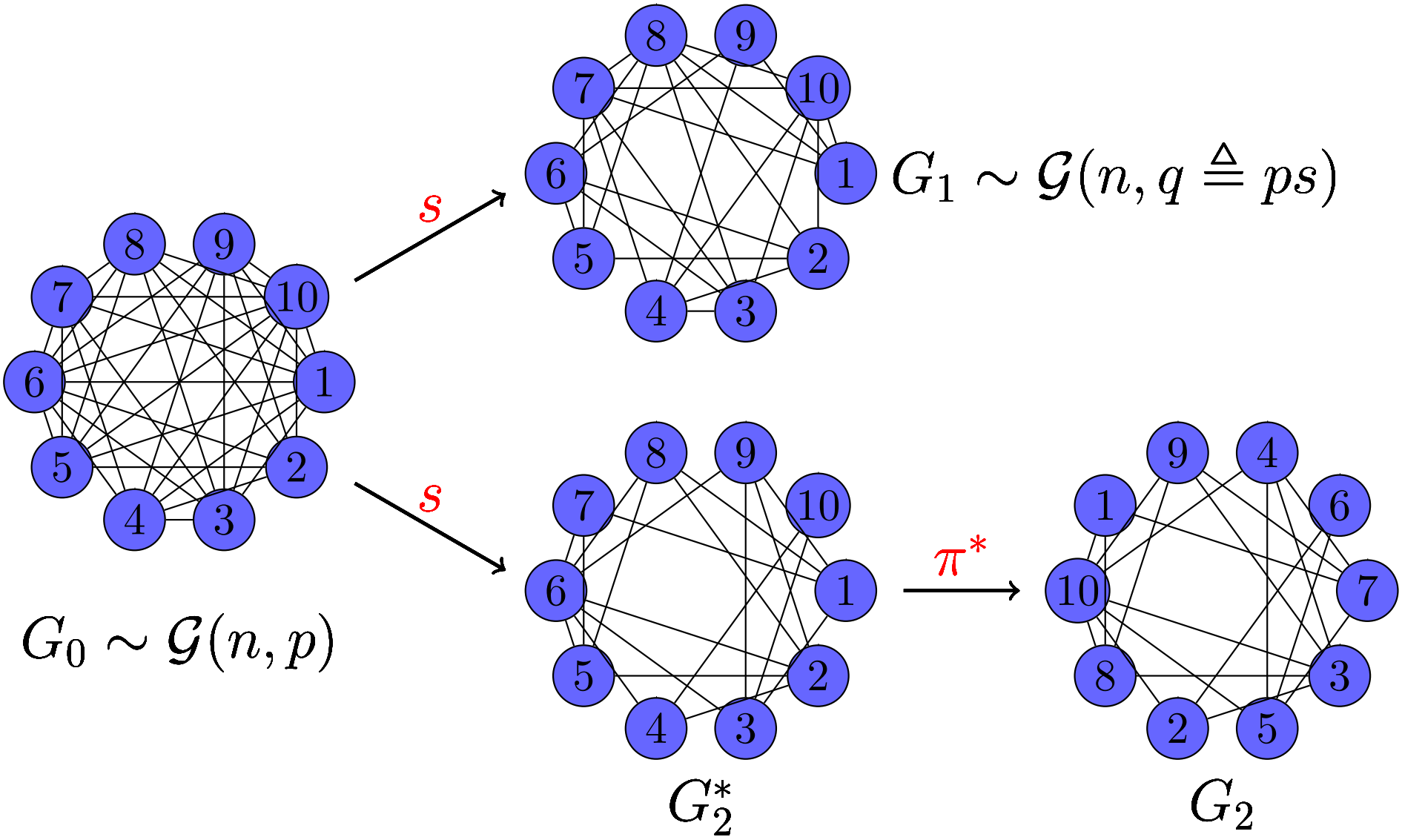
$G_1 \sim \mathcal{G}(n, q \triangleq ps)$

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

Correlated Erdős-Rényi random graphs model



Correlated Erdős-Rényi random graphs model



Prior results

Correlated ER graphs $\mathcal{G}(n, p; s)$: proposed by [Pedarsani-Grossglauser '11]

Theorem (Cullina-Kiyavash '18)

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

$$nps^2 - \log n \to +\infty$$

Prior results

Correlated ER graphs $\mathcal{G}(n, p; s)$: proposed by [Pedarsani-Grossglauser '11]

Theorem (Cullina-Kiyavash '18)

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

$$nps^2 - \log n \to +\infty$$

- Interpretation: Intersection graph $G_1 \wedge G_2^* \sim \mathcal{G}(n, ps^2)$ is connected
- $s = 1$: achieved in linear-time [Bolloás '82, Czajka-Pandurangan '08]
- $s < 1$: little is known for efficient algorithms

Prior results

Correlated ER graphs $\mathcal{G}(n,p;s)$: proposed by [Pedarsani-Grossglauser '11]

Theorem (Cullina-Kiyavash '18)

