

Correlated stochastic block models: graph matching and community recovery

Based on joint work with Anirudh Sridhar

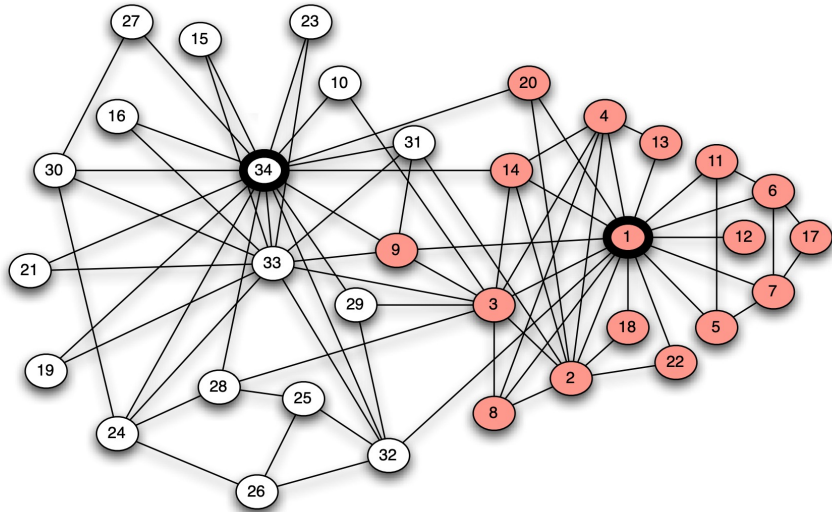
Miklós Z. Rácz



Workshop on Algorithmic Advances for Statistical Inference
with Combinatorial Structure, Simons Institute, Berkeley

