Correlated stochastic block models: graph matching and community recovery

Based on joint works with Julia Gaudio and Anirudh Sridhar

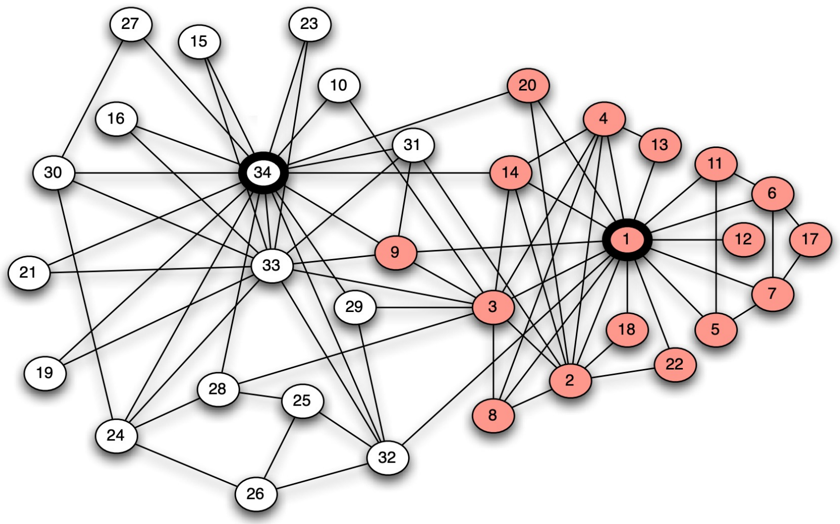
Miklós Z. Rácz

Northwestern University

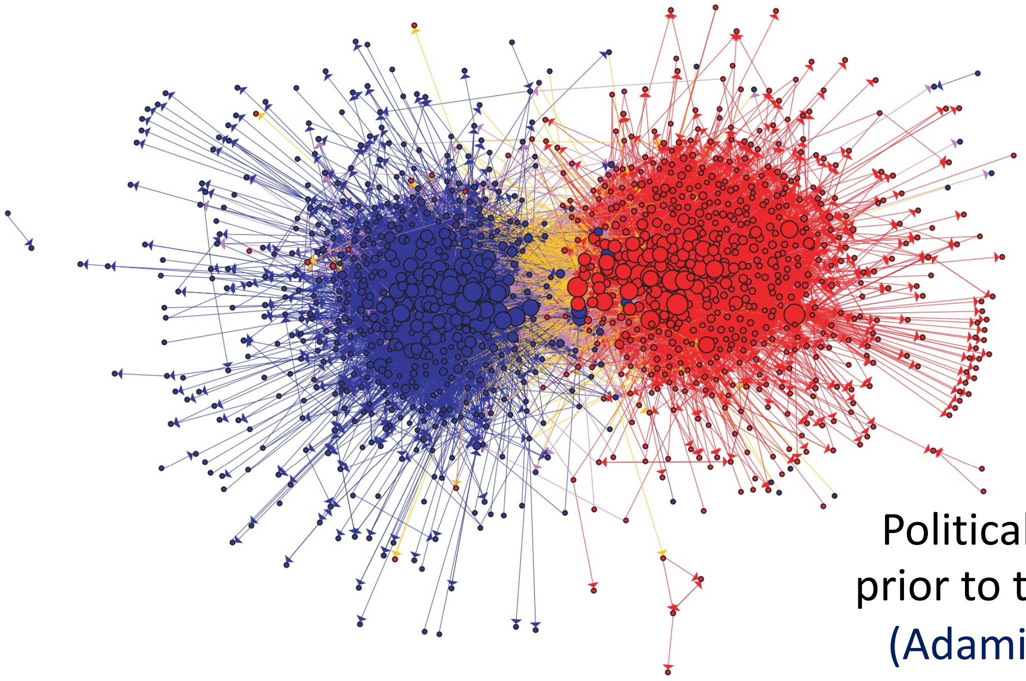
Oxford Discrete Mathematics and Probability Seminar

March 7, 2023

Recovering communities in networks

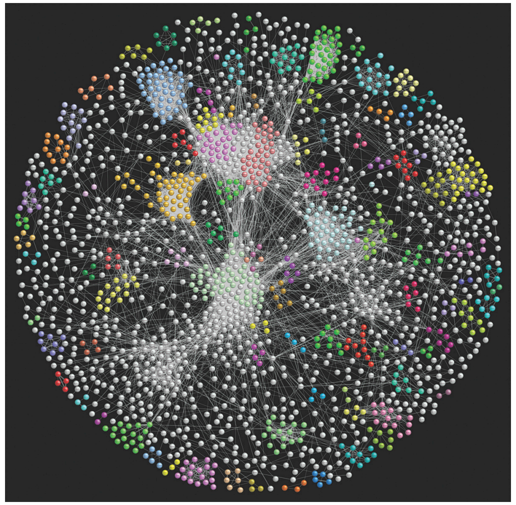


Zachary's karate club (1970-72; 1977)



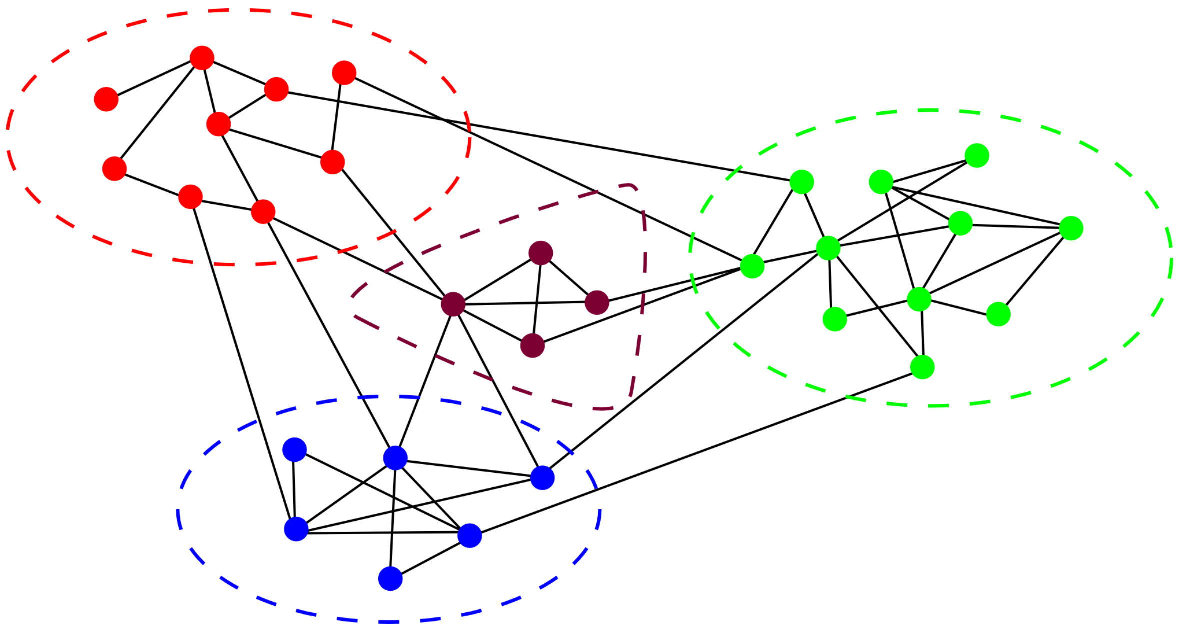
Political blogs in the US, prior to the 2004 elections (Adamic, Glance, 2005)

Facebook: 2.9+ billion monthly active users
2+ billion daily active users



Drosophila protein-protein interaction network (Guruharsha et al., 2011)

Stochastic block model (SBM)

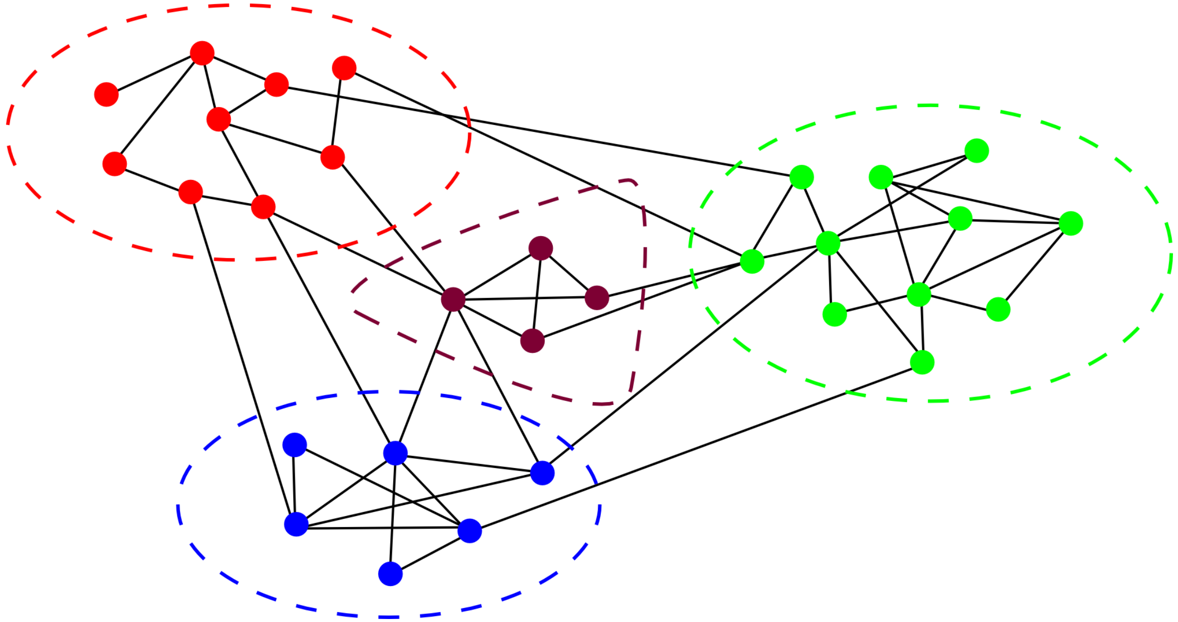


Holland, Laskey, Leinhardt (1983)

Many works in physics, statistics, probability, CS, info theory… including:

- Decelle, Krzakala, Moore, Zdeborová (2011)
- Mossel, Neeman, Sly (2012, 2013a,b, 2014)
- Massoulié (2014)
- Abbé, Bandeira, Hall (2014)
- Abbé, Sandon (2015a,b,c)
- Bordenave, Lelarge, Massoulié (2015)
- Abbé (2017)
- …

Stochastic block model (SBM)



Holland, Laskey, Leinhardt (1983)

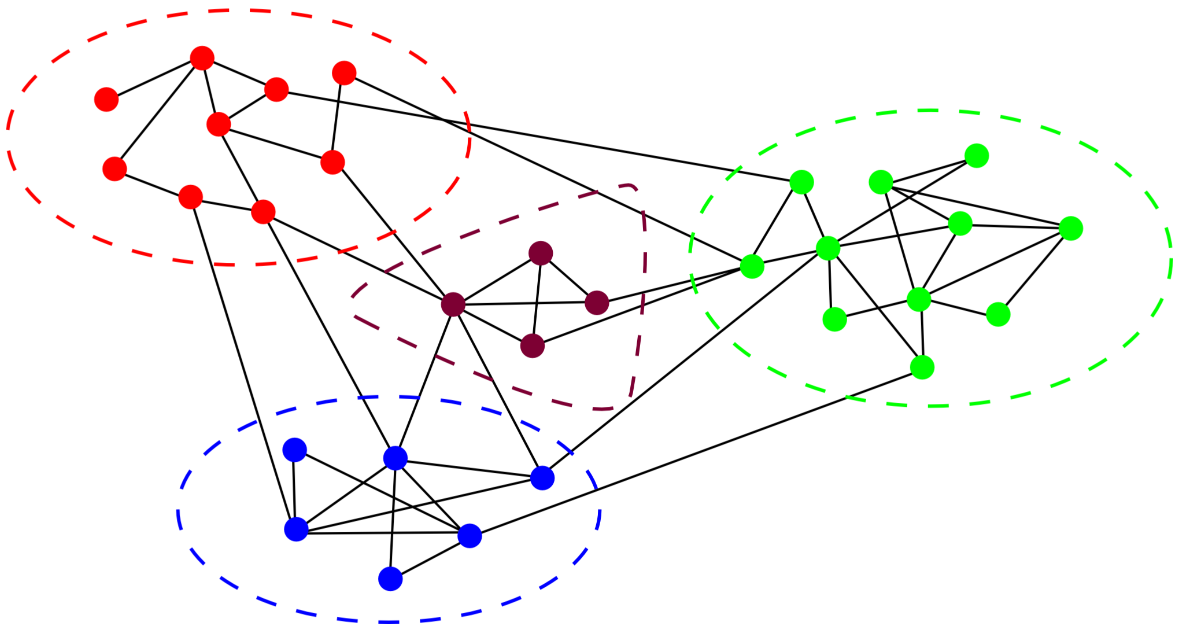
Many works in physics, statistics, probability, CS, info theory… including:

- Decelle, Krzakala, Moore, Zdeborová (2011)
- Mossel, Neeman, Sly (2012, 2013a,b, 2014)
- Massoulié (2014)
- Abbé, Bandeira, Hall (2014)
- Abbé, Sandon (2015a,b,c)
- Bordenave, Lelarge, Massoulié (2015)
- Abbé (2017)
- ...

Q: given the graph without community labels, can we recover the communities?

- Partial recovery?
- Almost exact recovery?
- Exact recovery?

Stochastic block model (SBM)



Holland, Laskey, Leinhardt (1983)

Many works in physics, statistics, probability, CS, info theory… including:

- Decelle, Krzakala, Moore, Zdeborová (2011)
- Mossel, Neeman, Sly (2012, 2013a,b, 2014)
- Massoulié (2014)
- Abbé, Bandeira, Hall (2014)
- Abbé, Sandon (2015a,b,c)
- Bordenave, Lelarge, Massoulié (2015)
- Abbé (2017)
- …

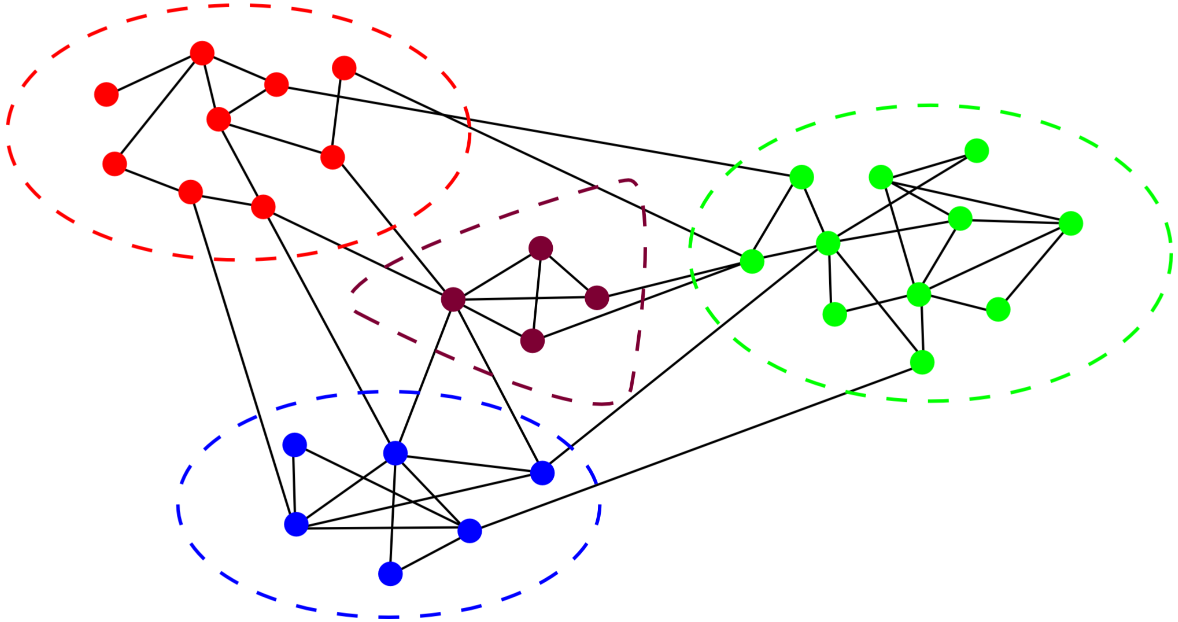
Q: given the graph without community labels, can we recover the communities?

- Partial recovery?
- Almost exact recovery?
- Exact recovery?

This talk: two balanced communities

- n nodes
- $\sigma_i \in \{+1, -1\}$ i.i.d. uniform community labels
- Given $\boldsymbol{\sigma} = \{\sigma_i\}$, edges drawn independently:
  - If $\sigma_i = \sigma_j$, then $i \sim j$ with prob. p
  - If $\sigma_i \neq \sigma_j$, then $i \sim j$ with prob. q

Stochastic block model (SBM)



Holland, Laskey, Leinhardt (1983)

Many works in physics, statistics, probability, CS, info theory… including:

- Decelle, Krzakala, Moore, Zdeborová (2011)
- Mossel, Neeman, Sly (2012, 2013a,b, 2014)
- Massoulié (2014)
- Abbé, Bandeira, Hall (2014)
- Abbé, Sandon (2015a,b,c)
- Bordenave, Lelarge, Massoulié (2015)
- Abbé (2017)
- …

Q: given the graph without community labels, can we recover the communities?