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

$$nps^2-\log n\to+\infty$$

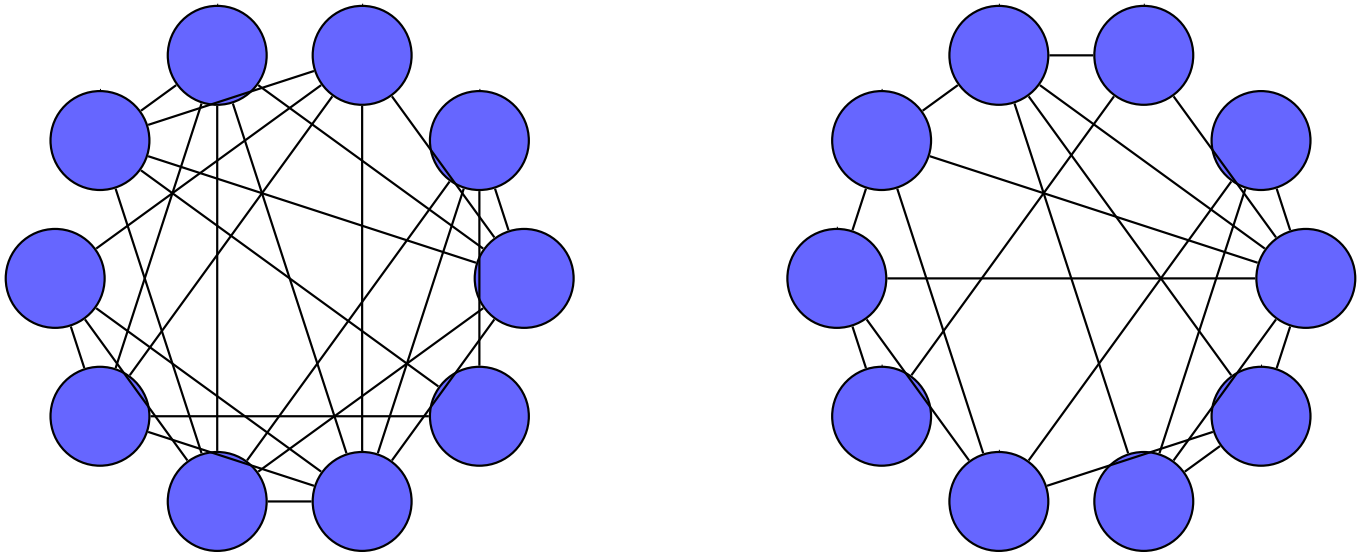
- Interpretation: Intersection graph $G_1\wedge G_2^*\sim\mathcal{G}(n,ps^2)$ is connected
- $s=1$: achieved in linear-time [Bolloás '82, Czajka-Pandurangan '08]
- $s<1$: little is known for efficient algorithms

Question

Is the IT-limit achievable in poly-time for $s<1$?