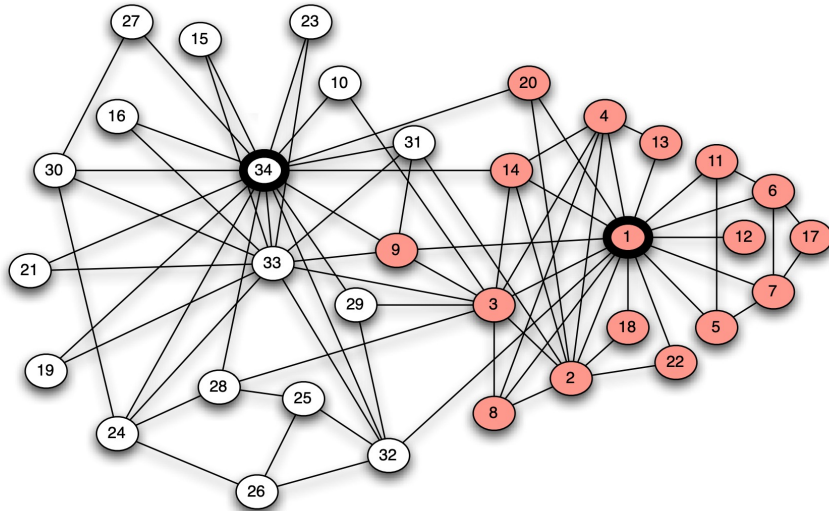
October 11, 2021

Recovering communities in networks

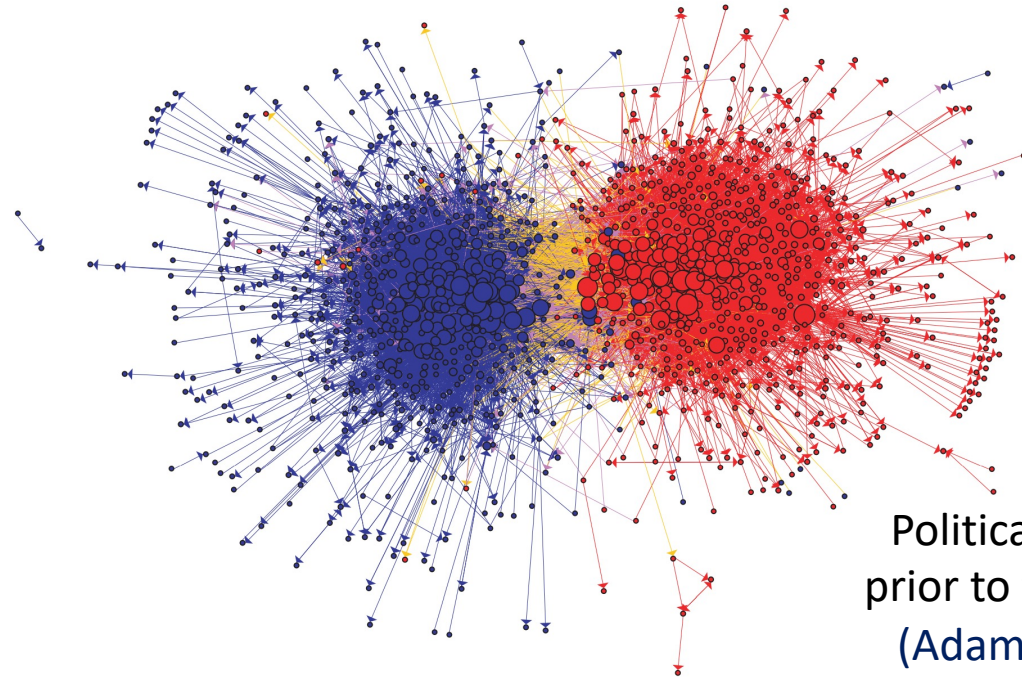


Zachary's karate club (1970-72; 1977)

Recovering communities in networks



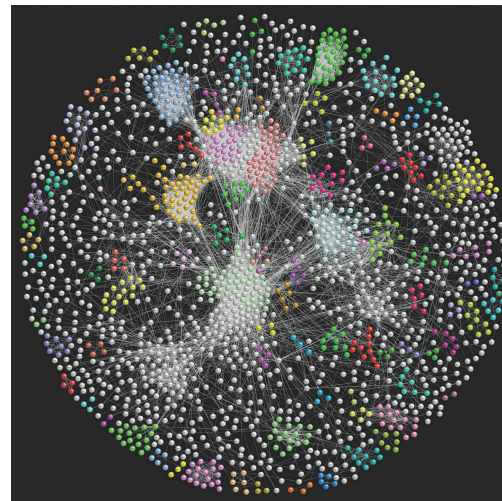
Zachary's karate club (1970-72; 1977)



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

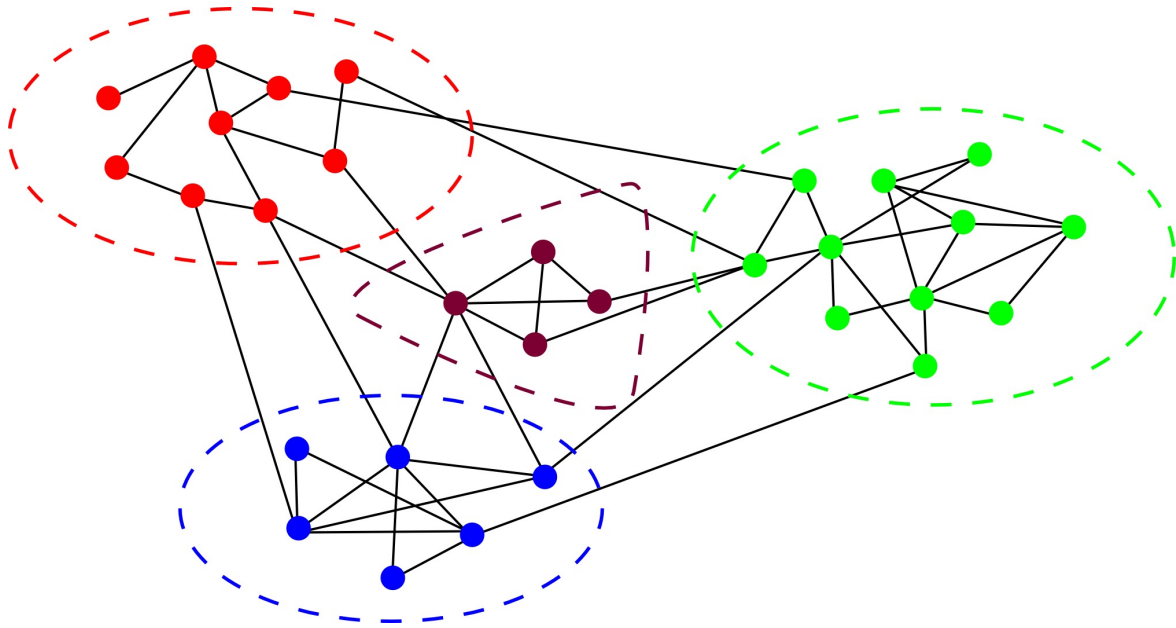


Facebook: 2.7+ billion monthly active users
1.8+ billion daily active users



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

Stochastic block model (SBM)