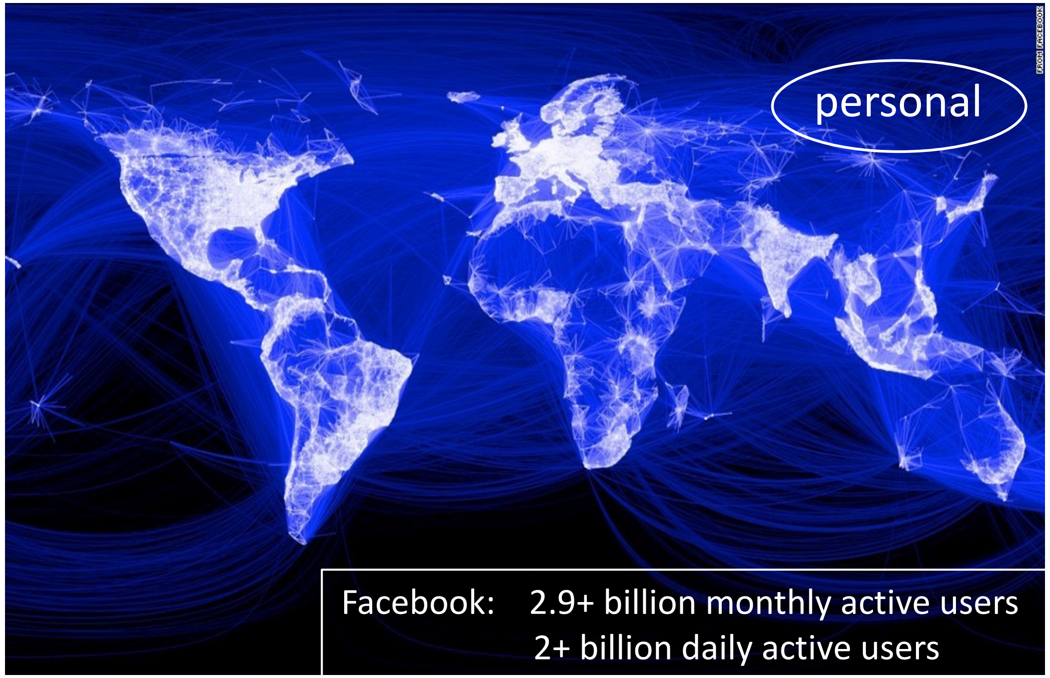
- Partial recovery?
- Almost exact recovery?
- Exact recovery?

$$G \sim SBM(n, p, q)$$

This talk: two balanced communities

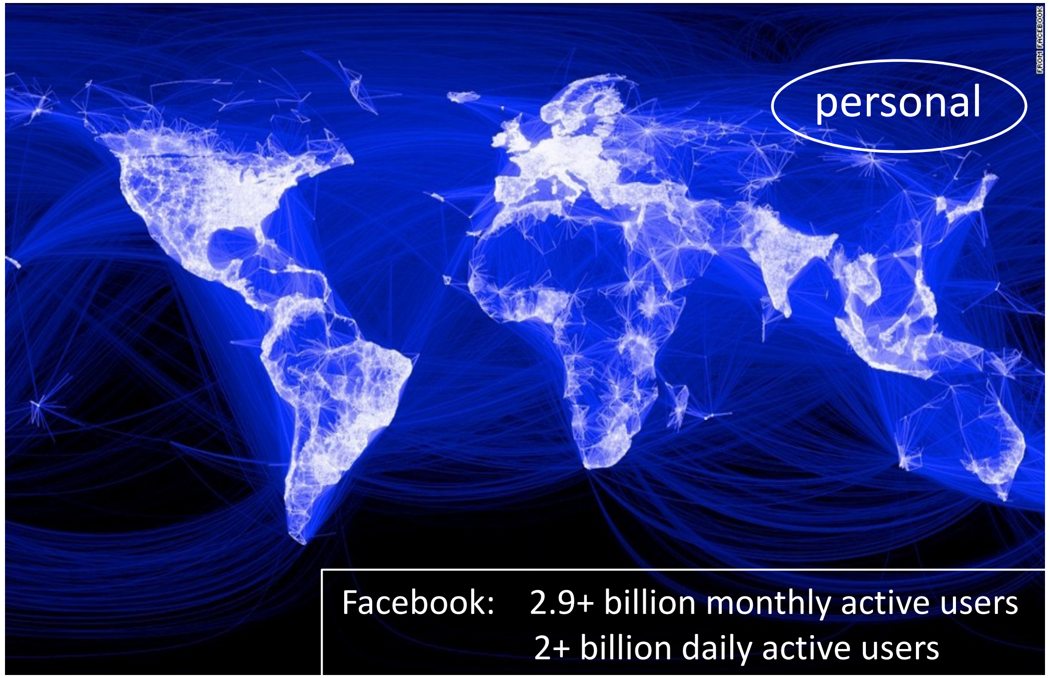
- n nodes
- $\sigma_i \in \{+1, -1\}$ i.i.d. uniform community labels
- Given $\boldsymbol{\sigma} = \{\sigma_i\}$, edges drawn independently:
  - If $\sigma_i = \sigma_j$, then $i \sim j$ with prob. p
  - If $\sigma_i \neq \sigma_j$, then $i \sim j$ with prob. q

Multiple correlated networks

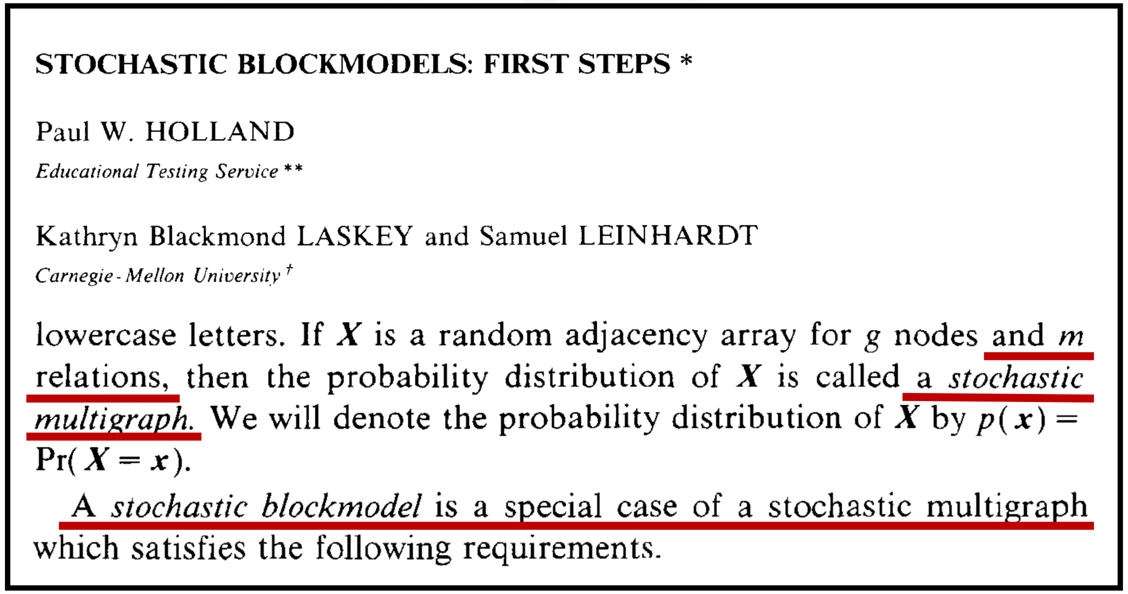


Q: can we synthesize information from multiple correlated networks to better recover communities?

Multiple correlated networks



professional



STOCHASTIC BLOCKMODELS: FIRST STEPS *

Paul W. HOLLAND

Educational Testing Service **

Kathryn Blackmond LASKEY and Samuel LEINHARDT

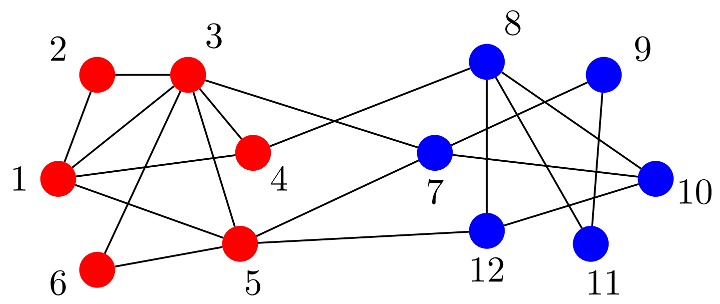
Carnegie-Mellon University †

lowercase letters. If X is a random adjacency array for g nodes and m relations, then the probability distribution of X is called a *stochastic multigraph*. We will denote the probability distribution of X by $p(x) = \Pr(X = x)$.

A *stochastic blockmodel* is a special case of a stochastic multigraph which satisfies the following requirements.

Q: can we synthesize information from multiple correlated networks to better recover communities?

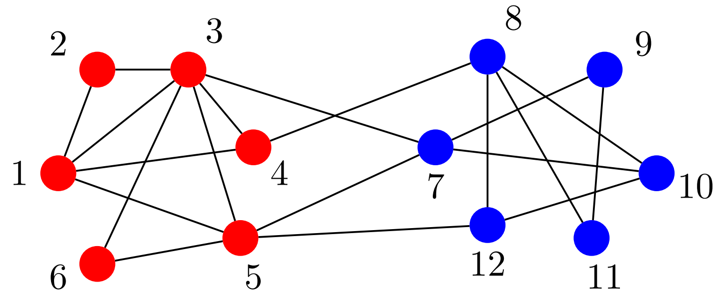
Correlated stochastic block model



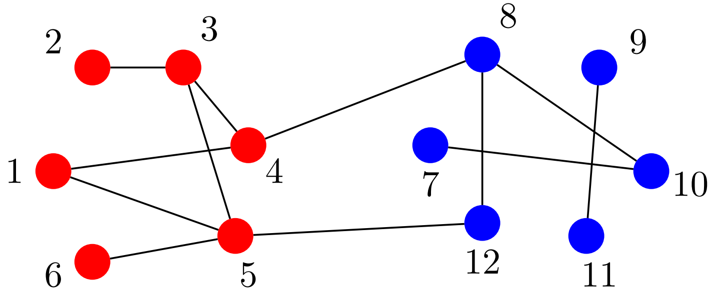
$$G \sim \mathrm{SBM}(n, p, q)$$

Correlated stochastic block model

subsample



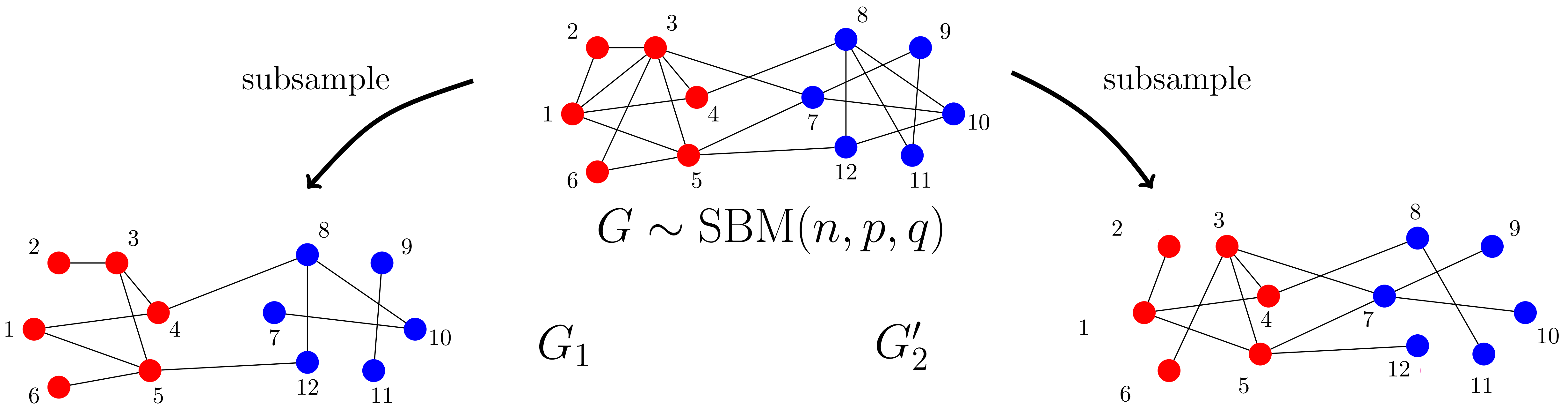
$$G \sim \mathrm{SBM}(n, p, q)$$



$$G_1$$

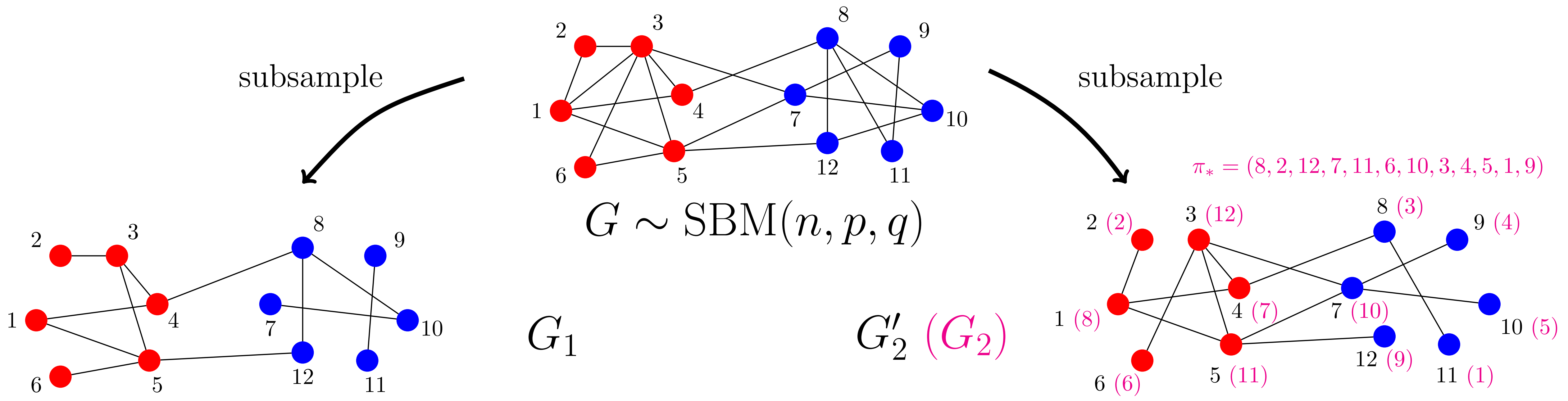
- Subsampling probability $s \in [0,1]$

Correlated stochastic block model



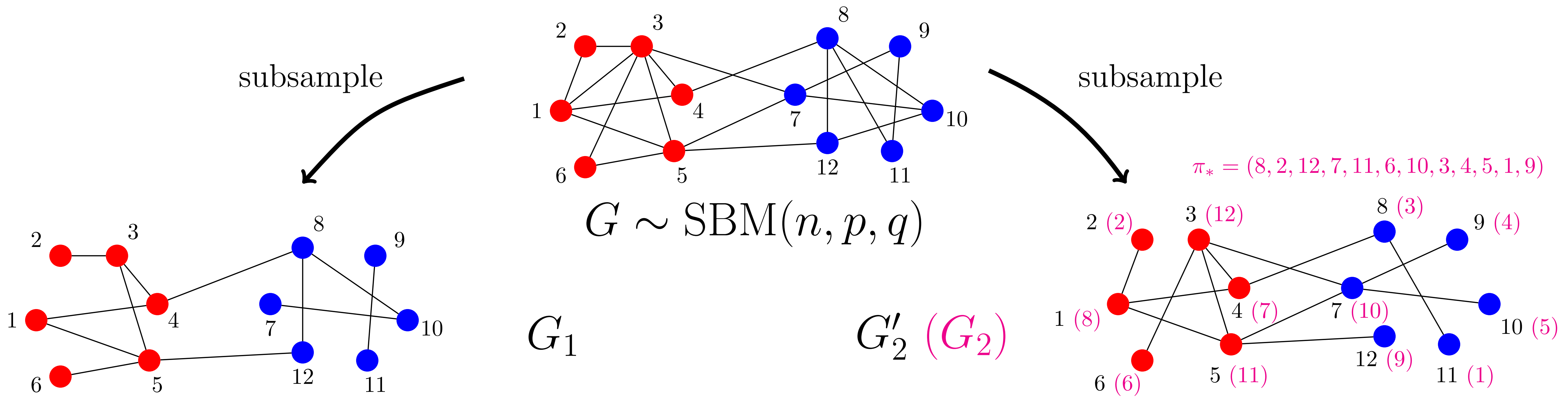
- Subsampling probability $s \in [0,1]$

Correlated stochastic block model



- Subsampling probability $s \in [0,1]$
- π_* uniformly random permutation of $[n]$

Correlated stochastic block model



- Subsampling probability $s \in [0,1]$
- π_* uniformly random permutation of $[n]$

$$\boxed{(G_1, G_2) \sim \mathrm{CSBM}(n,p,q,s)}$$

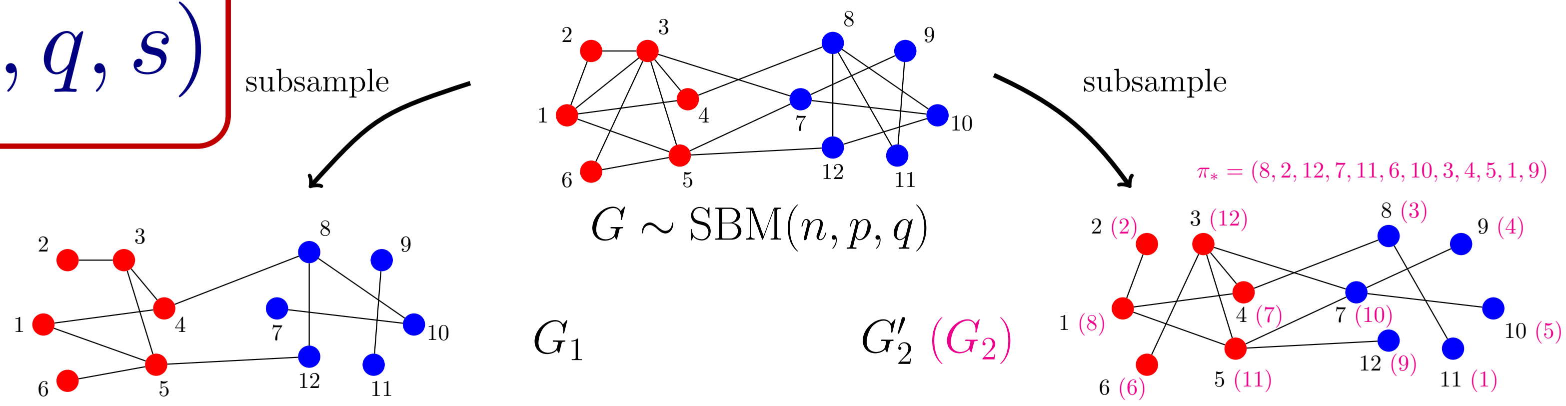
- Marginally $G_1, G_2 \sim SBM(n, ps, qs)$
- Corresponding edges are *correlated*

(Onaran, Garg, Erkip, 2016)

HLL83: (G_1, G_2') is a “pair-dependent SBM”

Correlated stochastic block model

$$(G_1, G_2) \sim \mathrm{CSBM}(n, p, q, s)$$



Main Q:

- given (G_1, G_2), when can we (exactly) recover the communities?
- can we do so in regimes where it is impossible to do so using only G_1?

Exact community recovery in the SBM

Need no isolated vertices $\Longrightarrow$ logarithmic degree regime: $p = a\log(n)/n$ and $q = b\log(n)/n$

Exact community recovery in the SBM

Need no isolated vertices $\Longrightarrow$ logarithmic degree regime: $p = a\log(n)/n$ and $q = b\log(n)/n$

Theorem (Abbé, Bandeira, Hall, 2014; Mossel, Neeman, Sly, 2014)

Consider the balanced two-community SBM: $G \sim \mathrm{SBM}\left(n, \frac{a\log n}{n}, \frac{b\log n}{n}\right)$

Exact recovery is **possible** (in polynomial time) if $\boxed{\left|\sqrt{a}-\sqrt{b}\right| > \sqrt{2}}$

Exact recovery is **impossible** if $\boxed{\left|\sqrt{a}-\sqrt{b}\right| < \sqrt{2}}$

Exact community recovery in the SBM

Need no isolated vertices $\Longrightarrow$ logarithmic degree regime: $p = a\log(n)/n$ and $q = b\log(n)/n$

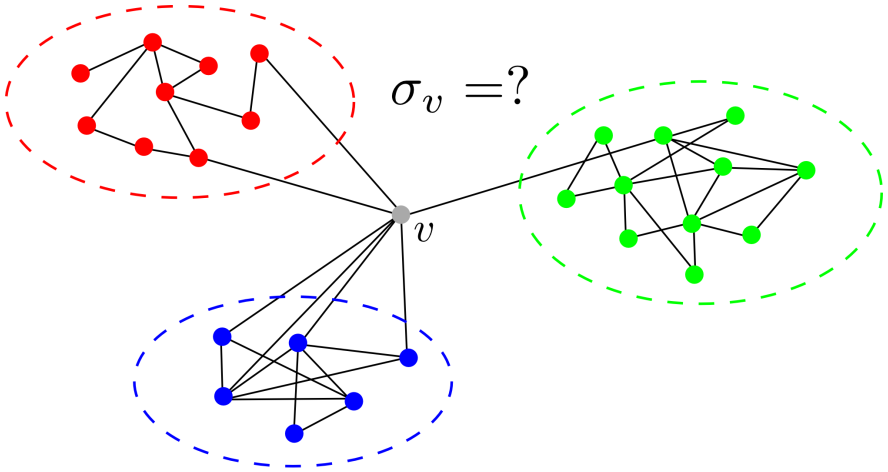
Theorem (Abbé, Bandeira, Hall, 2014; Mossel, Neeman, Sly, 2014)

Consider the balanced two-community SBM: $G \sim \mathrm{SBM}\left(n, \frac{a\log n}{n}, \frac{b\log n}{n}\right)$

Exact recovery is **possible** (in polynomial time) if $\boxed{\left|\sqrt{a}-\sqrt{b}\right| > \sqrt{2}}$

Exact recovery is **impossible** if $\boxed{\left|\sqrt{a}-\sqrt{b}\right| < \sqrt{2}}$