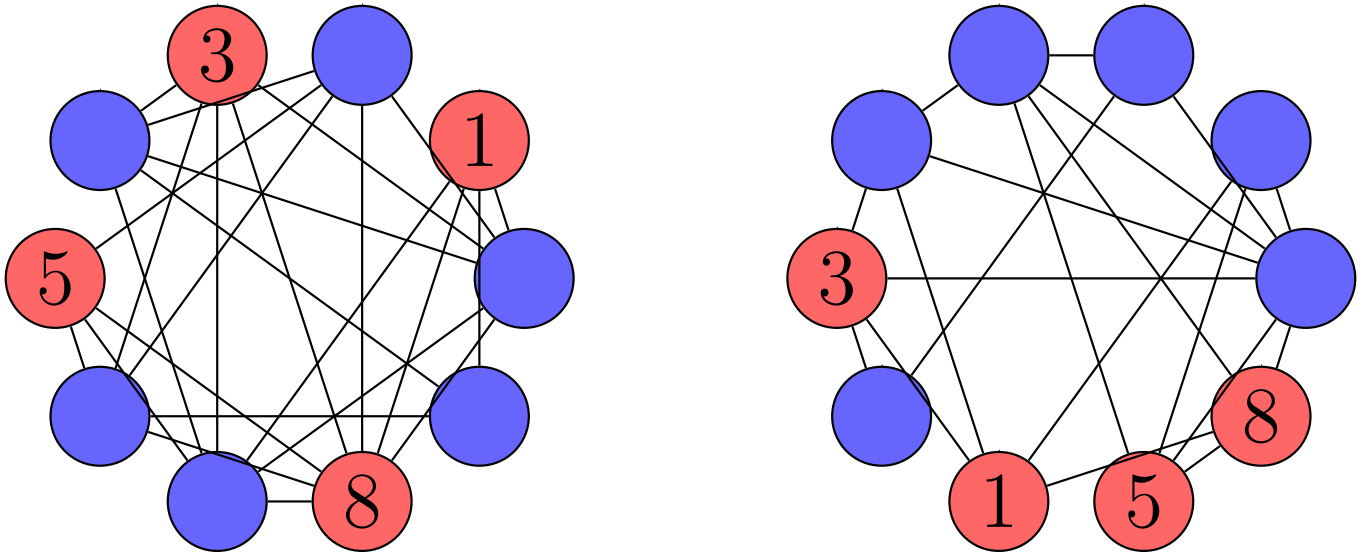
Seeded graph matching

An initial seed set of true pairs is revealed



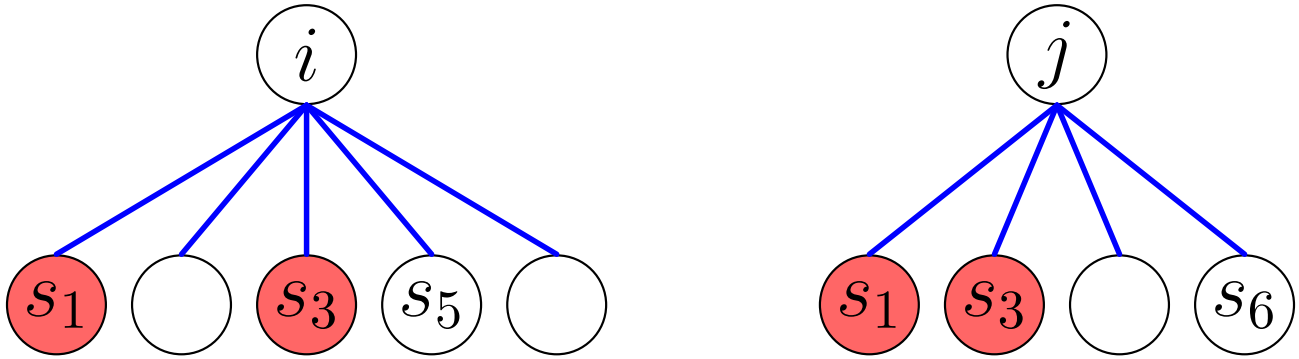
Seeded graph matching

An initial seed set of true pairs is revealed

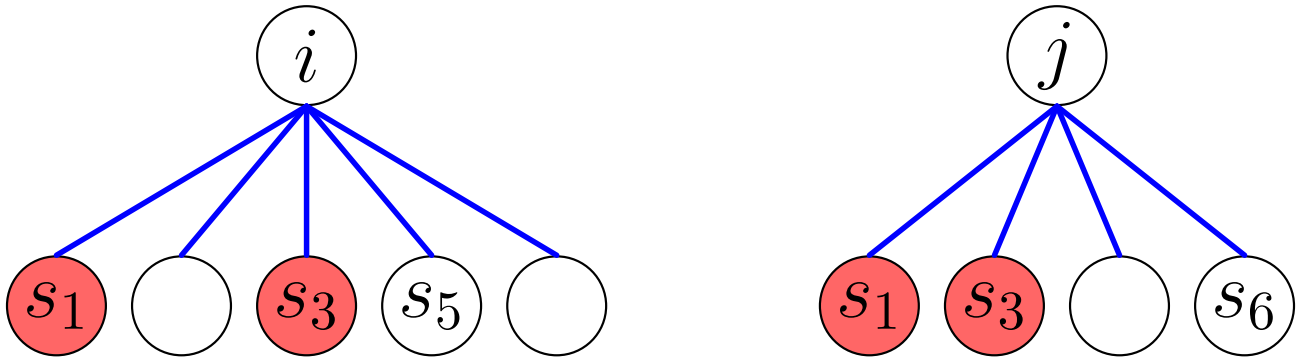


- Seeds are often available in practice
- Seedless algorithm + seeded graph matching:
  K seeds can be obtained in $n^{O(K)}$ time by exhaustive search

Previous ideas: counting the seeded common neighbors



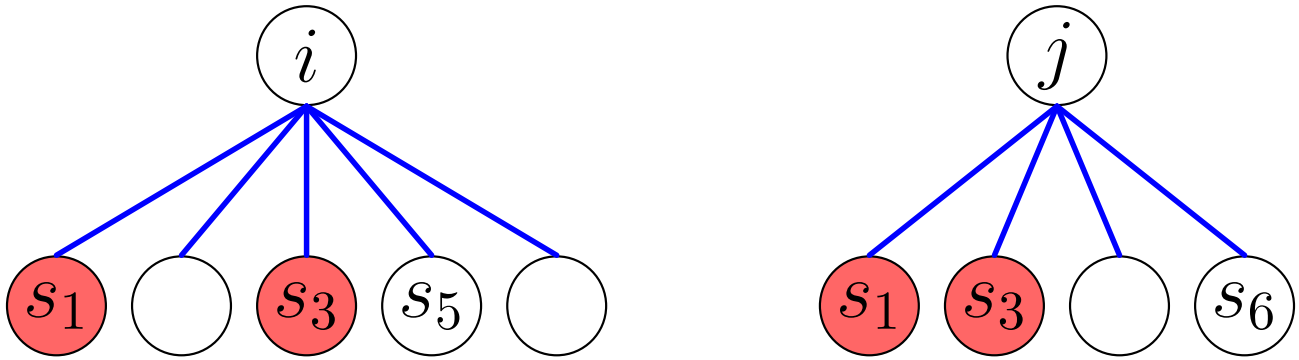
Previous ideas: counting the seeded common neighbors



Suppose each true pair is revealed as seeds w.p. α:

$$\#\text{ of seeded common neighbors} \sim \begin{cases} \text{Binom}\left(n-1,\ ps^2\alpha\right) & \text{if } i \text{ and } j \text{ are true match} \\ \text{Binom}\left(n-2,\ p^2s^2\alpha\right) & \text{if } i \text{ and } j \text{ are fake match} \end{cases}$$

Previous ideas: counting the seeded common neighbors



Suppose each true pair is revealed as seeds w.p. α:

$$\#\text{ of seeded common neighbors} \sim \begin{cases} \text{Binom}\left(n-1,\ ps^2\alpha\right) & \text{if } i \text{ and } j \text{ are true match} \\ \text{Binom}\left(n-2,\ p^2s^2\alpha\right) & \text{if } i \text{ and } j \text{ are fake match} \end{cases}$$