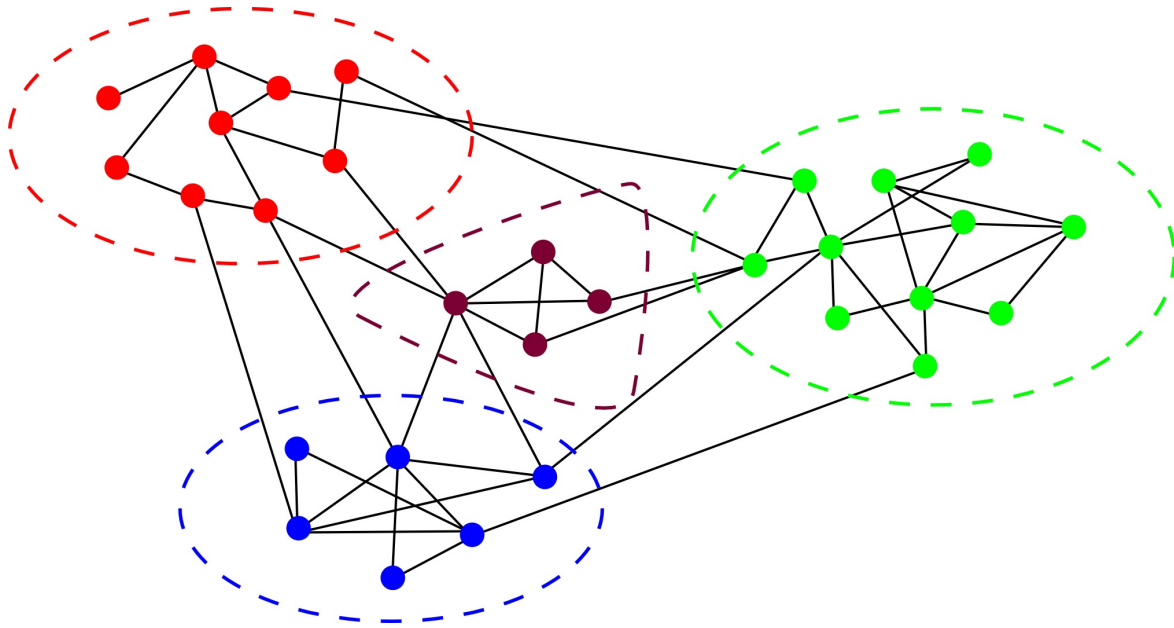


Holland, Laskey, Leinhardt (1983)

Many works in physics, probability, statistics, 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)



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

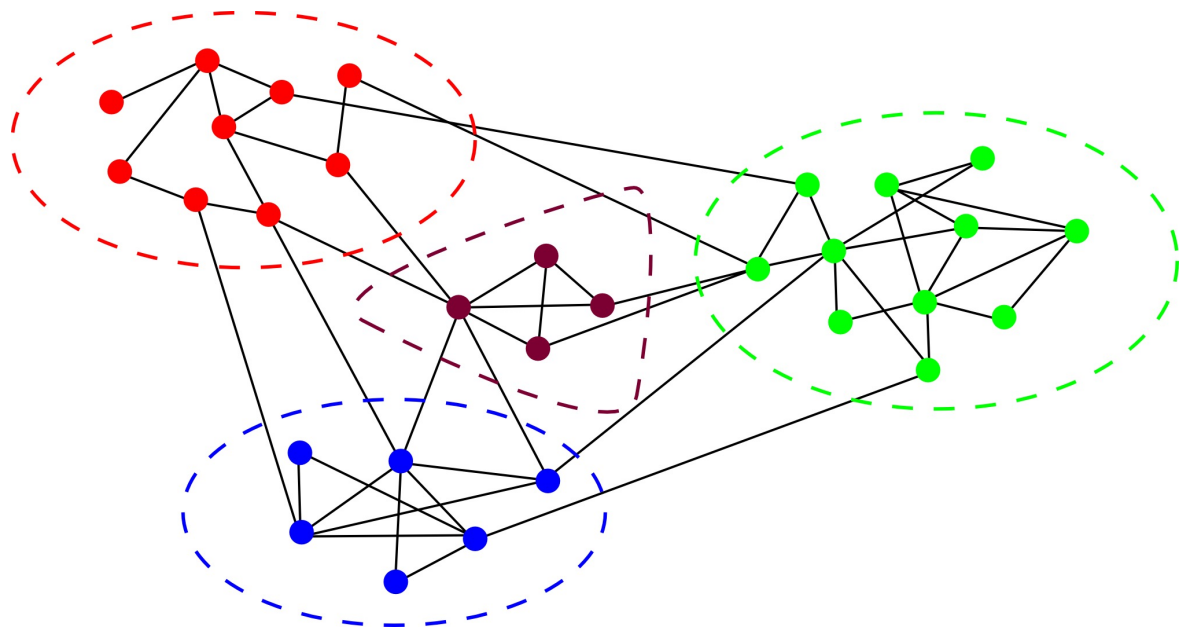
- Partial recovery?
- Almost exact recovery?
- Exact recovery?