Abbé, Sandon (2015): threshold for general SBMs

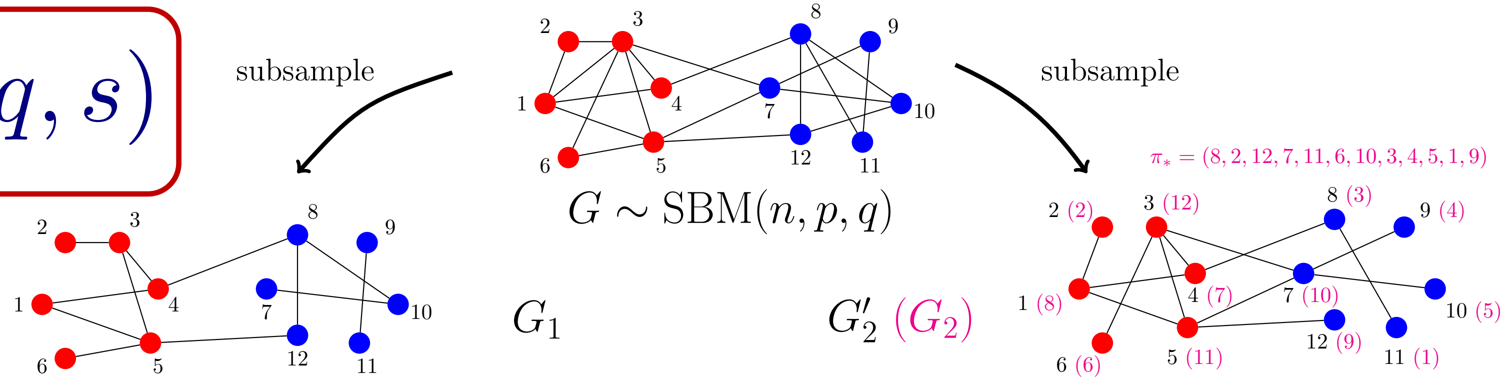


Intuition:

- Testing multivariate Poisson distributions
- Want error probability $n^{-1+o(1)}$
- Error exponent given by Chernoff-Hellinger divergence

Exact community recovery in correlated SBMs

$$\boxed{(G_1, G_2) \sim \mathrm{CSBM}(n, p, q, s)}$$

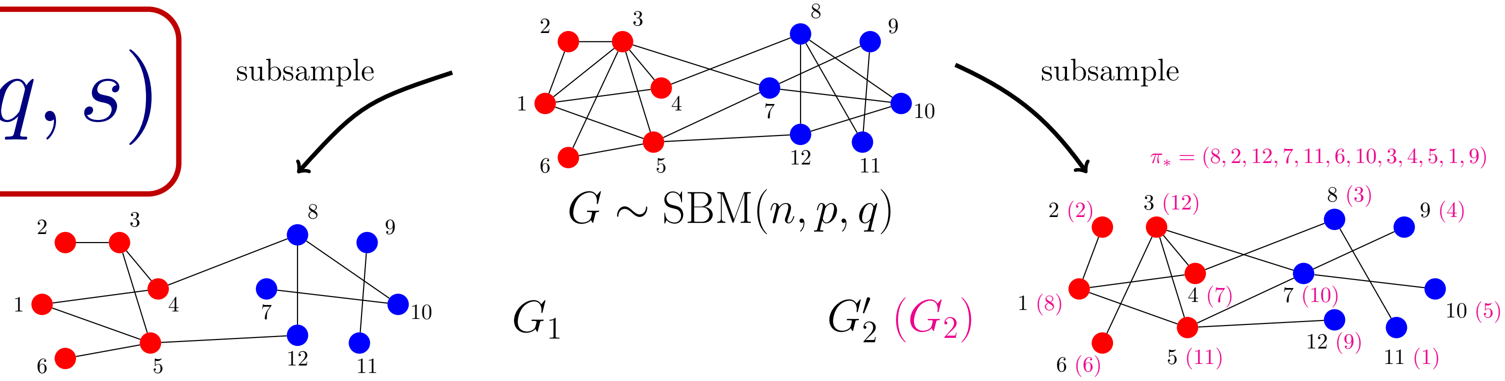


Since $G_1 \sim SBM(n, ps, qs)$, exact community recovery is possible from G_1 iff

$$\boxed{\left|\sqrt{a} - \sqrt{b}\right| > \sqrt{2/s}}$$

Exact community recovery in correlated SBMs

$$\boxed{(G_1, G_2) \sim \mathrm{CSBM}(n, p, q, s)}$$

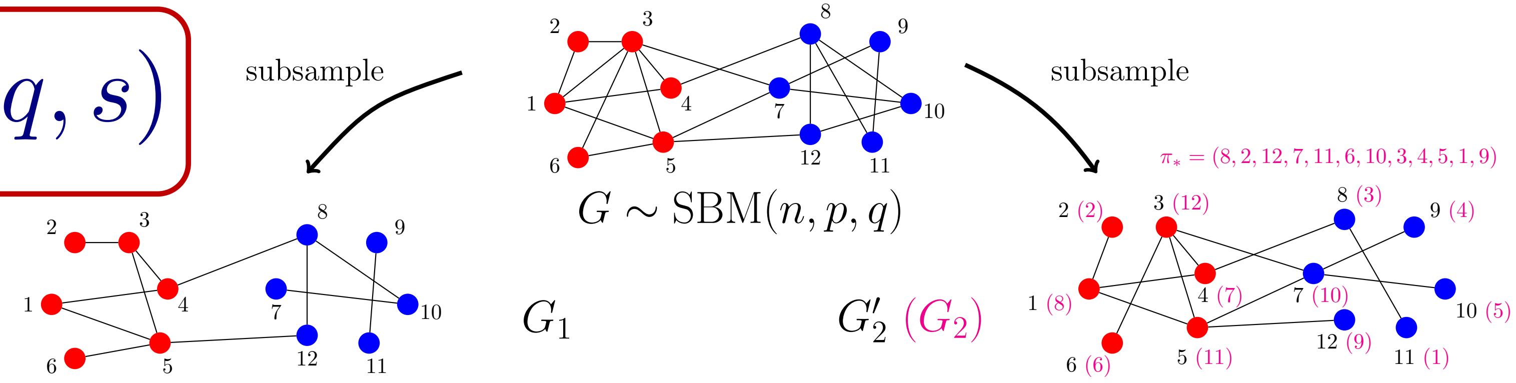


Since $G_1 \sim SBM(n, ps, qs)$, exact community recovery is possible from G_1 iff $\boxed{\left|\sqrt{a} - \sqrt{b}\right| > \sqrt{2/s}}$

How can we use both G_1 and G_2? **Suppose that π_* is known.**

Exact community recovery in correlated SBMs

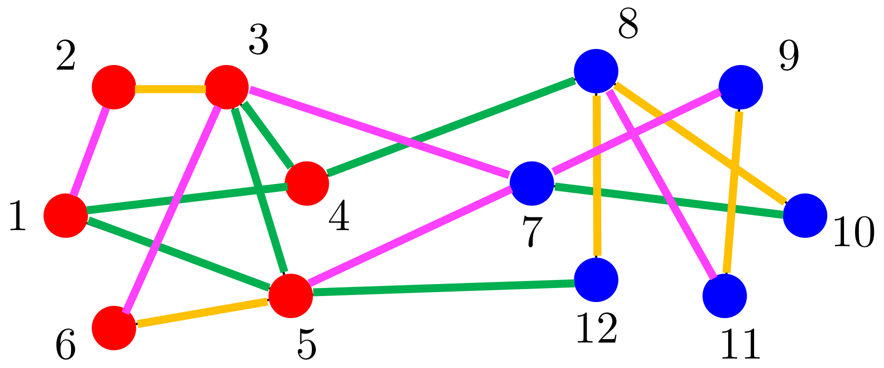
$$\boxed{(G_1, G_2) \sim \mathrm{CSBM}(n, p, q, s)}$$



Since $G_1 \sim SBM(n, ps, qs)$, exact community recovery is possible from G_1 iff $\boxed{\left|\sqrt{a} - \sqrt{b}\right| > \sqrt{2/s}}$

How can we use both G_1 and G_2? **Suppose that π_* is known.** Then:

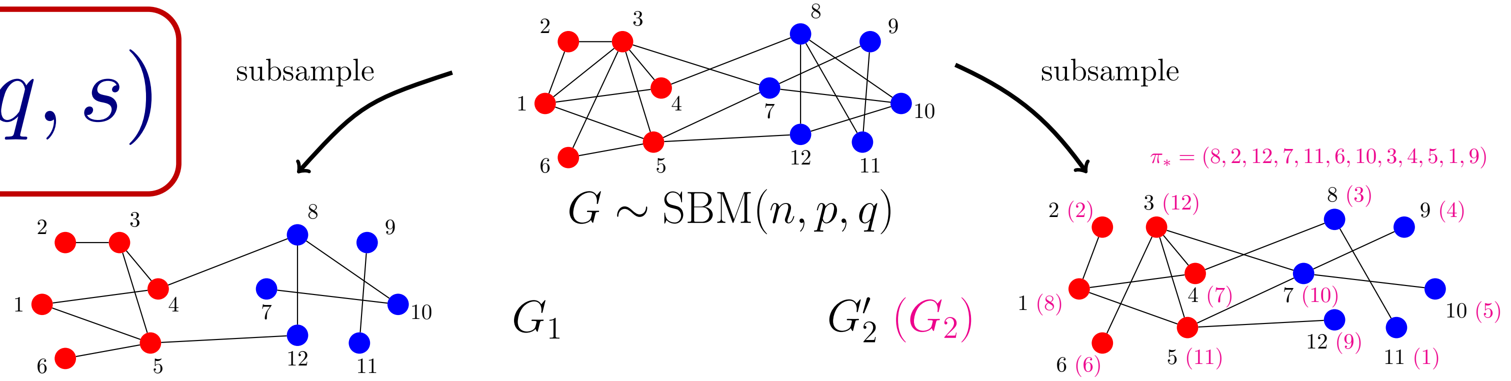
- in G_1 and G_2
- in G_1, not in G_2
- not in G_1, in G_2



$$G_1 \vee_{\pi_*} G_2 \sim \mathrm{SBM}\left(n, \frac{a(1-(1-s)^2)\log n}{n}, \frac{b(1-(1-s)^2)\log n}{n}\right)$$

Exact community recovery in correlated SBMs

$$\boxed{(G_1, G_2) \sim \mathrm{CSBM}(n, p, q, s)}$$



Since $G_1 \sim SBM(n, ps, qs)$, exact community recovery is possible from G_1 iff

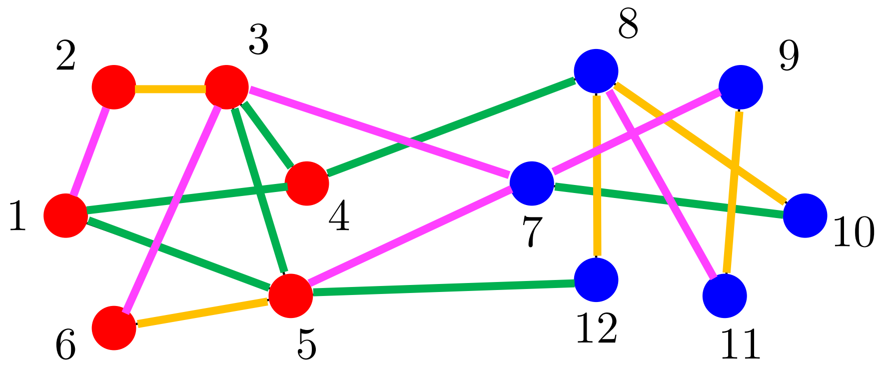
$$\boxed{\left|\sqrt{a} - \sqrt{b}\right| > \sqrt{2/s}}$$

How can we use both G_1 and G_2? **Suppose that π_* is known.** Then:

in G_1 and G_2

in G_1, not in G_2

not in G_1, in G_2



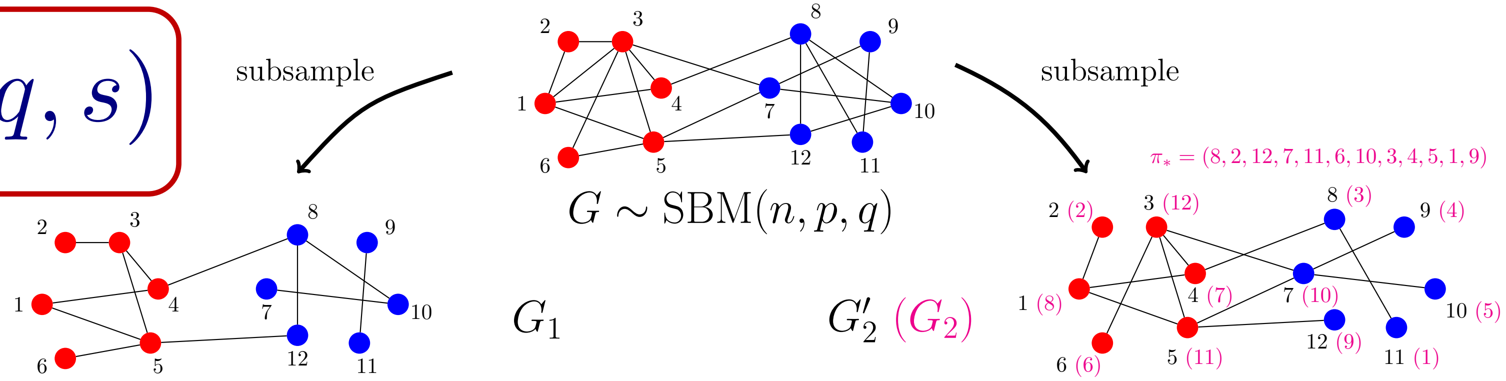
Thus exact community recovery is possible iff

$$\boxed{\left|\sqrt{a} - \sqrt{b}\right| > \sqrt{2/(1-(1-s)^2)}}$$

$$G_1 \vee_{\pi_*} G_2 \sim \mathrm{SBM}\left(n, \frac{a(1-(1-s)^2)\log n}{n}, \frac{b(1-(1-s)^2)\log n}{n}\right)$$

Exact community recovery in correlated SBMs

$$(G_1, G_2) \sim \mathrm{CSBM}(n, p, q, s)$$



Sin

H

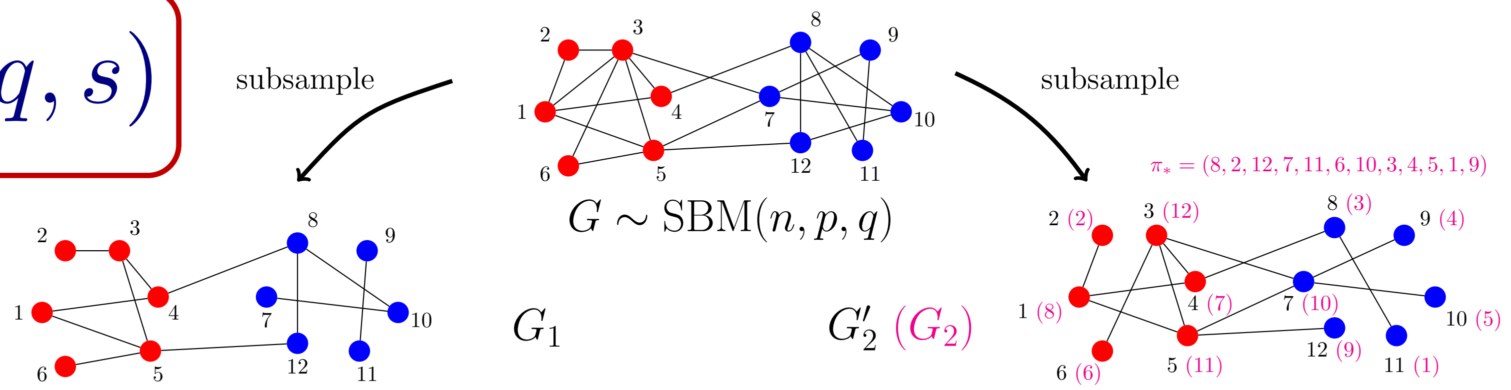
G_1

In particular, **if $\boldsymbol{\pi_*}$ is known** and $\boxed{\sqrt{2/s} > \left|\sqrt{a} - \sqrt{b}\right| > \sqrt{2/(1-(1-s)^2)}}$

then exact community recovery is possible from G_1 and G_2,
even though it is impossible from G_1 alone

Graph matching

$$(G_1, G_2) \sim \mathrm{CSBM}(n, p, q, s)$$

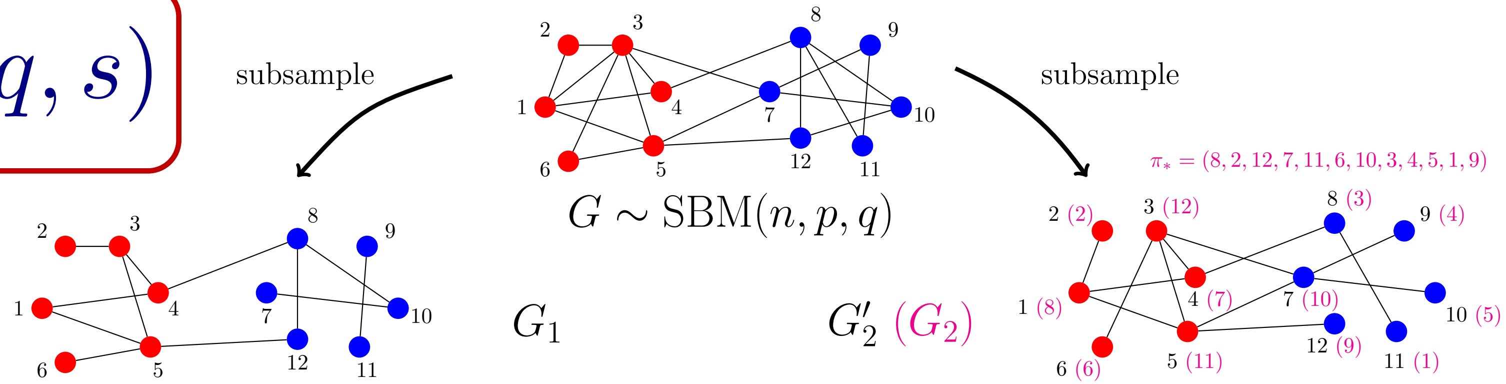


Main Q:

- given (G_1, G_2), when can we (exactly) recover the latent permutation π_*?

Graph matching

$$(G_1, G_2) \sim \mathrm{CSBM}(n, p, q, s)$$



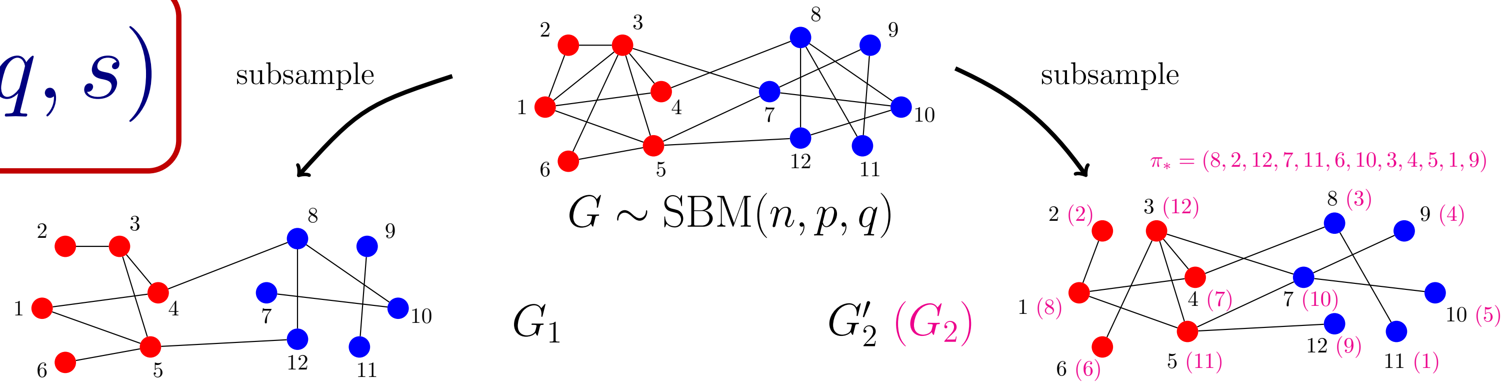
Main Q:

- given (G_1, G_2), when can we (exactly) recover the latent permutation π_*?