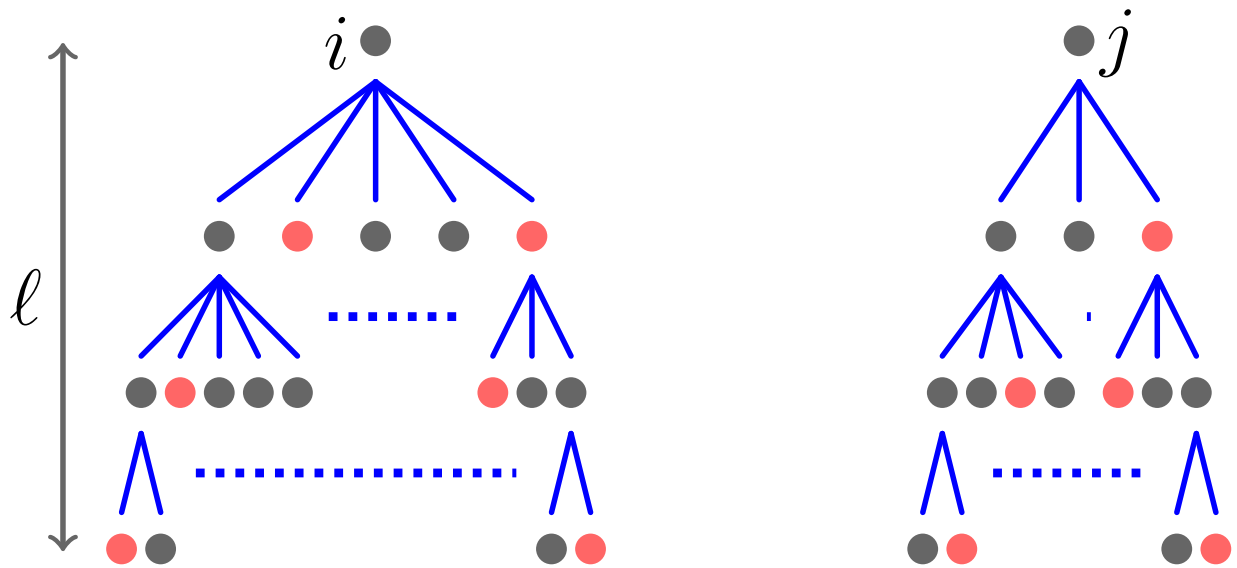
Need

$$\boxed{n\alpha ps^2 \gg 1 \Leftrightarrow n\alpha \gg \frac{1}{ps^2}}$$

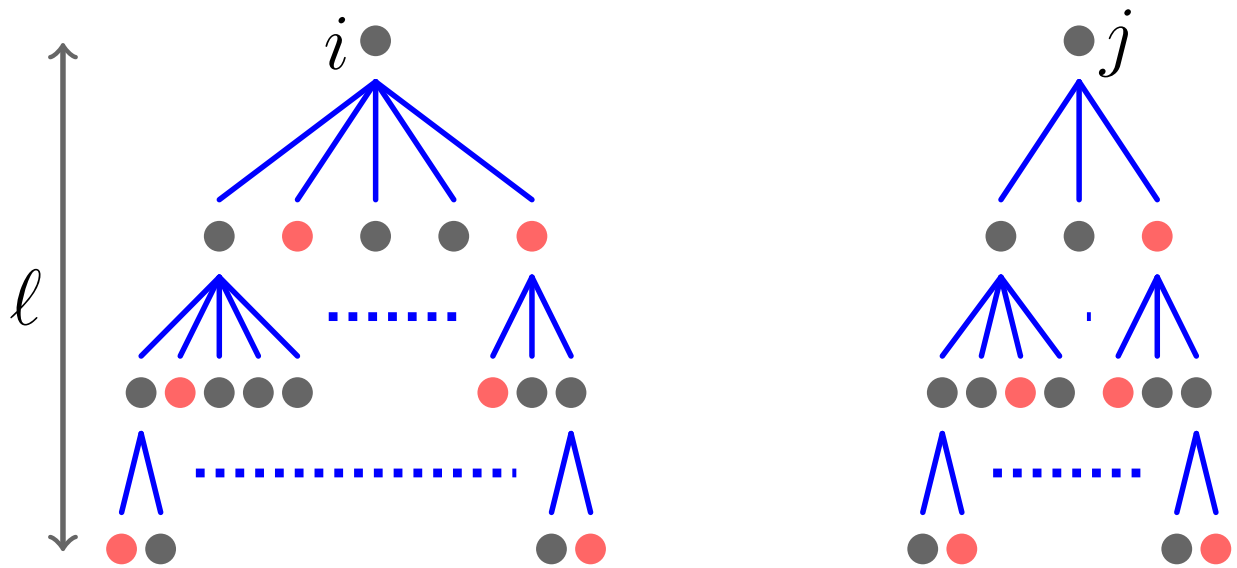
so that true pair has many seeded common neighbors

[Yartseva-Grossglauser '13, Korula-Lattanzi '14]

Our ideas: explore much larger neighborhoods



Our ideas: explore much larger neighborhoods



Our key idea

Match two vertices by comparing the set of seeded vertices in their ℓ-th neighborhoods

Main results

Theorem (Mossel-X. '18)

Suppose $s = \Theta(1)$. Exact recovery is attainable in polynomial time if

- *(Sparse regime) $np \le n^{\epsilon}$ for $\epsilon < 1/6$:*

$$nps^2 - \log n \to +\infty \quad \text{and} \quad \alpha n \ge n^{3\epsilon}$$

- *(Dense regime) $np = \Theta(n^a)$ for a fixed constant $a \in (0, 1]$:*

$$\alpha (nps^2)^{\lfloor 1/a \rfloor} \ge 300 \log n$$

Main results

Theorem (Mossel-X. '18)

Suppose $s = \Theta(1)$. Exact recovery is attainable in polynomial time if

- *(Sparse regime) $np \le n^{\epsilon}$ for $\epsilon < 1/6$:*

$$nps^2 - \log n \to +\infty \quad \text{and} \quad \alpha n \ge n^{3\epsilon}$$

- *(Dense regime) $np = \Theta(n^a)$ for a fixed constant $a \in (0, 1]$:*

$$\alpha (nps^2)^{\lfloor 1/a \rfloor} \ge 300 \log n$$

- $nps^2 - \log n \to +\infty$ is necessary as long as $1 - \alpha = \Omega(1)$

Main results

Theorem (Mossel-X. '18)

Suppose $s = \Theta(1)$. Exact recovery is attainable in polynomial time if

- *(Sparse regime) $np \le n^{\epsilon}$ for $\epsilon < 1/6$:*

$$nps^2 - \log n \to +\infty \quad \text{and} \quad \alpha n \ge n^{3\epsilon}$$

- *(Dense regime) $np = \Theta(n^a)$ for a fixed constant $a \in (0, 1]$:*

$$\alpha (nps^2)^{\lfloor 1/a \rfloor} \ge 300 \log n$$

- $nps^2 - \log n \to +\infty$ is necessary as long as $1 - \alpha = \Omega(1)$
- $1/a \in \mathbb{N}$: $\alpha n = \Omega(\log n) \Rightarrow n^{O(\log n)}$ time seedless algorithm