Holland, Laskey, Leinhardt (1983)

Many works in physics, probability, statistics, 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)



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

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

Holland, Laskey, Leinhardt (1983)

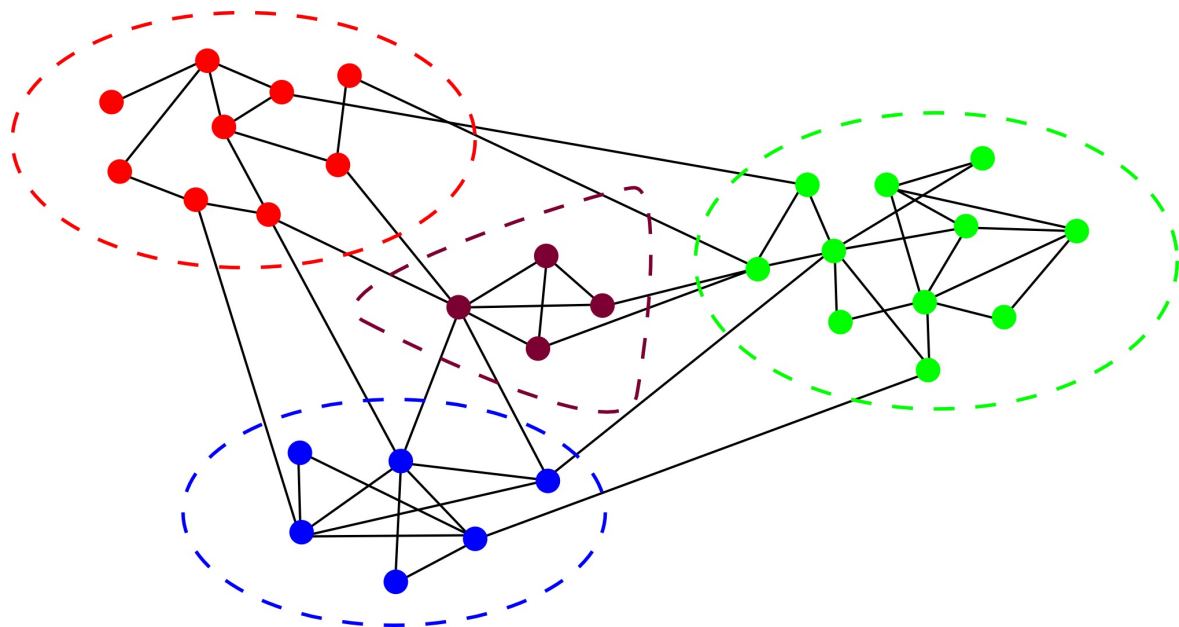
Many works in physics, probability, statistics, CS, info theory... including:

- 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)
- ...

This talk: two balanced communities

- n nodes
- $\sigma_i \in \{+1, -1\}$ i.i.d. uniform community labels
- Given $\sigma = \{\sigma_i\}$, 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)



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

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

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

Holland, Laskey, Leinhardt (1983)

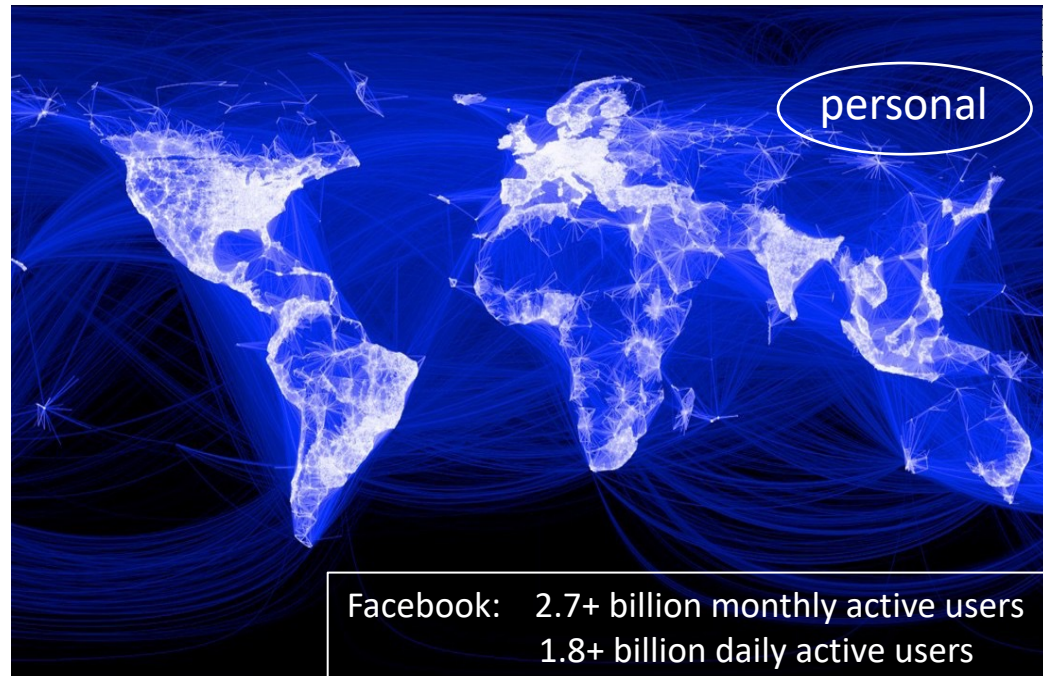
Many works in physics, probability, statistics, CS, info theory... including:

- 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)
- ...

This talk: two balanced communities

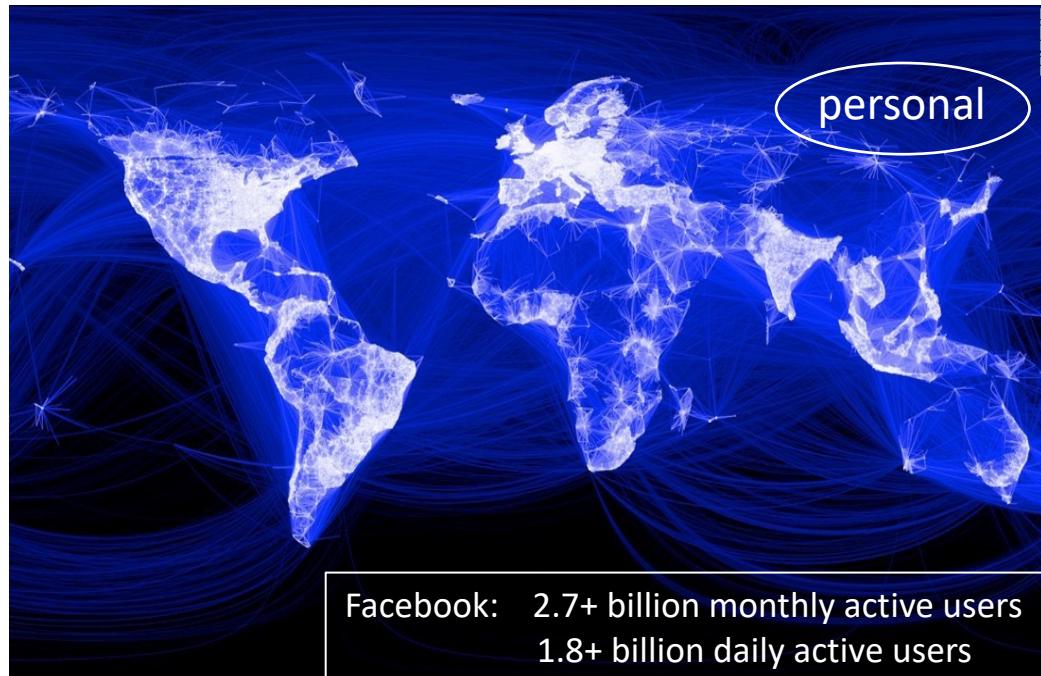
- n nodes
- $\sigma_i \in \{+1, -1\}$ i.i.d. uniform community labels
- Given $\sigma = \{\sigma_i\}$, edges drawn independently:
 - If $\sigma_i = \sigma_j$, then $i \sim j$ with prob. p
 - If $\sigma_i \neq \sigma_j$, then $i \sim j$ with prob. q

Multiple correlated networks



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

Multiple correlated networks



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

STOCHASTIC BLOCKMODELS: FIRST STEPS *

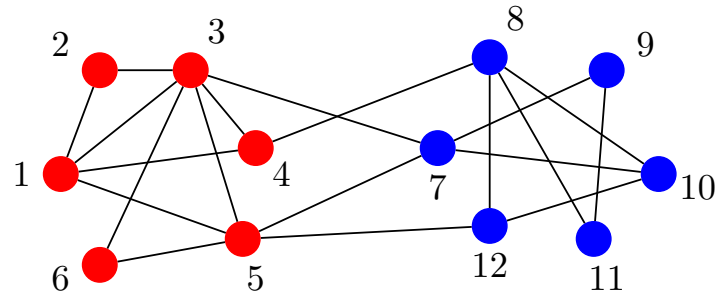
Paul W. HOLLAND
*Educational Testing Service***

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.