- Of significant independent interest

Graph matching

$$(G_1, G_2) \sim \mathrm{CSBM}(n, p, q, s)$$



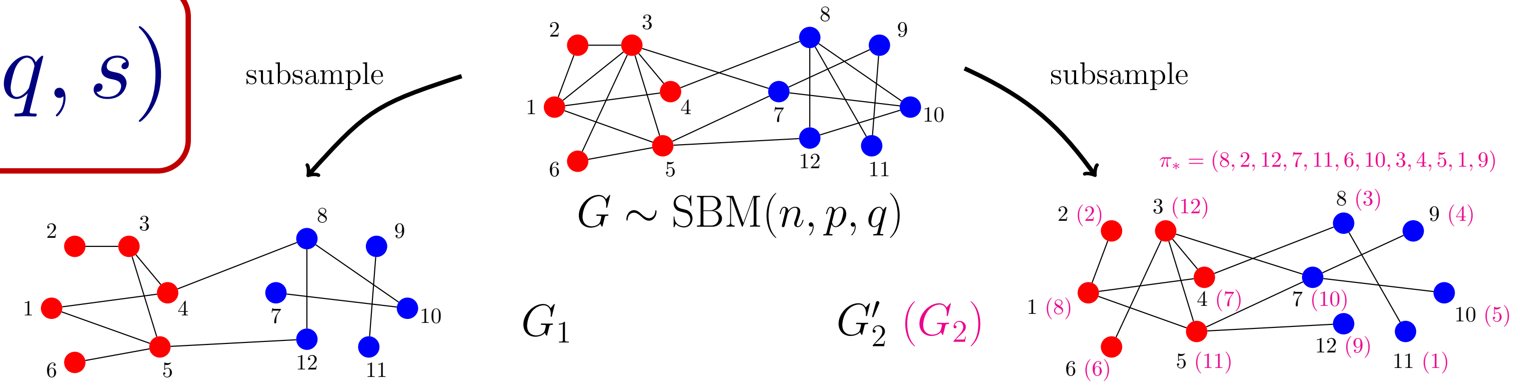
Main Q:

- given (G_1, G_2), when can we (exactly) recover the latent permutation π_*?

- Of significant independent interest
- Correlated Erdős-Rényi random graphs: Pedarsani, Grossglauser (2011)

Graph matching

$$(G_1, G_2) \sim \mathrm{CSBM}(n, p, q, s)$$



Main Q:

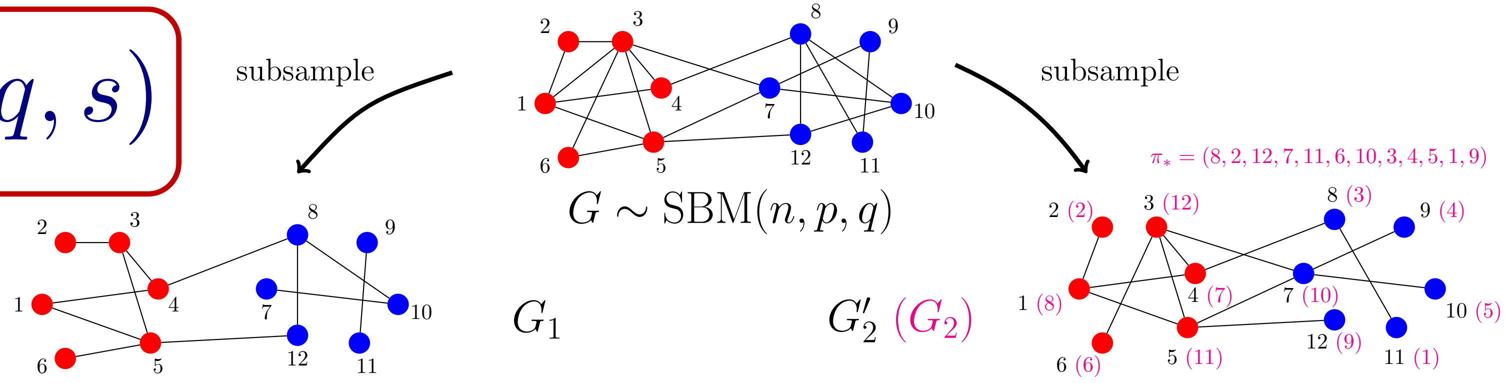
- given (G_1, G_2), when can we (exactly) recover the latent permutation π_*?

- Of significant independent interest
- Correlated Erdős-Rényi random graphs: Pedarsani, Grossglauser (2011)

- Many works in statistics/probability/CS/info theory… including:
  - Cullina, Kiyavash (2016, 2017)
  - Barak, Chou, Lei, Schramm, Sheng (2019)
  - Ding, Ma, Wu, Xu (2018)
  - Mossel, Xu (2019)
  - Fan, Mao, Wu, Xu (2019a,b)
  - Ganassali, Massoulié (2020)
  - Wu, Xu, Yu (2020, 2021)
  - Cullina, Kiyavash, Mittal, Poor (2020)
  - Mao, Rudelson, Tikhomirov (2021a,b)
  - Ganassali, Lelarge, Massoulié (2021)
  - Mao, Wu, Xu, Yu (2021,2022)
  - Ding, Du (2022a,b)

Correlated SBMs: graph matching and community recovery

$$(G_1, G_2) \sim \mathrm{CSBM}(n, p, q, s)$$



Main Q1 (community recovery):

- given (G_1, G_2), when can we (exactly) recover the communities?
- can we do so in regimes where it is impossible to do so using only G_1?

Main Q2 (graph matching):

- given (G_1, G_2), when can we (exactly) recover the latent permutation π_*?

Related work

Multi-layer networks/SBMs

- Holland, Laskey, Leinhardt (1983)
- Han, Xu, Airoldi (2015)
- Paul, Chen (2016, 2020a,b, 2021)
- Ali et al. (2019)
- Lei, Chen, Lynch (2019)
- Arroyo et al. (2020)
- Bhattacharyya, Chatterjee (2020)
- Chen, Liu, Ma (2020)
- …

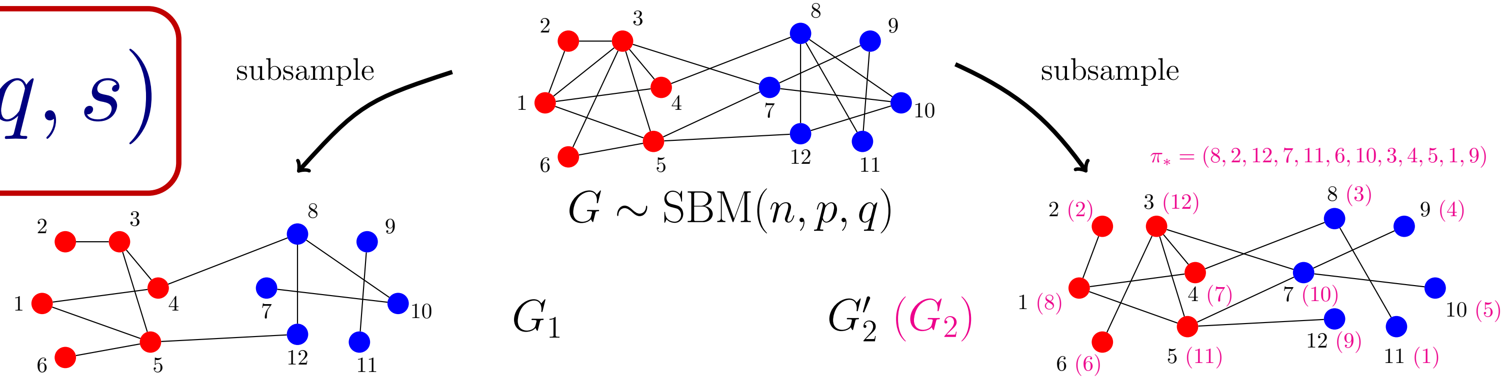
Contextual block models

- Kanade, Mossel, Schramm (2016)
- Mossel, Xu (2016)
- Zhang, Levina, Zhu (2016)
- Binkiewicz, Vogelstein, Rohe (2017)
- Deshpande, Sen, Montanari, Mossel (2018)
- Abbé, Fan, Wang (2020)
- Lu, Sen (2020)
- …

- Mayya, Reeves (2019)
- Ma, Nandy (2021)

Correlated SBMs: graph matching and community recovery

$$(G_1, G_2) \sim \mathrm{CSBM}(n, p, q, s)$$



Main Q1 (community recovery):

- given (G_1, G_2), when can we (exactly) recover the communities?
- can we do so in regimes where it is impossible to do so using only G_1?

Main Q2 (graph matching):

- given (G_1, G_2), when can we (exactly) recover the latent permutation π_*?

Results

Exact graph matching

Theorem (R., Sridhar, 2021)

Let $\hat{\pi}(G_1, G_2)$ be a vertex mapping that maximizes the number of agreeing edges between G_1 and G_2.

$$\widehat{\pi}(G_1, G_2) \in \arg\max_{\pi \in \mathcal{S}_n} \sum_{(i,j) \in \mathcal{E}} A_{i,j} B_{\pi(i),\pi(j)}$$

If $s^2 \left(\frac{a+b}{2} \right) > 1$ then $\lim_{n \to \infty} \mathbb{P}\left(\widehat{\pi}(G_1, G_2) = \pi_* \right) = 1$

Exact graph matching

Theorem (R., Sridhar, 2021)

Let $\hat{\pi}(G_1, G_2)$ be a vertex mapping that maximizes the number of agreeing edges between G_1 and G_2.

$$\widehat{\pi}(G_1, G_2) \in \arg\max_{\pi \in \mathcal{S}_n} \sum_{(i,j) \in \mathcal{E}} A_{i,j} B_{\pi(i),\pi(j)}$$

If $s^2 \left(\frac{a+b}{2} \right) > 1$ then $\lim_{n \to \infty} \mathbb{P}\left(\widehat{\pi}(G_1, G_2) = \pi_* \right) = 1$

- $\hat{\pi}$ is the MAP estimate for the correlated Erdős-Rényi model

Exact graph matching

Theorem (R., Sridhar, 2021)

Let $\hat{\pi}(G_1, G_2)$ be a vertex mapping that maximizes the number of agreeing edges between G_1 and G_2.

$$\widehat{\pi}(G_1, G_2) \in \arg\max_{\pi \in \mathcal{S}_n} \sum_{(i,j) \in \mathcal{E}} A_{i,j} B_{\pi(i),\pi(j)}$$

If $s^2 \left(\frac{a+b}{2} \right) > 1$ then $\lim_{n \to \infty} \mathbb{P}\left(\widehat{\pi}(G_1, G_2) = \pi_*\right) = 1$

- $\hat{\pi}$ is the MAP estimate for the correlated Erdős-Rényi model
- Cullina, Kiyavash (2016, 2017): exact graph matching for the correlated Erdős-Rényi model; see also Wu, Xu, Yu (2021)

Exact graph matching

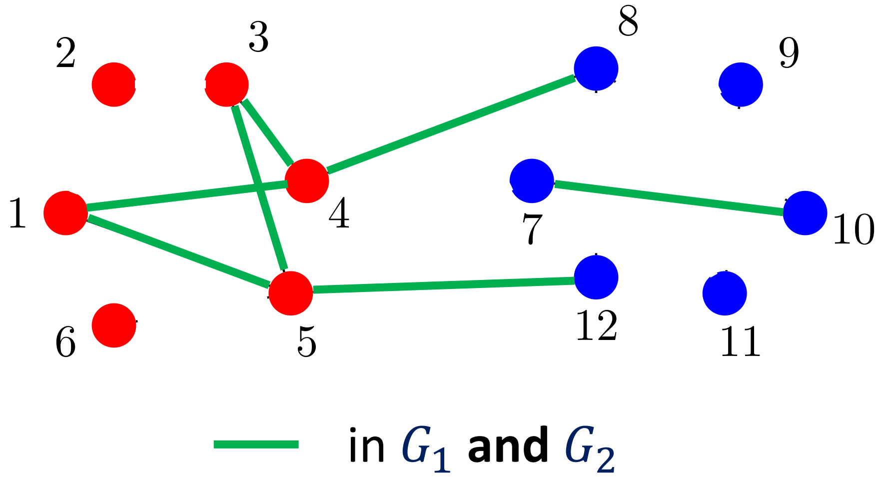
Theorem (R., Sridhar, 2021)

Let $\hat{\pi}(G_1, G_2)$ be a vertex mapping that maximizes the number of agreeing edges between G_1 and G_2.

$$\widehat{\pi}(G_1, G_2) \in \arg\max_{\pi \in \mathcal{S}_n} \sum_{(i,j) \in \mathcal{E}} A_{i,j} B_{\pi(i),\pi(j)}$$

If $\quad s^2 \left(\frac{a+b}{2} \right) > 1 \quad$ then $\quad \lim_{n \to \infty} \mathbb{P}\left(\widehat{\pi}(G_1, G_2) = \pi_* \right) = 1$

- $\hat{\pi}$ is the MAP estimate for the correlated Erdős-Rényi model
- Cullina, Kiyavash (2016, 2017): exact graph matching for the correlated Erdős-Rényi model; see also Wu, Xu, Yu (2021)
- Condition: the intersection graph is connected (whp)



Exact graph matching

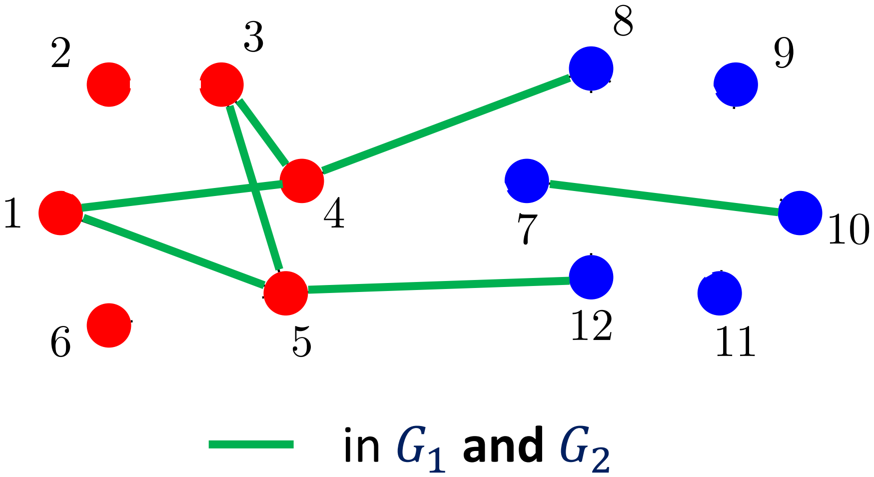
Theorem (R., Sridhar, 2021)

Let $\hat{\pi}(G_1, G_2)$ be a vertex mapping that maximizes the number of agreeing edges between G_1 and G_2.

$$\widehat{\pi}(G_1, G_2) \in \arg\max_{\pi \in \mathcal{S}_n} \sum_{(i,j) \in \mathcal{E}} A_{i,j} B_{\pi(i),\pi(j)}$$

If $\quad s^2 \left(\frac{a+b}{2} \right) > 1 \quad$ then $\quad \lim_{n \to \infty} \mathbb{P}\left(\widehat{\pi}(G_1, G_2) = \pi_* \right) = 1$

- $\hat{\pi}$ is the MAP estimate for the correlated Erdős-Rényi model
- Cullina, Kiyavash (2016, 2017): exact graph matching for the correlated Erdős-Rényi model; see also Wu, Xu, Yu (2021)
- Condition: the intersection graph is connected (whp)
- Onaran, Garg, Erkip (2016): same conclusion under stronger parameter assumptions and assuming all community labels are known



Exact graph matching – converse

Theorem (Cullina, Singhal, Kiyavash, Mittal, 2016)

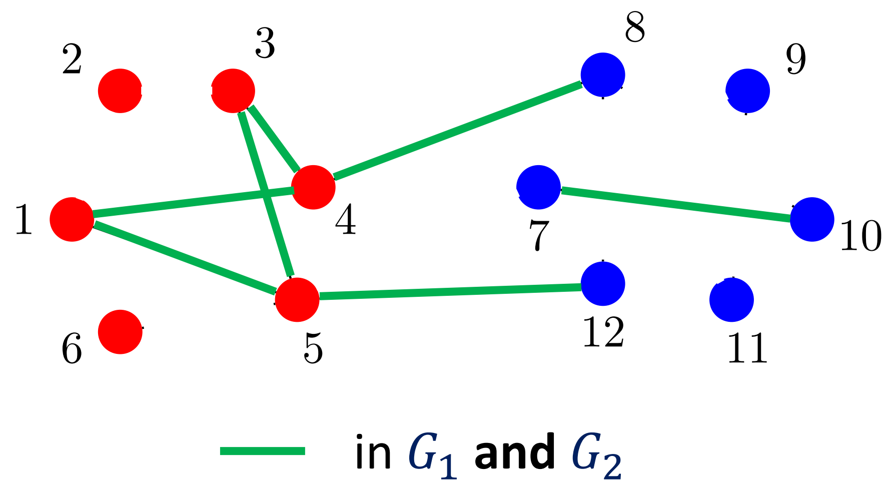
If $s^2\left(\frac{a+b}{2}\right) < 1$ then $\lim_{n\to\infty} \mathbb{P}\left(\widetilde{\pi}(G_1, G_2) = \pi_*\right) = 0$ for every estimator $\tilde{\pi}$

Exact graph matching – converse

Theorem (Cullina, Singhal, Kiyavash, Mittal, 2016)

If $\quad s^2 \left(\frac{a+b}{2} \right) < 1 \quad$ then $\quad \lim_{n \to \infty} \mathbb{P}\left(\widetilde{\pi}(G_1, G_2) = \pi_* \right) = 0 \quad$ for every estimator $\tilde{\pi}$

- Condition: the intersection graph is disconnected (whp)

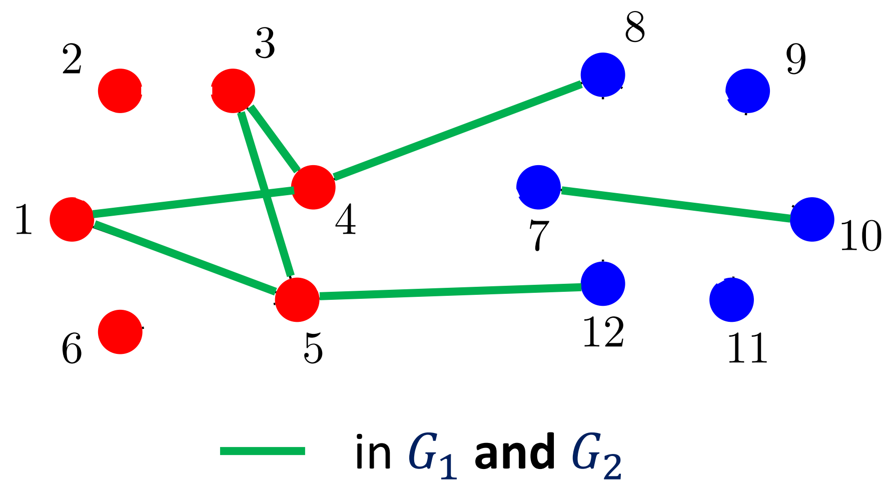


Exact graph matching – converse

Theorem (Cullina, Singhal, Kiyavash, Mittal, 2016)

If $s^2\left(\frac{a+b}{2}\right) < 1$ then $\lim_{n\to\infty} \mathbb{P}\left(\widetilde{\pi}(G_1, G_2) = \pi_*\right) = 0$ for every estimator $\tilde{\pi}$

- Condition: the intersection graph is disconnected (whp)
- In particular: the intersection graph has many *isolated vertices*