Main results

Theorem (Mossel-X. '18)

Suppose $s = \Theta(1)$. Exact recovery is attainable in polynomial time if

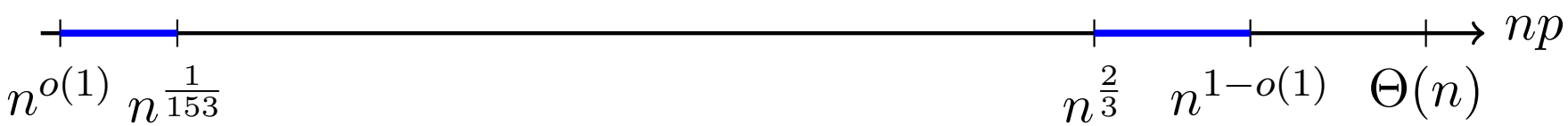
- *(Sparse regime) $np \le n^{\epsilon}$ for $\epsilon < 1/6$:*

$$nps^2 - \log n \to +\infty \quad \text{and} \quad \alpha n \ge n^{3\epsilon}$$

- *(Dense regime) $np = \Theta(n^a)$ for a fixed constant $a \in (0, 1]$:*

$$\alpha (nps^2)^{\lfloor 1/a \rfloor} \ge 300 \log n$$

- $nps^2 - \log n \to +\infty$ is necessary as long as $1 - \alpha = \Omega(1)$
- $1/a \in \mathbb{N}$: $\alpha n = \Omega(\log n) \Rightarrow n^{O(\log n)}$ time seedless algorithm
- [Barak-Chou-Lei-Schramm-Sheng '18]: $n^{O(\log n)}$ time seedless algorithm for



Main results

Theorem (Mossel-X. '18)

Suppose $s = \Theta(1)$. Exact recovery is attainable in polynomial time if

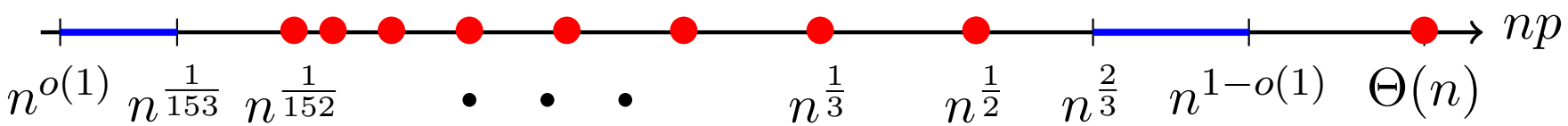
- *(Sparse regime) $np \le n^{\epsilon}$ for $\epsilon < 1/6$:*

$$nps^2 - \log n \to +\infty \quad \text{and} \quad \alpha n \ge n^{3\epsilon}$$

- *(Dense regime) $np = \Theta(n^a)$ for a fixed constant $a \in (0, 1]$:*

$$\alpha (nps^2)^{\lfloor 1/a \rfloor} \ge 300 \log n$$

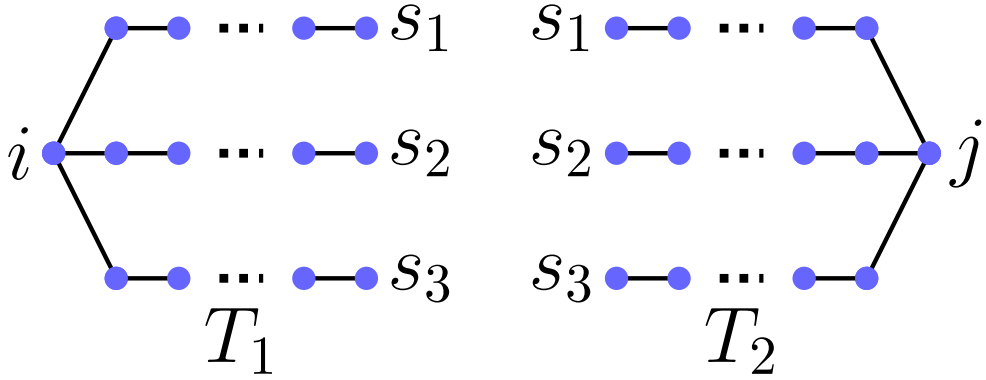
- $nps^2 - \log n \to +\infty$ is necessary as long as $1 - \alpha = \Omega(1)$
- $1/a \in \mathbb{N}$: $\alpha n = \Omega(\log n) \Rightarrow n^{O(\log n)}$ time seedless algorithm
- [Barak-Chou-Lei-Schramm-Sheng '18]: $n^{O(\log n)}$ time seedless algorithm for



Warm-up: explore $\sqrt{n}$-sized neighborhood

Match i and j if there are

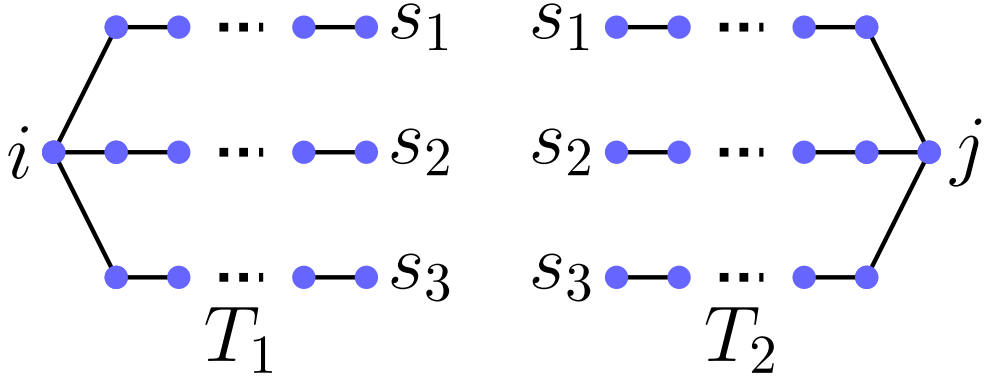
1. m independent ℓ-paths from i to a seeded vertex set
2. m independent ℓ-paths from j to the same seeded vertex set