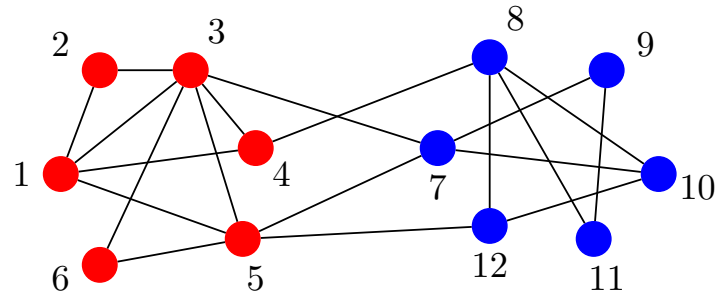
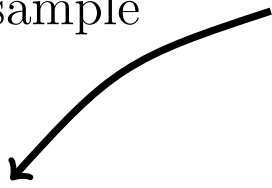
Correlated stochastic block model



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

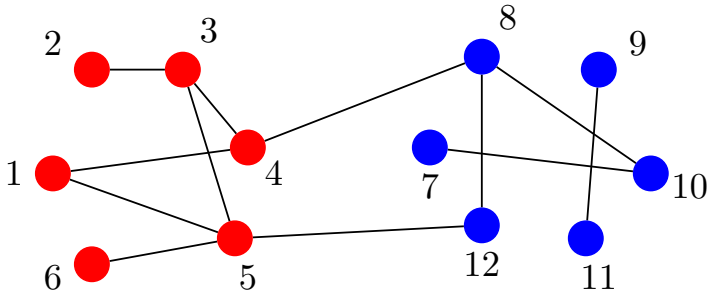
Correlated stochastic block model

subsample



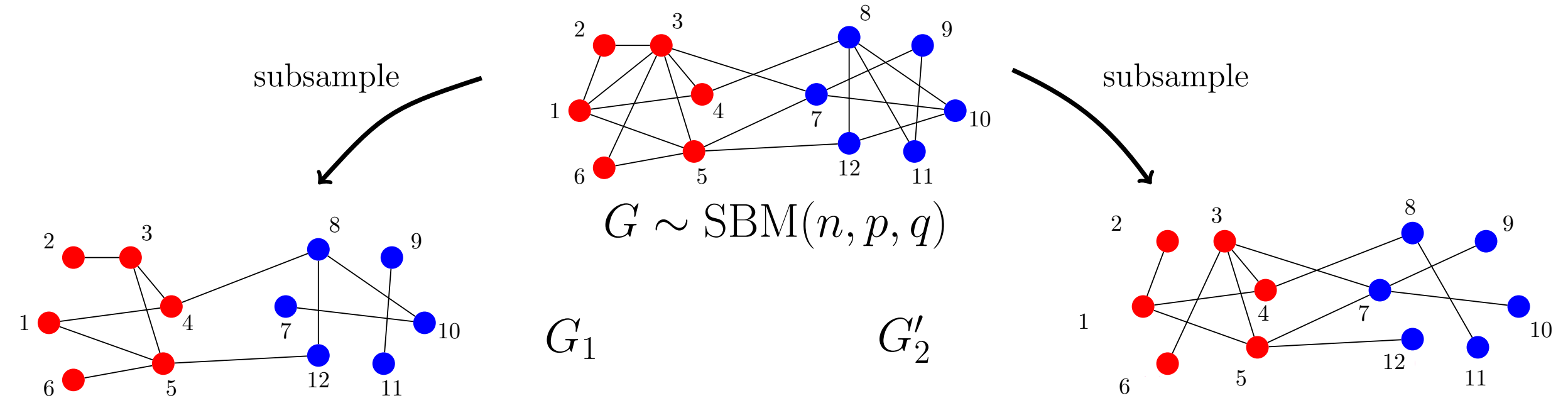
$$G \sim \text{SBM}(n, p, q)$$

G_1



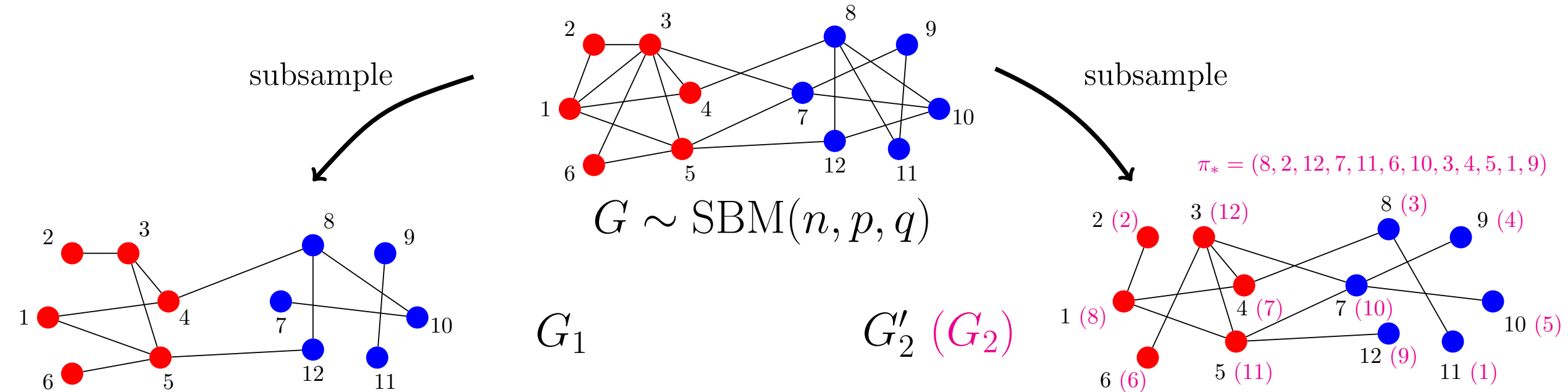
- Subsampling probability $s \in [0,1]$