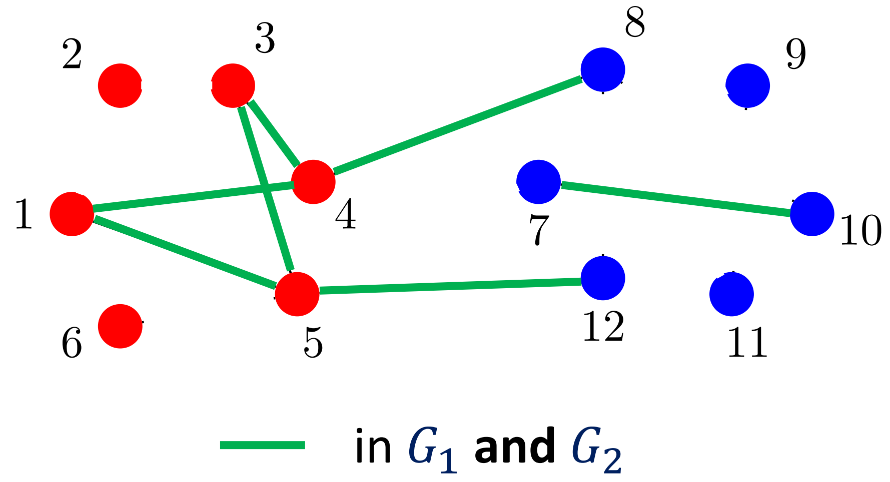


Exact graph matching – converse

Theorem (Cullina, Singhal, Kiyavash, Mittal, 2016)

If $\quad s^2\left(\frac{a+b}{2}\right) < 1 \quad$ then $\quad \lim_{n\to\infty} \mathbb{P}\left(\widetilde{\pi}(G_1, G_2) = \pi_*\right) = 0 \quad$ for every estimator $\tilde{\pi}$

- Condition: the intersection graph is disconnected (whp)
- In particular: the intersection graph has many *isolated vertices*
- These vertices have *non-overlapping neighborhoods* in G_1 and G_2'

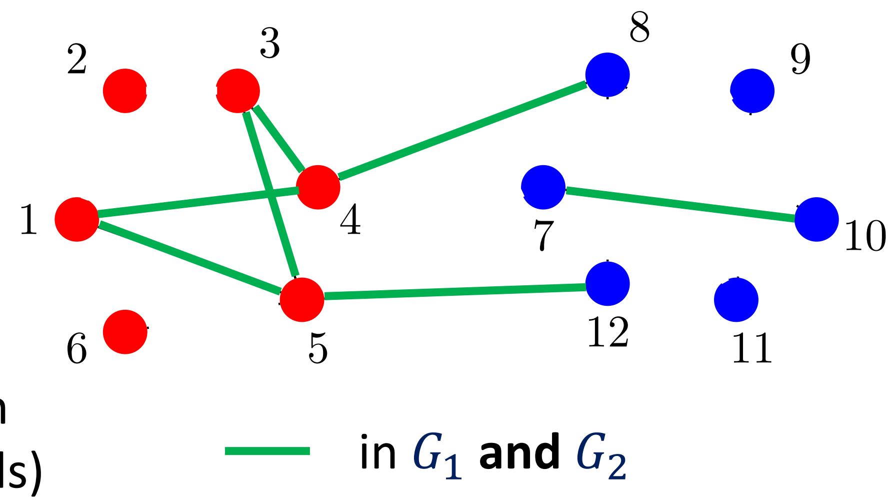


Exact graph matching – converse

Theorem (Cullina, Singhal, Kiyavash, Mittal, 2016)

If $s^2\left(\frac{a+b}{2}\right) < 1$ then $\lim_{n\to\infty} \mathbb{P}\left(\widetilde{\pi}(G_1, G_2) = \pi_*\right) = 0$ for every estimator $\tilde{\pi}$

- Condition: the intersection graph is disconnected (whp)
- In particular: the intersection graph has many *isolated vertices*
- These vertices have *non-overlapping neighborhoods* in G_1 and G_2'
- Such vertices are hard to match due to the lack of shared information (even for optimal estimators that have access to the community labels)



Exact community recovery

Theorem (R., Sridhar, 2021)

Exact community recovery is **possible**

If $s^2\left(\frac{a+b}{2}\right) > 1$ and $|\sqrt{a} - \sqrt{b}| > \sqrt{2/(1-(1-s)^2)}$

then there is an estimator $\widehat{\boldsymbol{\sigma}} = \widehat{\boldsymbol{\sigma}}(G_1, G_2)$ such that $\boxed{\lim_{n\to\infty} \mathbb{P}(\mathrm{ov}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = 1) = 1}$

Exact community recovery

Theorem (R., Sridhar, 2021)

Exact community recovery is **possible**

If $s^2\left(\frac{a+b}{2}\right) > 1$ and $\left|\sqrt{a}-\sqrt{b}\right| > \sqrt{2/(1-(1-s)^2)}$

then there is an estimator $\widehat{\boldsymbol{\sigma}} = \widehat{\boldsymbol{\sigma}}(G_1, G_2)$ such that $\lim_{n\to\infty} \mathbb{P}(\mathrm{ov}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = 1) = 1$

Proof: can recover π_* whp; then run a community recovery algorithm on the union of the matched graphs.

Exact community recovery

Theorem (R., Sridhar, 2021)

Exact community recovery is **possible**

If $s^2 \left(\frac{a+b}{2} \right) > 1$ and $\left| \sqrt{a} - \sqrt{b} \right| > \sqrt{2/(1-(1-s)^2)}$

then there is an estimator $\widehat{\boldsymbol{\sigma}} = \widehat{\boldsymbol{\sigma}}(G_1, G_2)$ such that $\lim_{n \to \infty} \mathbb{P}(\mathrm{ov}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = 1) = 1$

Proof: can recover π_* whp; then run a community recovery algorithm on the union of the matched graphs.

Theorem (R., Sridhar, 2021)

Exact community recovery is **impossible**

If $\left| \sqrt{a} - \sqrt{b} \right| < \sqrt{2/(1-(1-s)^2)}$

then for any estimator $\widetilde{\boldsymbol{\sigma}} = \widetilde{\boldsymbol{\sigma}}(G_1, G_2)$ we have that $\lim_{n \to \infty} \mathbb{P}(\mathrm{ov}(\widetilde{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = 1) = 0$

Exact community recovery

Theorem (R., Sridhar, 2021)

Exact community recovery is **possible**

If $s^2\left(\frac{a+b}{2}\right) > 1$ and $|\sqrt{a}-\sqrt{b}| > \sqrt{2/(1-(1-s)^2)}$

then there is an estimator $\widehat{\boldsymbol{\sigma}} = \widehat{\boldsymbol{\sigma}}(G_1, G_2)$ such that $\lim_{n\to\infty} \mathbb{P}(\mathrm{ov}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = 1) = 1$

Proof: can recover π_* whp; then run a community recovery algorithm on the union of the matched graphs.

Theorem (R., Sridhar, 2021)

Exact community recovery is **impossible**

If $|\sqrt{a}-\sqrt{b}| < \sqrt{2/(1-(1-s)^2)}$

then for any estimator $\widetilde{\boldsymbol{\sigma}} = \widetilde{\boldsymbol{\sigma}}(G_1, G_2)$ we have that $\lim_{n\to\infty} \mathbb{P}(\mathrm{ov}(\widetilde{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = 1) = 0$

Proof: even if π_* is known, it is impossible to exactly recover the communities from $G_1 \vee_{\pi_*} G_2$

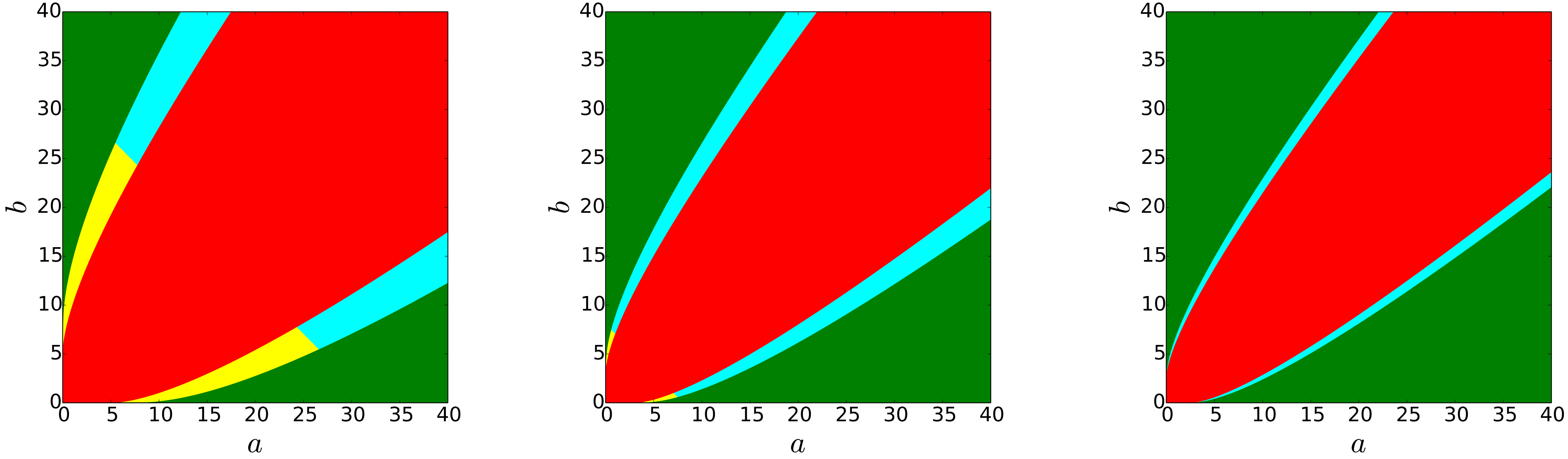
Phase diagrams

Exact community recovery possible from G_1

Exact community recovery impossible from (G_1, G_2)

Exact community recovery impossible from G_1, possible from (G_1, G_2)

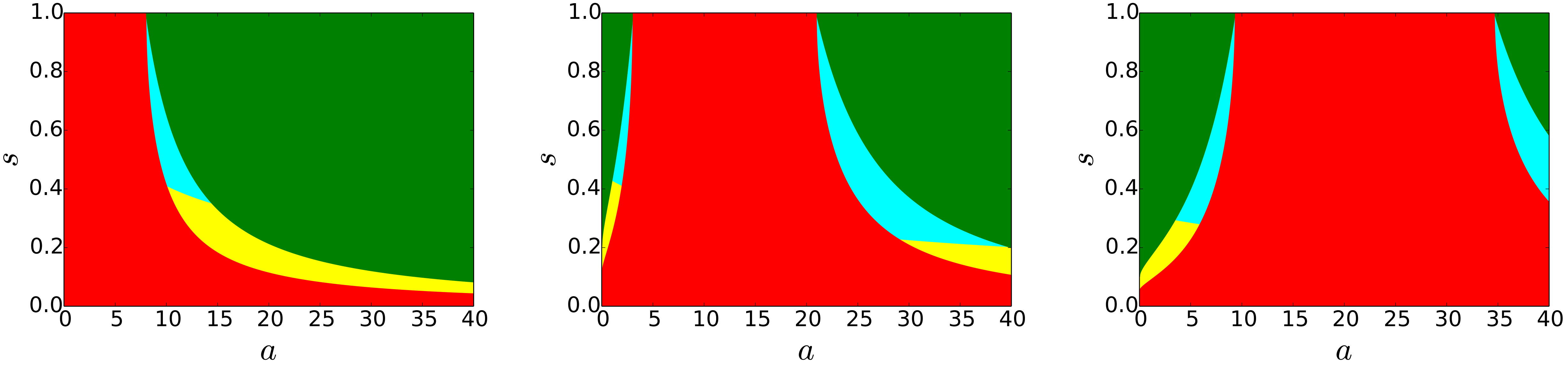
Exact community recovery impossible from G_1, exact recovery of π_* impossible



$s = 0.25$ $s = 0.5$ $s = 0.75$

Phase diagrams

- Exact community recovery possible from G_1
- Exact community recovery impossible from G_1, possible from (G_1, G_2)
- Exact community recovery impossible from (G_1, G_2)
- Exact community recovery impossible from G_1, exact recovery of π_* impossible



$b = 2$ $\qquad$ $b = 10$ $\qquad$ $b = 20$

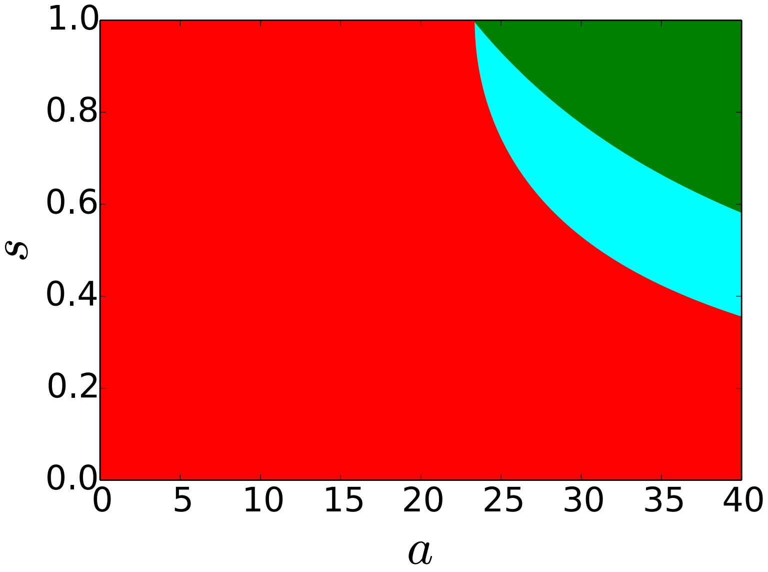
Phase diagrams

Exact community recovery possible from G_1

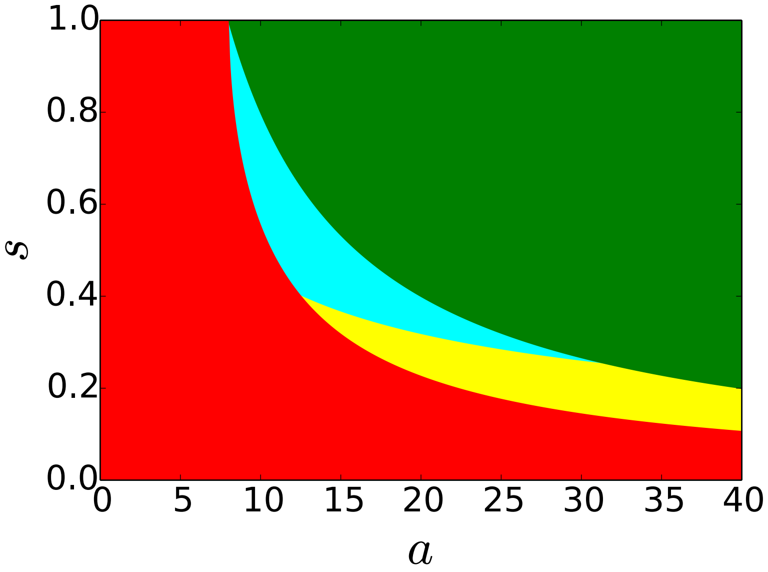
Exact community recovery impossible from (G_1, G_2)

Exact community recovery impossible from G_1, possible from (G_1, G_2)

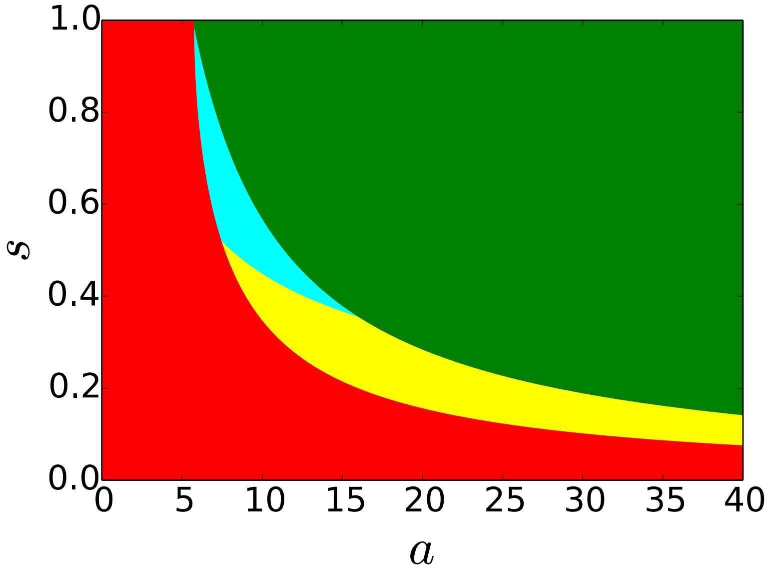
Exact community recovery impossible from G_1, exact recovery of π_* impossible



$a/b = 2$



$a/b = 4$



$a/b = 6$

Proof (graph matching)

Estimator

A, B: adjacency matrices of G_1, G_2

$$\widehat{\pi}(G_1, G_2) \in \arg\max_{\pi \in \mathcal{S}_n} \sum_{(i,j) \in \mathcal{E}} A_{i,j} B_{\pi(i), \pi(j)}$$

Estimator

A, B: adjacency matrices of G_1, G_2

$$\widehat{\pi}(G_1, G_2) \in \arg\max_{\pi \in \mathcal{S}_n} \sum_{(i,j) \in \mathcal{E}} A_{i,j} B_{\pi(i),\pi(j)}$$

$$\widehat{\pi}(G_1, G_2) \in \arg\max_{\pi \in \mathcal{S}_n} \sum_{e \in \mathcal{E}} A_e B_{\tau(e)}$$

Permutation $\pi \in \mathcal{S}_n$ on vertices

$\Updownarrow$ $\tau = \ell(\pi)$

Lifted permutation $\tau: \mathcal{E} \to \mathcal{E}$ on vertex pairs

Estimator

A, B: adjacency matrices of G_1, G_2

$$\widehat{\pi}(G_1, G_2) \in \arg\max_{\pi \in \mathcal{S}_n} \sum_{(i,j) \in \mathcal{E}} A_{i,j} B_{\pi(i),\pi(j)}$$

Permutation $\pi \in \mathcal{S}_n$ on vertices

$$\Updownarrow \qquad \tau = \ell(\pi)$$

$$\widehat{\pi}(G_1, G_2) \in \arg\max_{\pi \in \mathcal{S}_n} \sum_{e \in \mathcal{E}} A_e B_{\tau(e)}$$

Lifted permutation $\tau\colon \mathcal{E} \to \mathcal{E}$ on vertex pairs

$$X(\tau) := \sum_{e \in \mathcal{E}} A_e B_{\tau_*(e)} - \sum_{e \in \mathcal{E}} A_e B_{\tau(e)} = \sum_{e \in \mathcal{E} : \tau(e) \neq \tau_*(e)} \left(A_e B_{\tau_*(e)} - A_e B_{\tau(e)} \right)$$

Estimator

A, B: adjacency matrices of G_1, G_2

$$\widehat{\pi}(G_1, G_2) \in \arg\max_{\pi \in \mathcal{S}_n} \sum_{(i,j)\in\mathcal{E}} A_{i,j} B_{\pi(i),\pi(j)}$$

Permutation $\pi \in \mathcal{S}_n$ on vertices

$\Updownarrow$ $\quad \tau = \ell(\pi)$

Lifted permutation $\tau: \mathcal{E} \to \mathcal{E}$ on vertex pairs

$$\widehat{\pi}(G_1, G_2) \in \arg\max_{\pi \in \mathcal{S}_n} \sum_{e\in\mathcal{E}} A_e B_{\tau(e)}$$

$$X(\tau) := \sum_{e\in\mathcal{E}} A_e B_{\tau_*(e)} - \sum_{e\in\mathcal{E}} A_e B_{\tau(e)} = \sum_{e\in\mathcal{E}:\tau(e)\neq\tau_*(e)} \left(A_e B_{\tau_*(e)} - A_e B_{\tau(e)}\right)$$