Warm-up: explore $\sqrt{n}$-sized neighborhood

Match i and j if there are

1. m independent ℓ-paths from i to a seeded vertex set
2. m independent ℓ-paths from j to the same seeded vertex set



- True pairs: in $G_1 \wedge G_2^*$, need $\alpha \left(nps^2\right)^\ell \gg m$
- Fake pairs: in $G_1 \vee G_2^*$, ensure no copy of $T_1 \cup T_2$ (subgraph count)

$$\ell = \left\lfloor \left(\frac{1}{2} - \epsilon\right) \frac{\log n}{\log(nps^2)} \right\rfloor \quad \text{and} \quad m = \left\lceil \frac{2}{\epsilon} \right\rceil$$

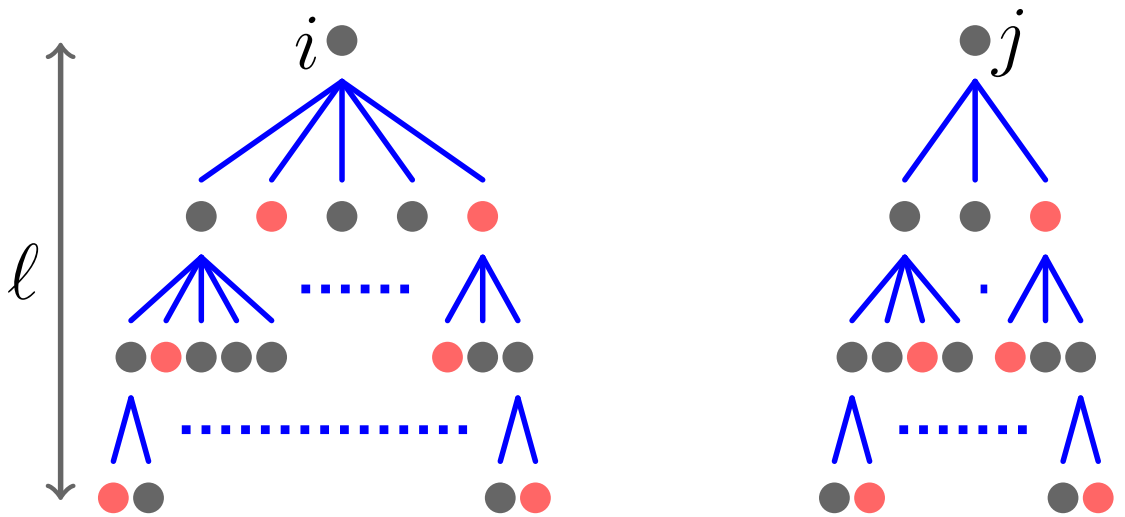
- Need

$$\boxed{\alpha n^{1/2-\epsilon} \geq n^{2\epsilon} \Leftrightarrow \alpha n \geq n^{1/2+3\epsilon}}$$

Dense graph: beyond $\sqrt{n}$-sized neighborhood

$w_{ij} = \#$ of seeded vertices within ℓ-hops from both i in G_1 and j in G_2

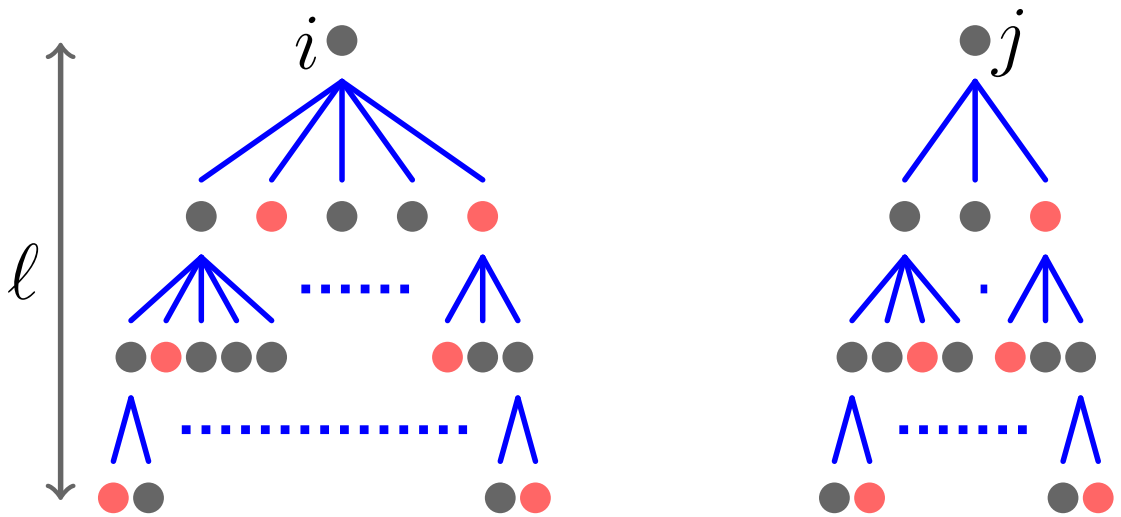
Match i to j if $j \in \arg\max_k w_{ik}$