Correlated stochastic block model



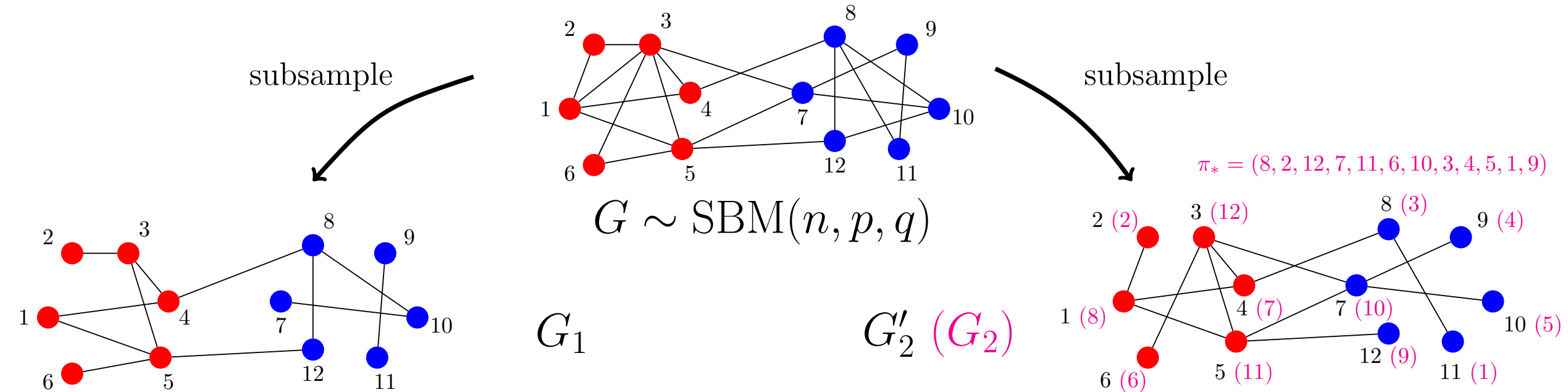
- Subsampling probability $s \in [0,1]$

Correlated stochastic block model



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

Correlated stochastic block model



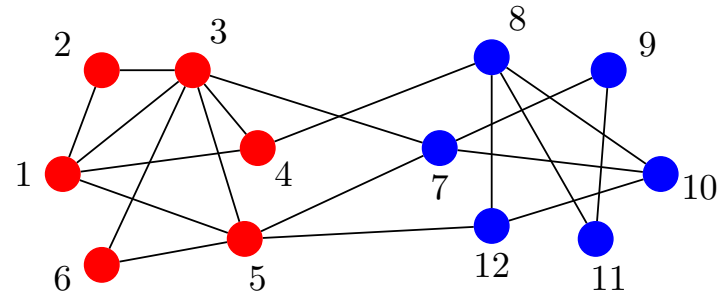
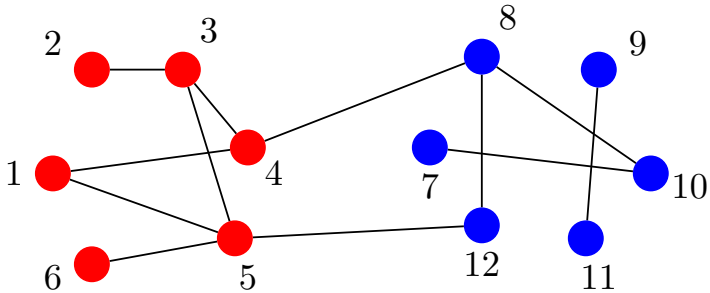
- Subsampling probability $s \in [0, 1]$
- π_* uniformly random permutation of $[n]$