If $X(\tau) > 0$ for every $\tau \neq \tau_*$, then $\hat{\pi} = \pi_*$

Permutations

Let S_{k_1,k_2} denote the set of lifted permutations such that

- k_1 vertices are mismatched in V_+
- k_2 vertices are mismatched in V_-

(relative to π_*)

Permutations

Let S_{k_1,k_2} denote the set of lifted permutations such that

- k_1 vertices are mismatched in V_+
- k_2 vertices are mismatched in V_-

(relative to π_*)

From vertex mismatches to edge mismatches:

$$M^+(\tau) := \left|\left\{e \in \mathcal{E}^+(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)\right\}\right|$$
$$M^-(\tau) := \left|\left\{e \in \mathcal{E}^-(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)\right\}\right|$$

Permutations

Let S_{k_1,k_2} denote the set of lifted permutations such that

- k_1 vertices are mismatched in V_+
- k_2 vertices are mismatched in V_-

(relative to π_*)

From vertex mismatches to edge mismatches:

$$M^+(\tau) := \left|\left\{e \in \mathcal{E}^+(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)\right\}\right|$$
$$M^-(\tau) := \left|\left\{e \in \mathcal{E}^-(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)\right\}\right|$$

Assume that the communities are approximately balanced (this happens whp).

$$\mathcal{F}_\epsilon := \left\{\left(1 - \frac{\epsilon}{2}\right)\frac{n}{2} \leq |V_+|, |V_-| \leq \left(1 + \frac{\epsilon}{2}\right)\frac{n}{2}\right\}$$

Lemma

When $k_1 \leq \frac{\epsilon}{2}|V_+|$ and $k_2 \leq \frac{\epsilon}{2}|V_-|$:

$$M^+(\tau) \geq (1-\epsilon)\frac{n}{2}(k_1 + k_2),$$
$$M^-(\tau) \geq (1-\epsilon)\frac{n}{2}(k_1 + k_2).$$

Permutations

Let S_{k_1,k_2} denote the set of lifted permutations such that

- k_1 vertices are mismatched in V_+
- k_2 vertices are mismatched in V_-

(relative to π_*)

From vertex mismatches to edge mismatches:

$$M^+(\tau) := \left|\left\{e \in \mathcal{E}^+(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)\right\}\right|$$
$$M^-(\tau) := \left|\left\{e \in \mathcal{E}^-(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)\right\}\right|$$

Assume that the communities are approximately balanced (this happens whp).

$$\mathcal{F}_\epsilon := \left\{\left(1 - \frac{\epsilon}{2}\right)\frac{n}{2} \leq |V_+|, |V_-| \leq \left(1 + \frac{\epsilon}{2}\right)\frac{n}{2}\right\}$$

Lemma

When $k_1 \leq \frac{\epsilon}{2}|V_+|$ and $k_2 \leq \frac{\epsilon}{2}|V_-|$:

$$M^+(\tau) \geq (1-\epsilon)\frac{n}{2}(k_1 + k_2),$$
$$M^-(\tau) \geq (1-\epsilon)\frac{n}{2}(k_1 + k_2).$$

In general:

$$M^+(\tau) \geq (1-\epsilon)\frac{n}{4}(k_1 + k_2),$$
$$M^-(\tau) \geq (1-\epsilon)\frac{n}{4}(k_1 + k_2).$$

Analysis

Claim

If $s^2\left(\frac{a+b}{2}\right) > 1$ then there exists $\delta > 0$ such that

$$\boxed{\mathbb{P}\left(\widehat{\tau} \in S_{k_1,k_2} \,|\, \boldsymbol{\sigma}, \tau_*\right) \mathbf{1}(\mathcal{F}_\epsilon) \leq n^{-\delta(k_1+k_2)}.}$$

Analysis

Claim

If $s^2\left(\frac{a+b}{2}\right) > 1$ then there exists $\delta > 0$ such that

$$\boxed{\mathbb{P}\left(\widehat{\tau} \in S_{k_1,k_2} \,|\, \boldsymbol{\sigma}, \tau_*\right) \mathbf{1}(\mathcal{F}_\epsilon) \leq n^{-\delta(k_1+k_2)}.}$$

Proof sketch:

- Union bound gives factor of $|S_{k_1,k_2}| \leq n^{k_1+k_2}$

Analysis

Claim

If $s^2\left(\frac{a+b}{2}\right) > 1$ then there exists $\delta > 0$ such that

$$\boxed{\mathbb{P}\left(\widehat{\tau} \in S_{k_1,k_2} \,|\, \boldsymbol{\sigma}, \tau_*\right) \mathbf{1}(\mathcal{F}_\epsilon) \leq n^{-\delta(k_1+k_2)}.}$$

Proof sketch:

- Union bound gives factor of $\left|S_{k_1,k_2}\right| \leq n^{k_1+k_2}$
- Individual bound boils down to bounds on the probability-generating function:

$$\mathbb{P}\left(\widehat{\tau} = \tau \,|\, \boldsymbol{\sigma}, \tau_*\right) \leq \mathbb{P}\left(X(\tau) \leq 0 \,|\, \boldsymbol{\sigma}, \tau_*\right) = \mathbb{P}\left(n^{-X(\tau)/2} \geq 1 \,\middle|\, \boldsymbol{\sigma}, \tau_*\right)$$
$$\leq \mathbb{E}\left[\left(1/\sqrt{n}\right)^{X(\tau)} \,\middle|\, \boldsymbol{\sigma}, \tau_*\right]$$

Analysis

Claim

If $s^2\left(\frac{a+b}{2}\right) > 1$ then there exists $\delta > 0$ such that

$$\boxed{\mathbb{P}\left(\widehat{\tau} \in S_{k_1,k_2} \,\middle|\, \boldsymbol{\sigma}, \tau_*\right) \mathbf{1}(\mathcal{F}_\epsilon) \le n^{-\delta(k_1+k_2)}.}$$

Proof sketch:

- Union bound gives factor of $\left|S_{k_1,k_2}\right| \le n^{k_1+k_2}$
- Individual bound boils down to bounds on the probability-generating function:

$$\begin{aligned}
\mathbb{P}\left(\widehat{\tau} = \tau \,\middle|\, \boldsymbol{\sigma}, \tau_*\right) \le \mathbb{P}\left(X(\tau) \le 0 \,\middle|\, \boldsymbol{\sigma}, \tau_*\right) &= \mathbb{P}\left(n^{-X(\tau)/2} \ge 1 \,\middle|\, \boldsymbol{\sigma}, \tau_*\right) \\
&\le \mathbb{E}\left[\left(1/\sqrt{n}\right)^{X(\tau)} \,\middle|\, \boldsymbol{\sigma}, \tau_*\right] \\
&\le \exp\left(-(1-\epsilon)s^2\left(aM^+(\tau) + bM^-(\tau)\right)\frac{\log n}{n}\right)
\end{aligned}$$

Generating function

$$M^+(\tau) := \left|\left\{e \in \mathcal{E}^+(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)\right\}\right|,$$
$$M^-(\tau) := \left|\left\{e \in \mathcal{E}^-(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)\right\}\right|,$$
$$Y^+(\tau) := \sum_{e \in \mathcal{E}^+(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)} A_e B_{\tau_*(e)},$$
$$Y^-(\tau) := \sum_{e \in \mathcal{E}^-(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)} A_e B_{\tau_*(e)}.$$

Joint generating function

$$\boxed{\Phi^\tau(\theta, \omega, \zeta) := \mathbb{E}\left[\theta^{X(\tau)} \omega^{Y^+(\tau)} \zeta^{Y^-(\tau)} \,\middle|\, \boldsymbol{\sigma}, \tau_*\right]}$$

The PGF of only $X(\tau)$ only works when $s^2(a+b)/2 > 2$

Generating function

$$M^+(\tau) := \left|\{e \in \mathcal{E}^+(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)\}\right|,$$
$$M^-(\tau) := \left|\{e \in \mathcal{E}^-(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)\}\right|,$$
$$Y^+(\tau) := \sum_{e \in \mathcal{E}^+(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)} A_e B_{\tau_*(e)},$$
$$Y^-(\tau) := \sum_{e \in \mathcal{E}^-(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)} A_e B_{\tau_*(e)}.$$

Joint generating function

$$\boxed{\Phi^\tau(\theta, \omega, \zeta) := \mathbb{E}\left[\theta^{X(\tau)} \omega^{Y^+(\tau)} \zeta^{Y^-(\tau)} \,\middle|\, \boldsymbol{\sigma}, \tau_*\right]}$$

The PGF of only $X(\tau)$ only works when $s^2(a+b)/2 > 2$

Lemma

For any $\varepsilon \in (0,1)$ and $1 \leq \omega, \zeta \leq 3$, and for all n large enough:

$$\Phi^\tau\left(1/\sqrt{n}, \omega, \zeta\right) \leq \exp\left(-(1-\epsilon)s^2\left(\alpha M^+(\tau) + \beta M^-(\tau)\right)\frac{\log n}{n}\right)$$

Generating function

$$M^+(\tau) := \left|\left\{e \in \mathcal{E}^+(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)\right\}\right|,$$
$$M^-(\tau) := \left|\left\{e \in \mathcal{E}^-(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)\right\}\right|,$$
$$Y^+(\tau) := \sum_{e \in \mathcal{E}^+(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)} A_e B_{\tau_*(e)},$$
$$Y^-(\tau) := \sum_{e \in \mathcal{E}^-(\boldsymbol{\sigma}) : \tau(e) \neq \tau_*(e)} A_e B_{\tau_*(e)}.$$

Joint generating function

$$\boxed{\Phi^\tau(\theta, \omega, \zeta) := \mathbb{E}\left[\theta^{X(\tau)} \omega^{Y^+(\tau)} \zeta^{Y^-(\tau)} \,\middle|\, \boldsymbol{\sigma}, \tau_*\right]}$$

The PGF of only $X(\tau)$ only works when $s^2(a+b)/2 > 2$

Analysis:

- Decompose according to cycles of $\tau_*^{-1} \circ \tau$; independence across cycles
- For correlated Erdős-Rényi: explicit formulas
- For correlated SBM: recursive bounds

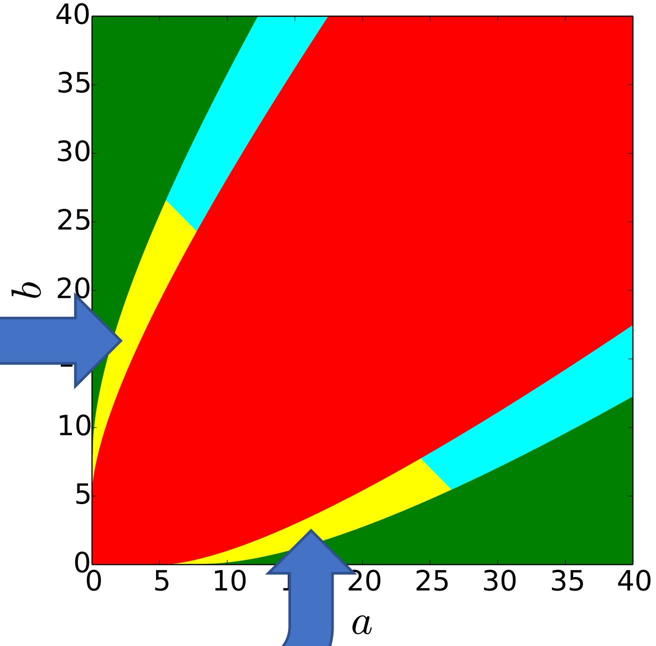
Lemma

For any $\varepsilon \in (0,1)$ and $1 \leq \omega, \zeta \leq 3$, and for all n large enough:

$$\Phi^\tau\left(1/\sqrt{n}, \omega, \zeta\right) \leq \exp\left(-(1-\epsilon)s^2\left(\alpha M^+(\tau) + \beta M^-(\tau)\right)\frac{\log n}{n}\right)$$

The interplay between community recovery and graph matching

Closing the gap for exact community recovery



$s = 0.25$

What happens in this region?

- Exact community recovery possible from G_1
- Exact community recovery impossible from (G_1, G_2)
- Exact community recovery impossible from G_1, possible from (G_1, G_2)
- Exact community recovery impossible from G_1, exact recovery of π_* impossible

- Exact community recovery is impossible from G_1
- Exact graph matching is impossible
- **Q:** is exact community recovery from (G_1, G_2) possible?

Interplay btw community recovery and graph matching

Theorem (Gaudio, R., Sridhar, 2022)

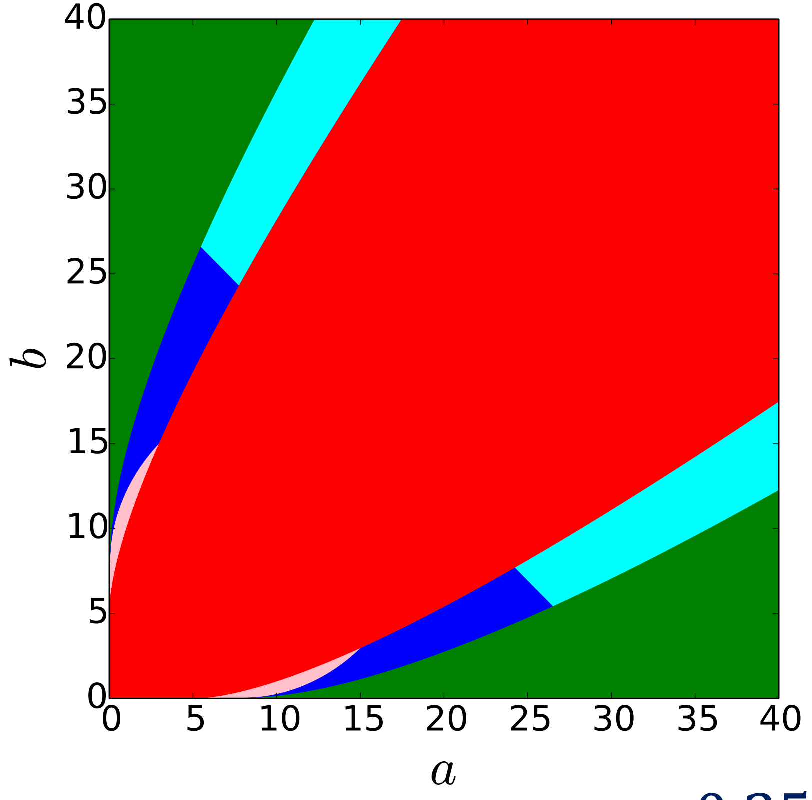
In the regime where $\left|\sqrt{a}-\sqrt{b}\right| > \sqrt{2/(1-(1-s)^2)}$,

the threshold for exact community recovery is given by:

$$\boxed{s^2\left(\frac{a+b}{2}\right) + s(1-s)\left(\frac{\sqrt{a}-\sqrt{b}}{\sqrt{2}}\right)^2 = 1}$$

graph matching

community recovery



$s = 0.25$

Phase diagrams

Exact graph matching:

possible

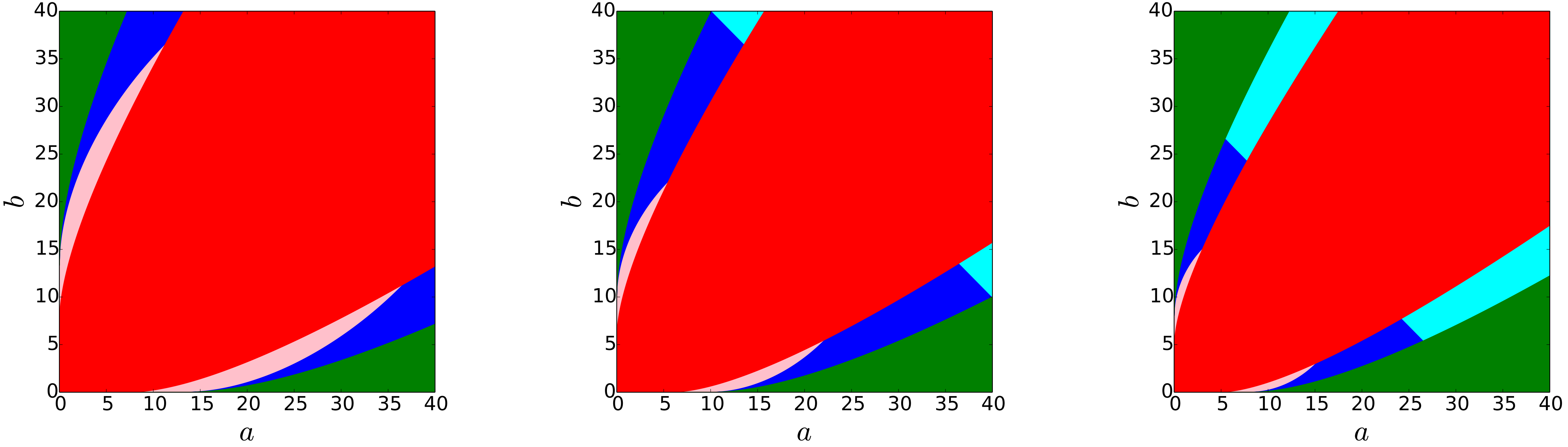
impossible

Exact community recovery possible from G_1

Exact community recovery impossible from G_1, possible from (G_1, G_2)

Exact community recovery impossible from (G_1, G_2)

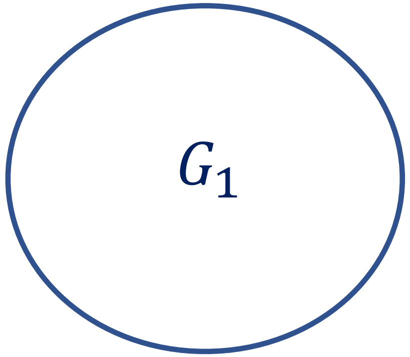
(though possible if π_* were known)