Dense graph: beyond $\sqrt{n}$-sized neighborhood

$w_{ij} = \#$ of seeded vertices within ℓ-hops from both i in G_1 and j in G_2

Match i to j if $j \in \arg\max_k w_{ik}$

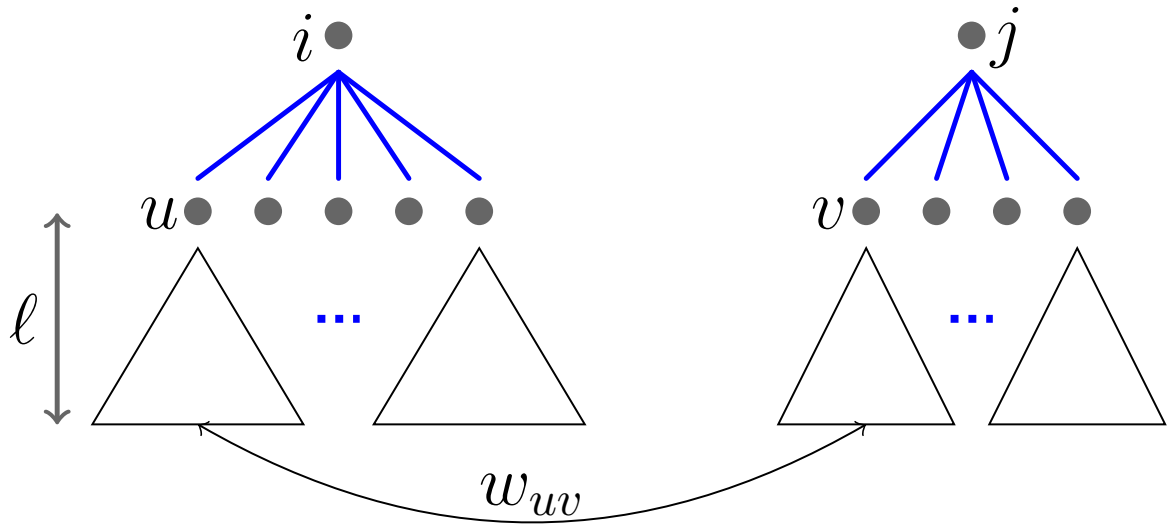


- $G_1 \wedge G_2^*$: the typical size of NB $\approx (nps^2)^\ell$
- $G_1 \vee G_2^*$: the typical size of intersection of two NBs $\approx (nps)^\ell \frac{(nps)^\ell}{n}$
- Choose

$$\ell = \left\lfloor \frac{\log n}{\log(np)} \right\rfloor \quad (\text{diameter}(G) - 1) \Rightarrow \alpha(nps^2)^\ell \gg \log n$$

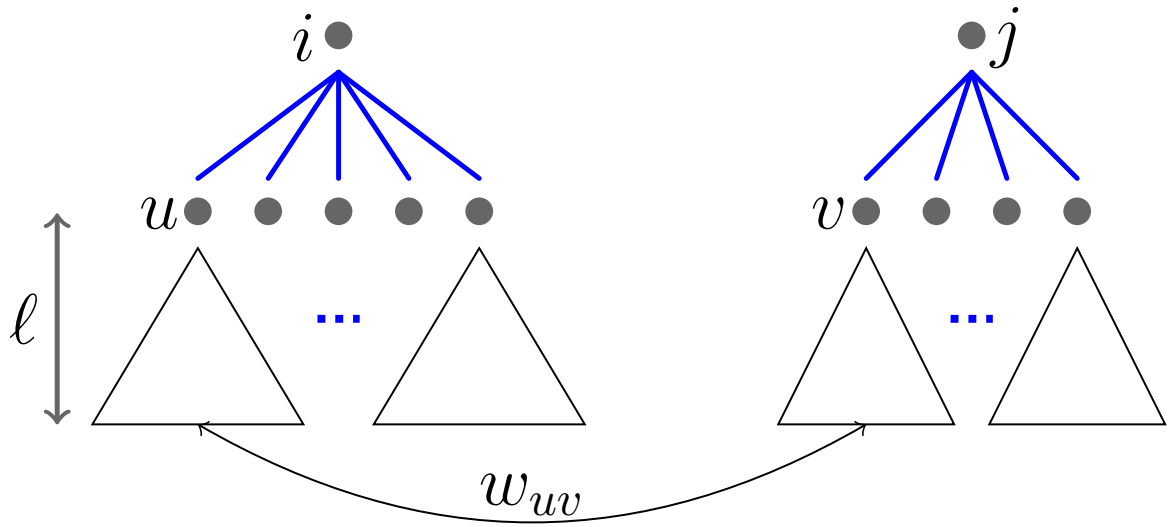
Sparse graph: beyond $\sqrt{n}$-sized neighborhood