$s = 0.15$ $s = 0.20$ $s = 0.25$

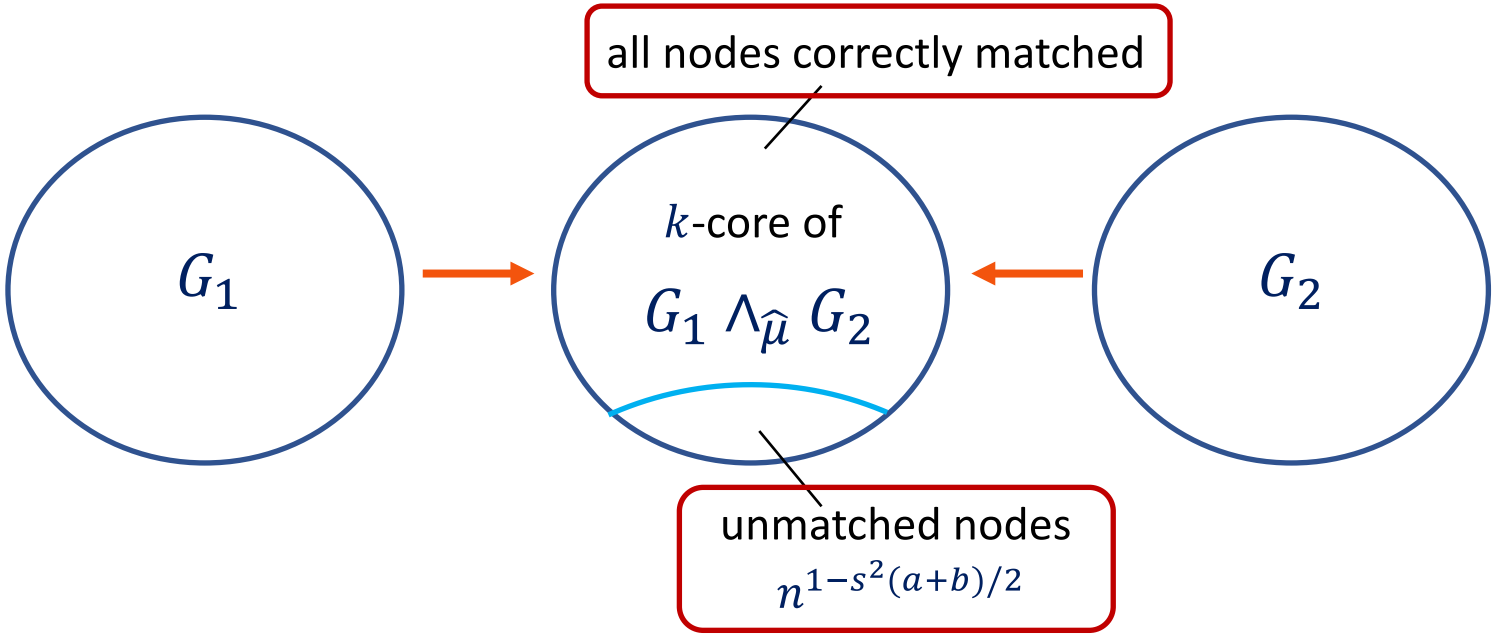
Algorithm

1. Almost exact labeling of G_1
   [Mossel, Neeman, Sly, 2014]



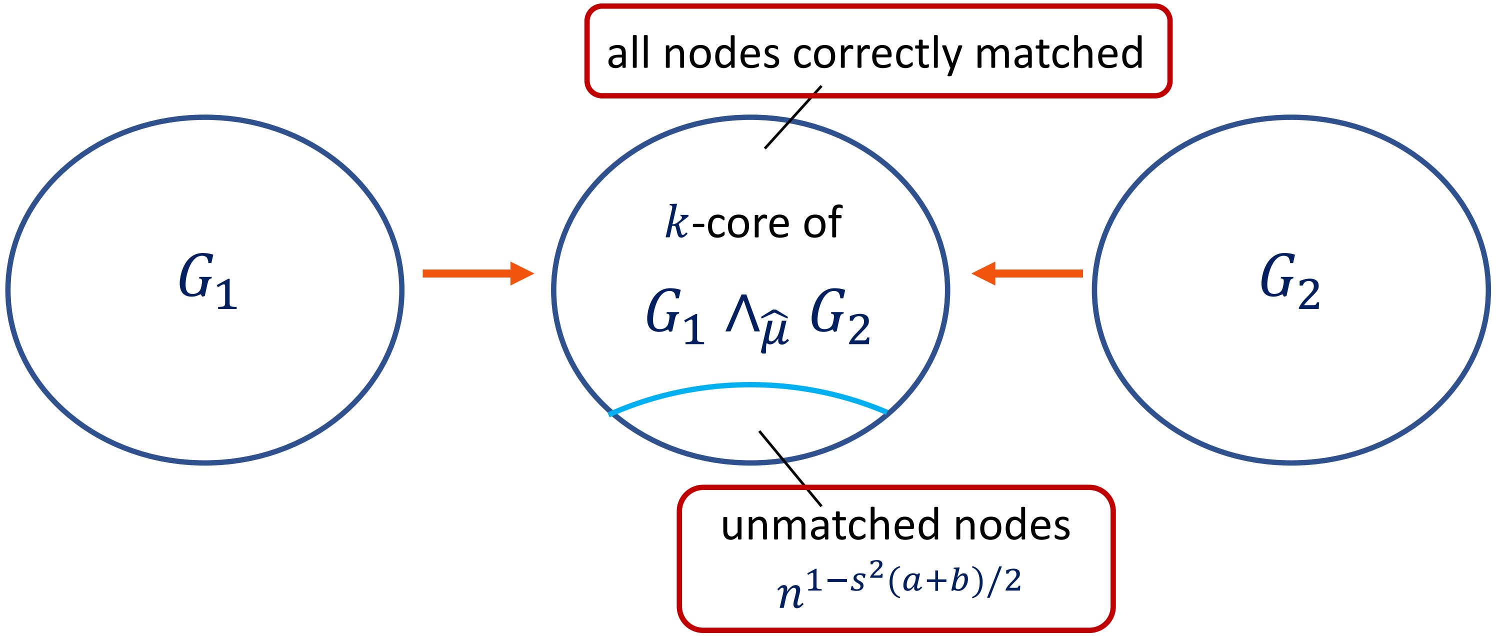
Algorithm

1. Almost exact labeling of G_1
   [Mossel, Neeman, Sly, 2014]
2. Partial almost exact graph matching $\hat{\mu}$
   [Cullina, Kiyavash, Mittal, Poor, 2020]



Algorithm

1. Almost exact labeling of G_1
   [Mossel, Neeman, Sly, 2014]
2. Partial almost exact graph matching $\hat{\mu}$
   [Cullina, Kiyavash, Mittal, Poor, 2020]

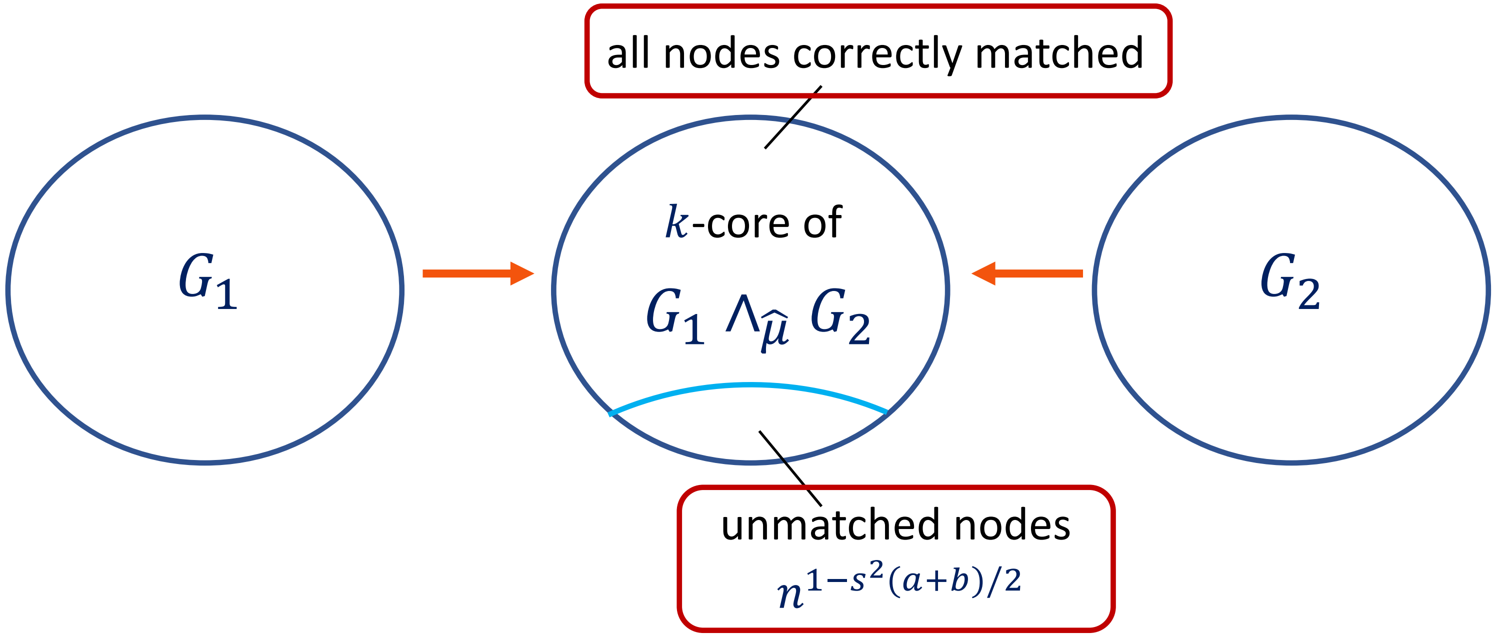


Remarks on the k-core estimator:

- Works well for correlated inhomogeneous random graphs [R., Sridhar, 2023]
- Closely related to densest subgraph estimator [Ding, Du, 2022a,b]

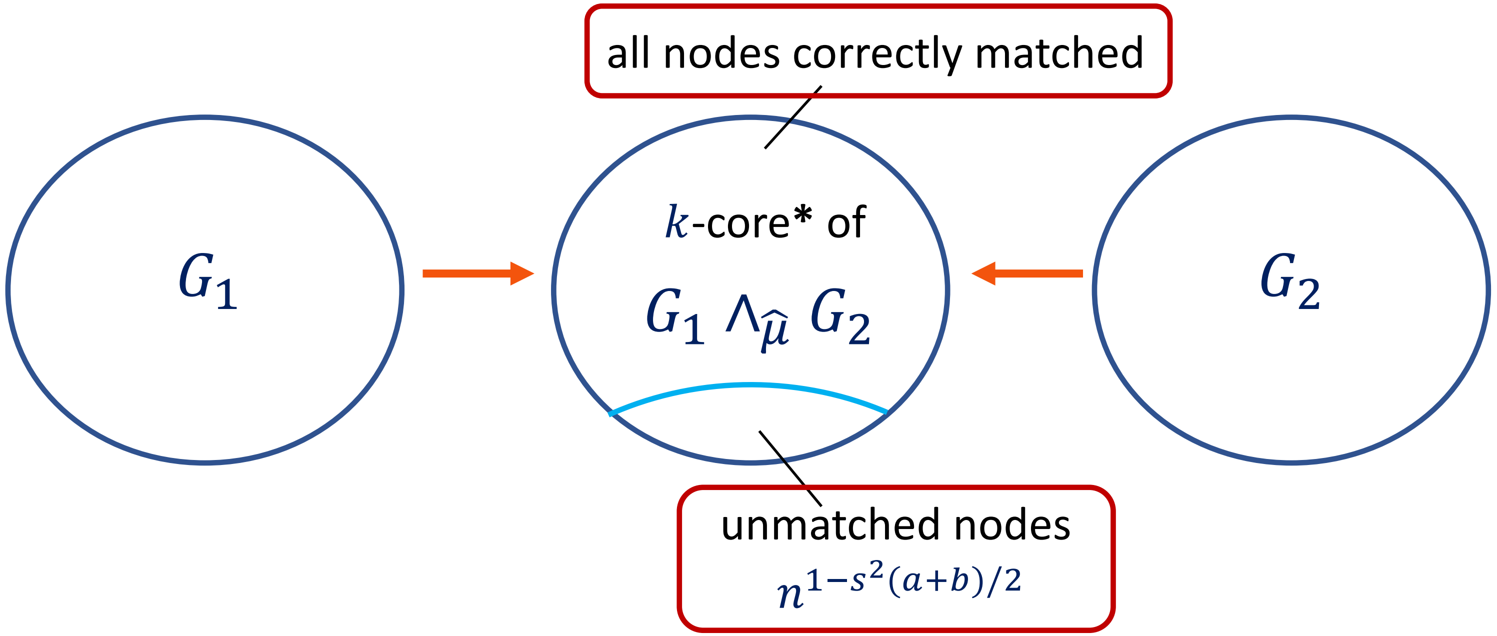
Algorithm

1. Almost exact labeling of G_1
   [Mossel, Neeman, Sly, 2014]
2. Partial almost exact graph matching $\hat{\mu}$
   [Cullina, Kiyavash, Mittal, Poor, 2020]



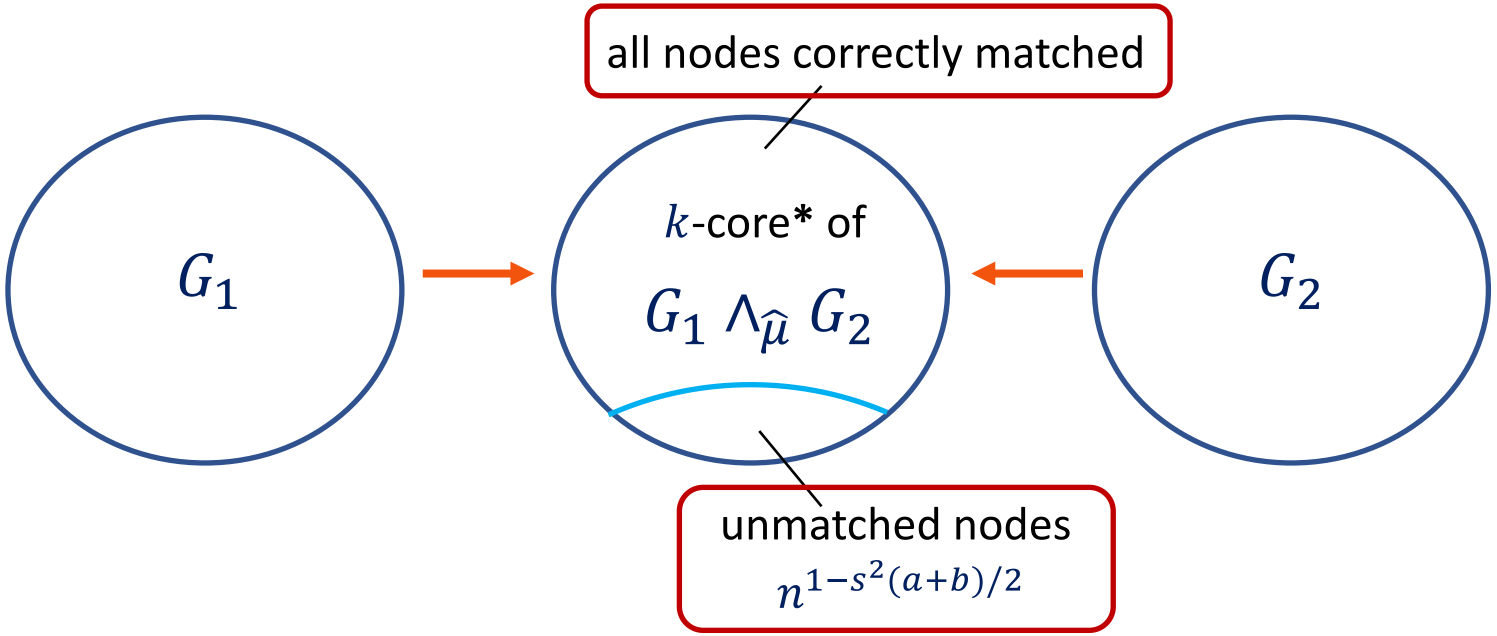
Algorithm

1. Almost exact labeling of G_1
   [Mossel, Neeman, Sly, 2014]
2. Partial almost exact graph matching $\hat{\mu}$
   [Cullina, Kiyavash, Mittal, Poor, 2020]



Algorithm

1. Almost exact labeling of G_1
   [Mossel, Neeman, Sly, 2014]
2. Partial almost exact graph matching $\hat{\mu}$
   [Cullina, Kiyavash, Mittal, Poor, 2020]

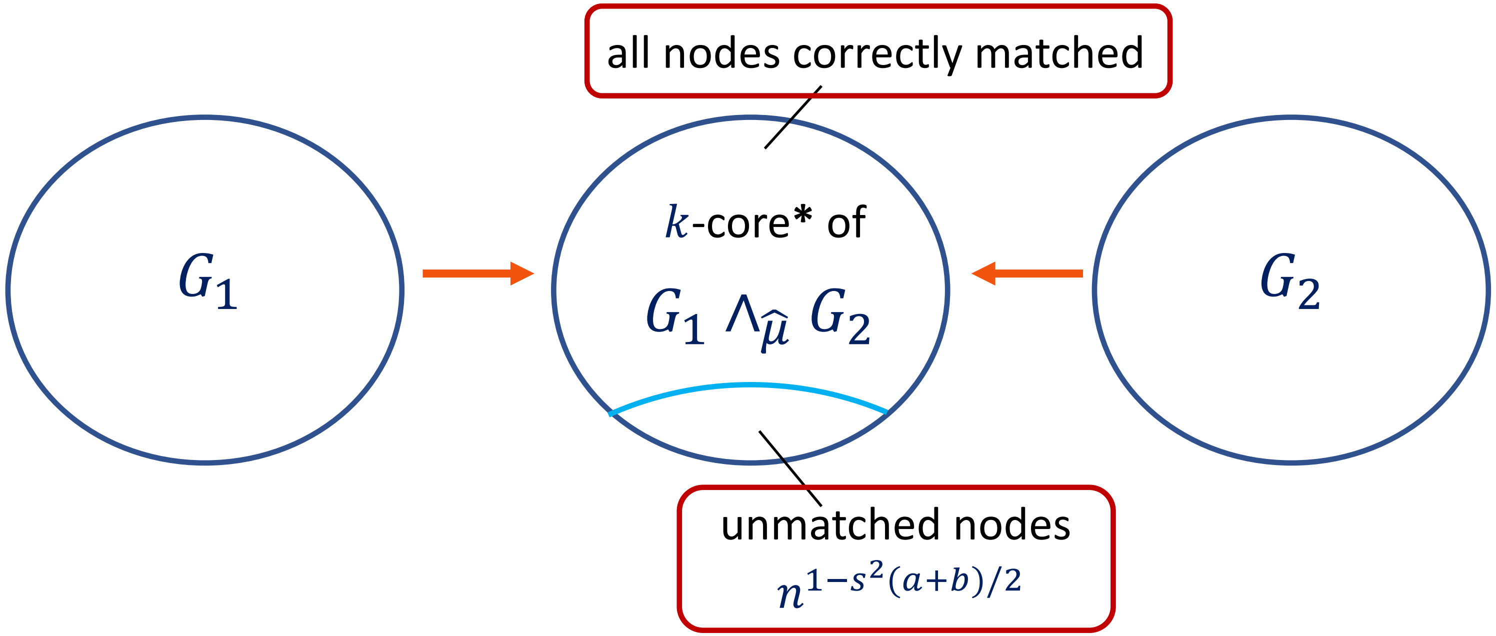


3. For matched nodes in G_1:

- Consider $G_1 \vee_{\hat{\mu}} G_2$
- Use majority vote among neighbors in $G_1 \vee_{\hat{\mu}} G_2$ to refine labels

Algorithm

1. Almost exact labeling of G_1
   [Mossel, Neeman, Sly, 2014]
2. Partial almost exact graph matching $\hat{\mu}$
   [Cullina, Kiyavash, Mittal, Poor, 2020]



3. For matched nodes in G_1:

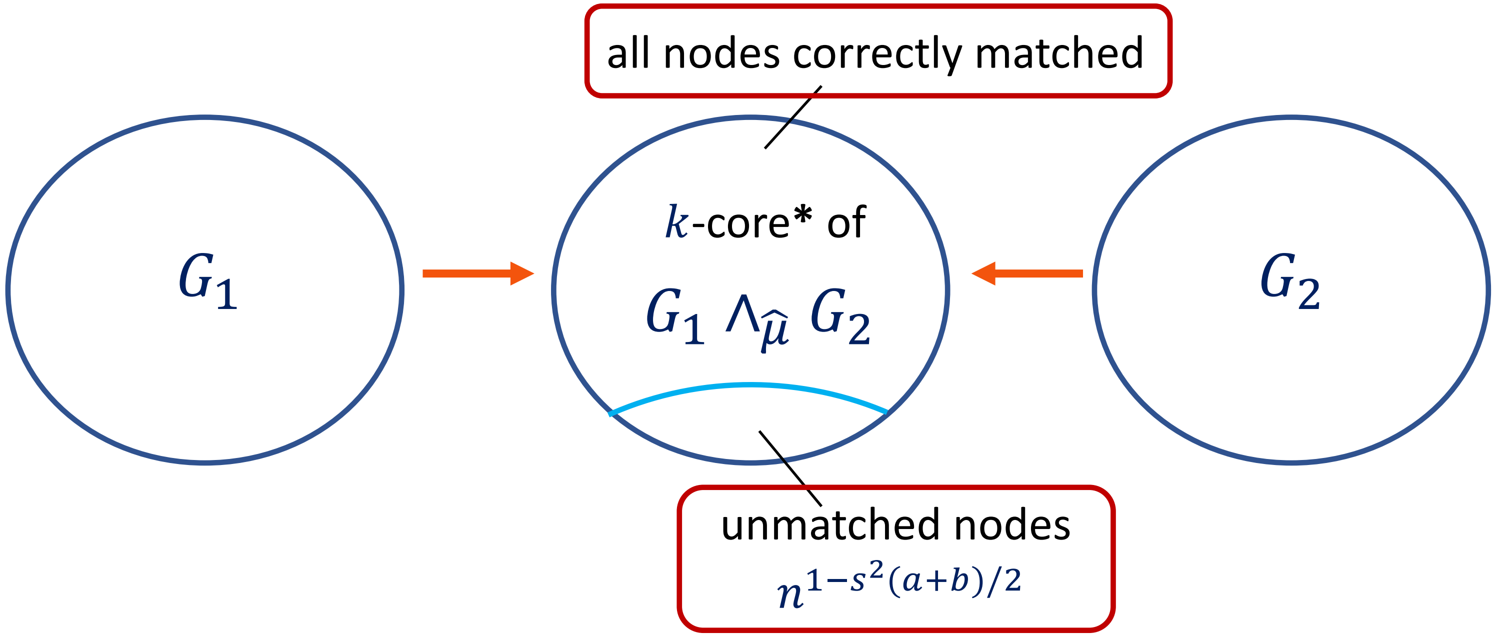
- Consider $G_1 \vee_{\hat{\mu}} G_2$
- Use majority vote among neighbors in $G_1 \vee_{\hat{\mu}} G_2$ to refine labels

4. For unmatched nodes in G_1:

- Use majority vote among neighbors in G_1

Algorithm

1. Almost exact labeling of G_1
   [Mossel, Neeman, Sly, 2014]
2. Partial almost exact graph matching $\hat{\mu}$
   [Cullina, Kiyavash, Mittal, Poor, 2020]



3. For matched nodes in G_1:
- Consider $G_1 \vee_{\hat{\mu}} G_2$
- Use majority vote among neighbors in $G_1 \vee_{\hat{\mu}} G_2$ to refine labels

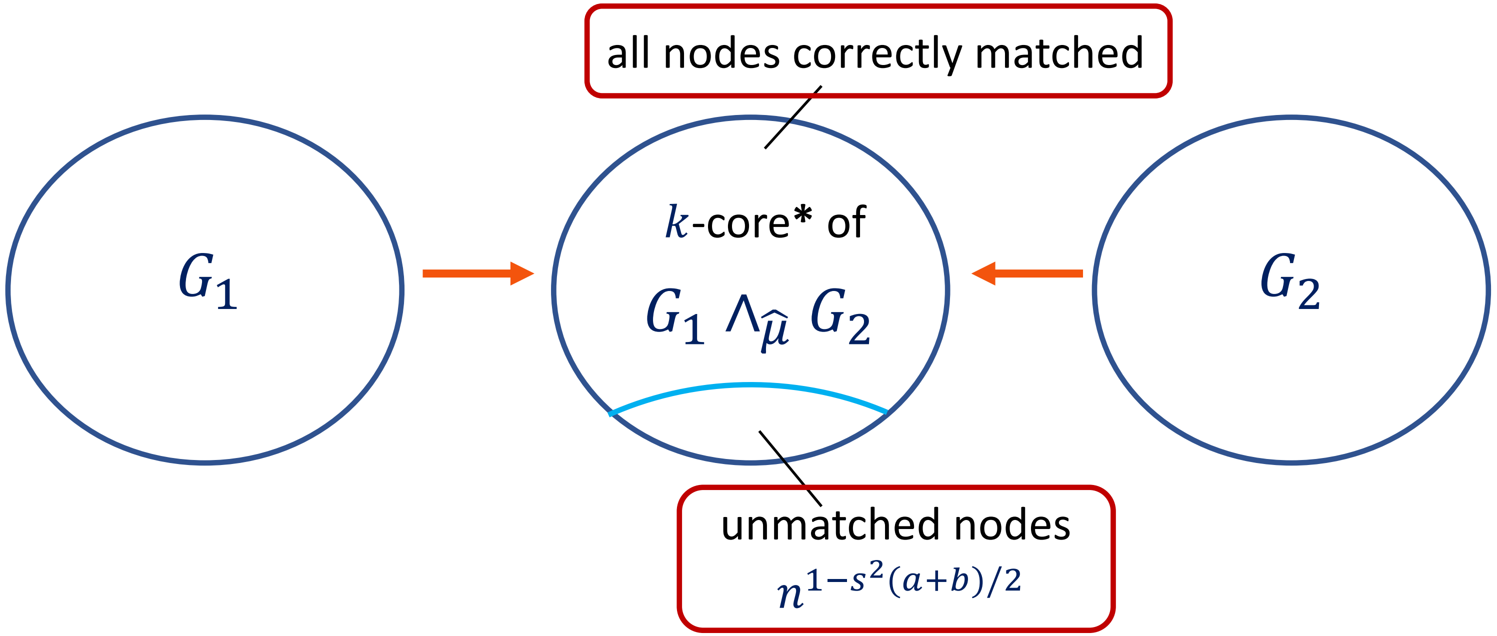
4. For unmatched nodes in G_1:
- Use majority vote among neighbors in G_1

relevant quantity is

$$s(1-s)\left(\frac{\sqrt{a}-\sqrt{b}}{\sqrt{2}}\right)^2$$

Algorithm

1. Almost exact labeling of G_1
   [Mossel, Neeman, Sly, 2014]
2. Partial almost exact graph matching $\hat{\mu}$
   [Cullina, Kiyavash, Mittal, Poor, 2020]



3. For matched nodes in G_1:
- Consider $G_1 \vee_{\hat{\mu}} G_2$
- Use majority vote among neighbors in $G_1 \vee_{\hat{\mu}} G_2$ to refine labels

$$s^2\left(\frac{a+b}{2}\right) + s(1-s)\left(\frac{\sqrt{a}-\sqrt{b}}{\sqrt{2}}\right)^2 = 1$$

4. For unmatched nodes in G_1:
- Use majority vote among neighbors in G_1

relevant quantity is

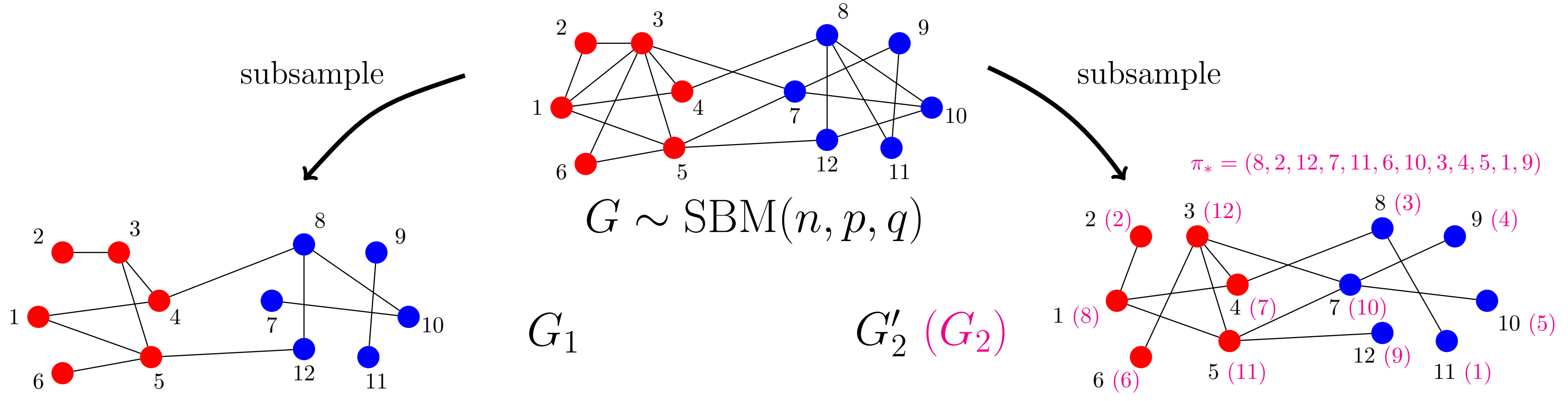
$$s(1-s)\left(\frac{\sqrt{a}-\sqrt{b}}{\sqrt{2}}\right)^2$$

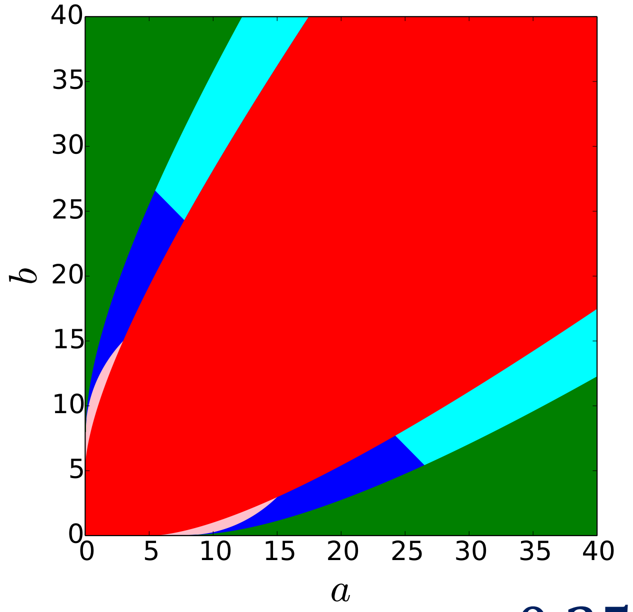
Impossibility argument sketch

- S_*: singletons in the intersection graph $G_1 \wedge_{\pi_*} G_2$
- Key: $|S_*| \asymp n^{1-s^2(a+b)/2}$
- MAP estimator fails even if given:
  - All community labels in G_2
  - S_*
  - π_* on $[n] \setminus S_*$
- Proof uses careful second moment analysis

Open problems / future directions

Efficient algorithms





$s = 0.25$

- Current algorithms for (exact) graph matching are not efficient
- Do there exist efficient algorithms for graph matching?

Exciting and promising recent developments for efficient graph matching for correlated Erdős—Rényi random graphs:

- Mao, Rudelson, Tikhomirov (2021)
- Mao, Wu, Xu, Yu (2022)

Beyond exact community recovery

- Almost exact recovery? → Improved error rate?
- Partial recovery? → Improved fraction recovered?
- Community detection? → Lower threshold?

Beyond exact community recovery

- Almost exact recovery? → Improved error rate?
- Partial recovery? → Improved fraction recovered?
- Community detection? → Lower threshold?

(Gaudio, R., Sridhar; in progress)

- Optimal error rate for almost exact recovery
- Beating KS w/ two correlated SBMs

Beyond exact community recovery

- Almost exact recovery?
- Partial recovery?
- Community detection?

→ Improved error rate?
→ Improved fraction recovered?
→ Lower threshold?

(Gaudio, R., Sridhar; in progress)

- Optimal error rate for almost exact recovery
- Beating KS w/ two correlated SBMs

Open problem

Predict the threshold for community detection from two correlated SBMs

Beyond exact community recovery

- Almost exact recovery? → Improved error rate?
- Partial recovery? → Improved fraction recovered?
- Community detection? → Lower threshold?

(Gaudio, R., Sridhar; in progress)

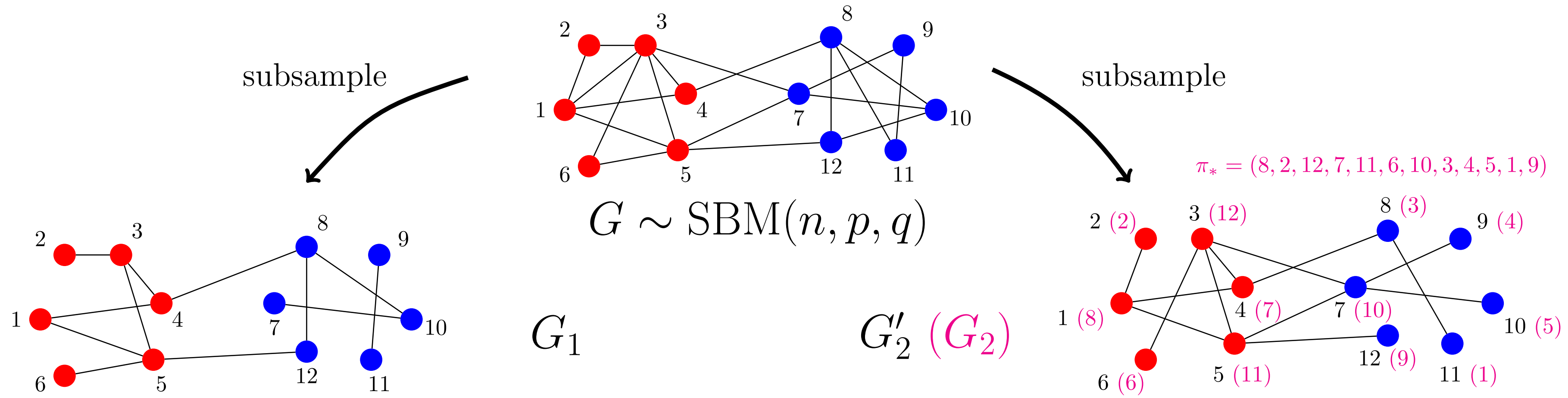
- Optimal error rate for almost exact recovery
- Beating KS w/ two correlated SBMs

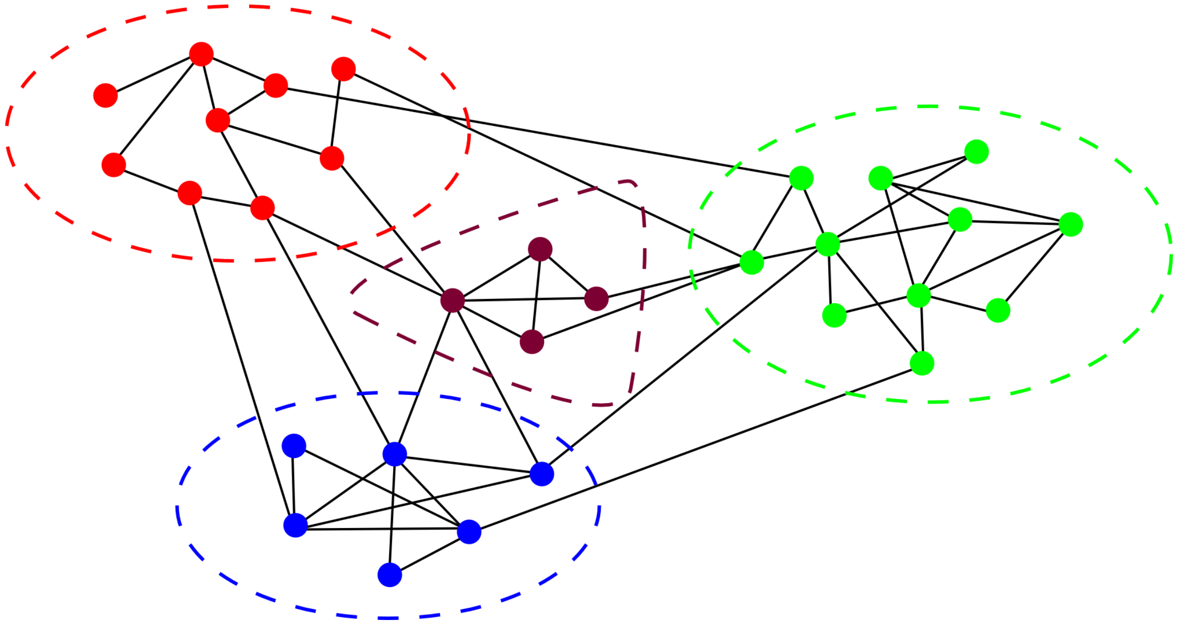
Open problem

Predict the threshold for community detection from two correlated SBMs

Challenge:
interplay between community recovery and graph matching

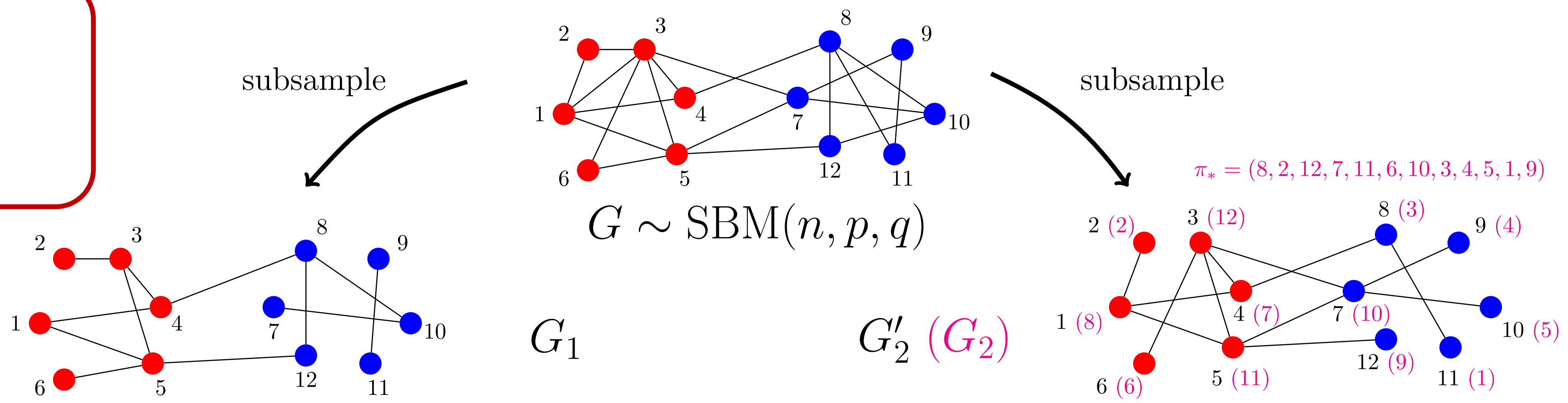
General correlated SBMs





k communities, general parameters:
- graph matching?
- (exact) community recovery?

Summary



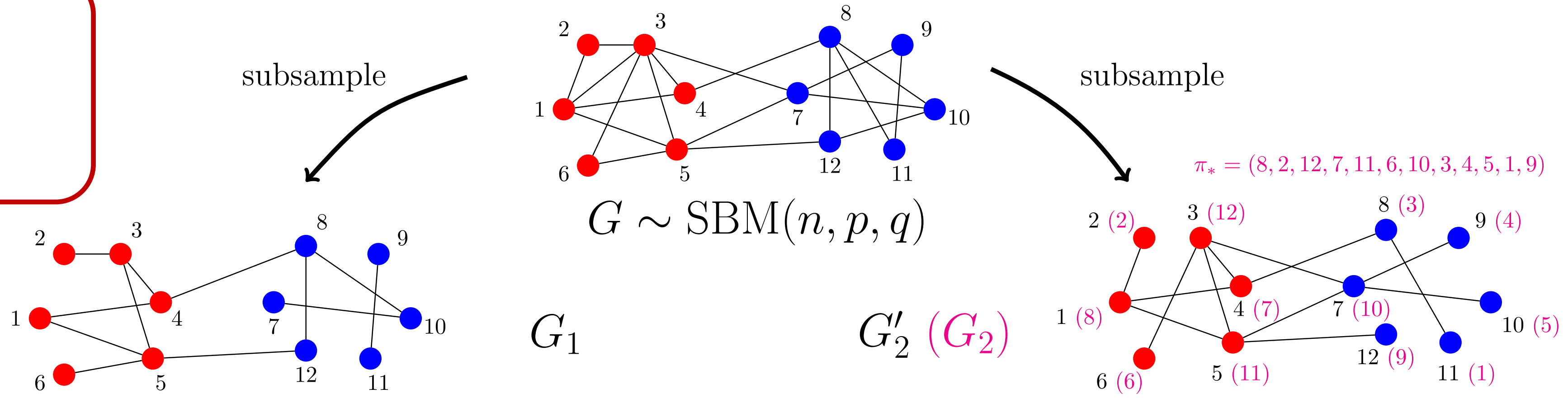
- **Correrelated SBMs:** determined the fundamental limits of

  exact graph matching and **exact community recovery**

- **Exact community recovery** possible in regimes where it is not possible from G_1 alone

- **Correlated random graphs:** many challenges and applications

Summary



- **Correlated SBMs:** determined the fundamental limits of

  exact graph matching and **exact community recovery**

- **Exact community recovery** possible in regimes where it is not possible from G_1 alone

- **Correlated random graphs:** many challenges and applications

Thank you!