Issue: Large neighborhoods of nearby vertices i and j have large overlaps



Sparse graph: beyond $\sqrt{n}$-sized neighborhood

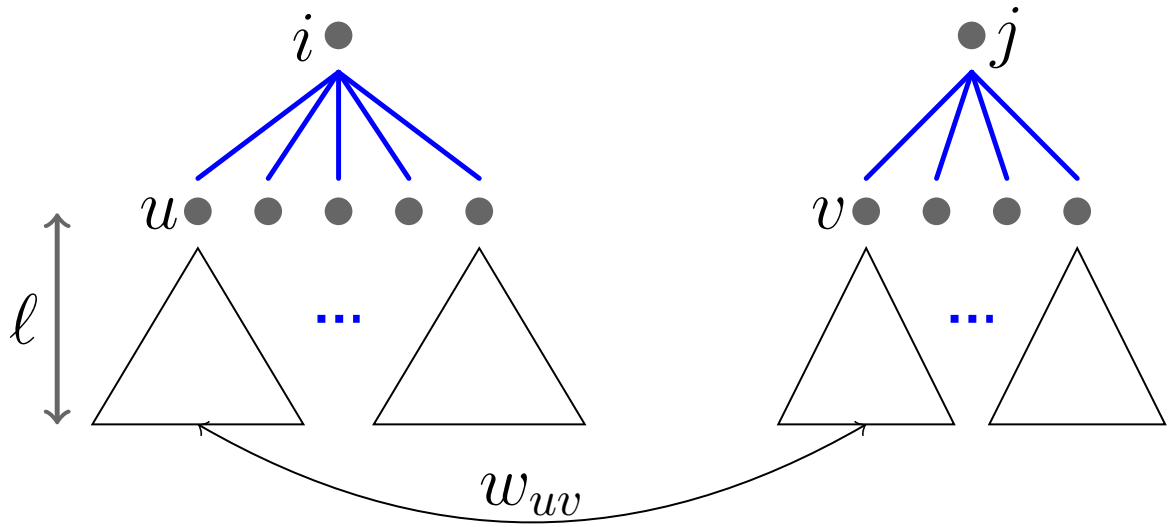
Issue: Large neighborhoods of nearby vertices i and j have large overlaps



Match i to j if $Z_{ij} \triangleq \sum_{u,v} \underbrace{A_{iu}B_{jv}\mathbf{1}_{\{w_{uv}\geq\eta\}}}_{\text{dependency}}$ is large

Sparse graph: beyond $\sqrt{n}$-sized neighborhood

Issue: Large neighborhoods of nearby vertices i and j have large overlaps

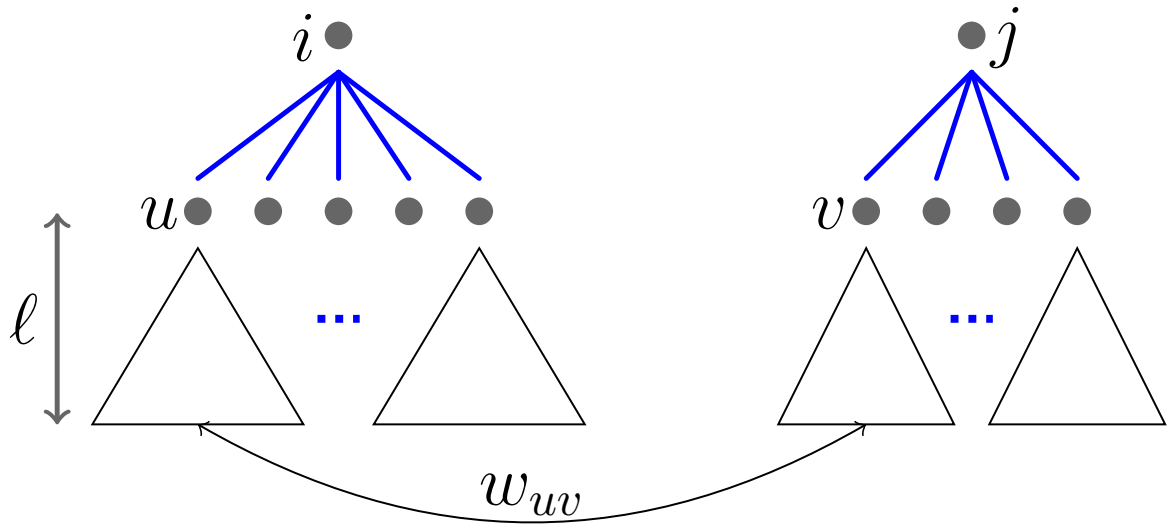


Match i to j if $Z_{ij} \triangleq \sum_{u,v} \underbrace{A_{iu}B_{jv}\mathbf{1}_{\{w_{uv}\geq\eta\}}}_{\text{dependency}}$ is large

Set $\ell = \left\lfloor \frac{(1-\epsilon)\log n}{\log(np)} \right\rfloor \Rightarrow \alpha n^{1-\epsilon} \geq n^{2\epsilon}$

Sparse graph: beyond $\sqrt{n}$-sized neighborhood

Issue: Large neighborhoods of nearby vertices i and j have large overlaps



Match i to j if $Z_{ij} \triangleq \sum_{u,v} \underbrace{A_{iu}B_{jv}\mathbf{1}_{\{w_{uv}\geq\eta\}}}_{\text{dependency}}$ is large

Set $\ell = \left\lfloor \frac{(1-\epsilon)\log n}{\log(np)} \right\rfloor \Rightarrow \alpha n^{1-\epsilon} \geq n^{2\epsilon}$

Analysis: "replica" trick + multivariate polynomial concentration [Vu '02]

Concluding remarks

- Seeded graph matching can achieve in poly-time the IT limit

$$nps^2 - \log n \to +\infty$$

- The # of seeds needed for poly-time recovery can be as low as

$$\begin{cases} n^{3\epsilon}, & \text{if } np \le n^{\epsilon} \\ \Omega\left(\log n\right), & \text{if } np = \Theta\left(n^{1/k}\right), k \in \mathbb{N} \end{cases}$$

Concluding remarks

- Seeded graph matching can achieve in poly-time the IT limit

$$nps^2 - \log n \to +\infty$$

- The # of seeds needed for poly-time recovery can be as low as

$$\begin{cases} n^{3\epsilon}, & \text{if } np \le n^{\epsilon} \\ \Omega\left(\log n\right), & \text{if } np = \Theta\left(n^{1/k}\right), k \in \mathbb{N} \end{cases}$$

Subsequent work

- $nps^2 - \log n \to +\infty$ achievable in poly-time under seedless case?
  [J. Ding, Z. Ma, Y. Wu, & X. *Efficient random graph matching via degree profiles.* arXiv:1811.07821. Result: $1 - s \lesssim \log^{-2}(np)$ or $1 - s \lesssim \log^{-2/3}(n)$]
- The minimum number of seeds needed for poly-time recovery?