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

(Onaran, Garg, Erkip, 2016)

Correlated stochastic block model

subsample

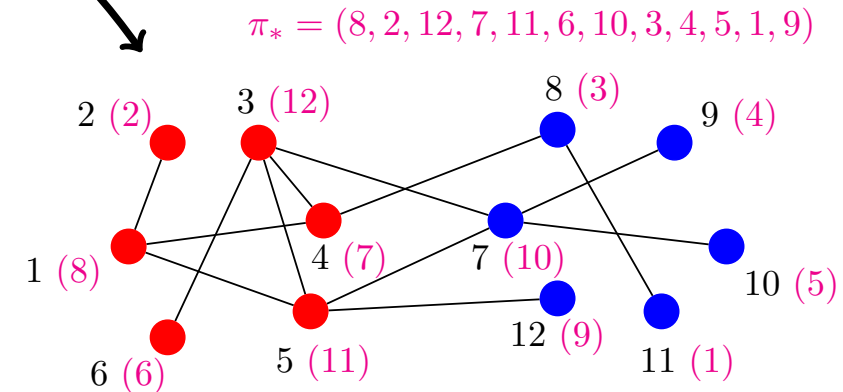


$$G \sim \text{SBM}(n, p, q)$$

G_1

G'_2 (G_2)

subsample



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

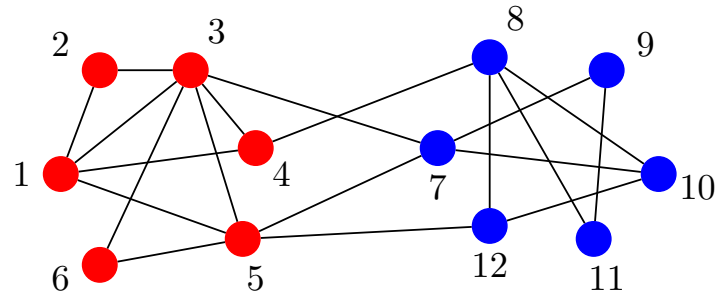
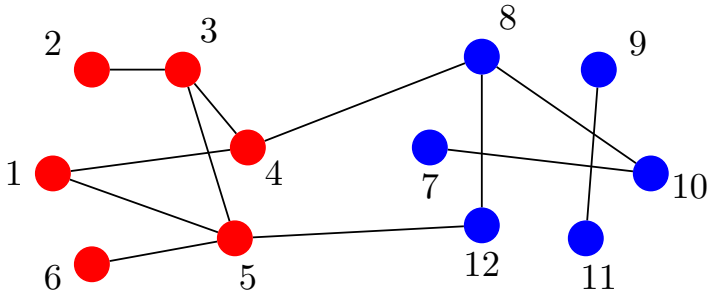
- Marginally $G_1, G_2 \sim \text{SBM}(n, ps, qs)$
- Corresponding edges are *correlated*

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

(Onaran, Garg, Erkip, 2016)

Correlated stochastic block model

subsample

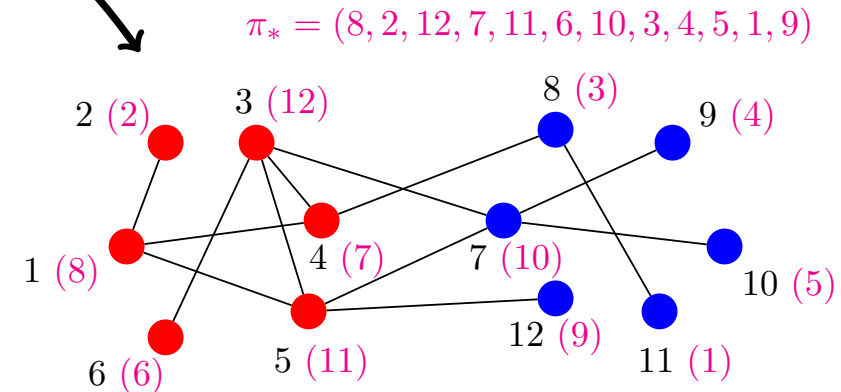


$$G \sim \text{SBM}(n, p, q)$$

G_1

G'_2 (G_2)

subsample



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

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

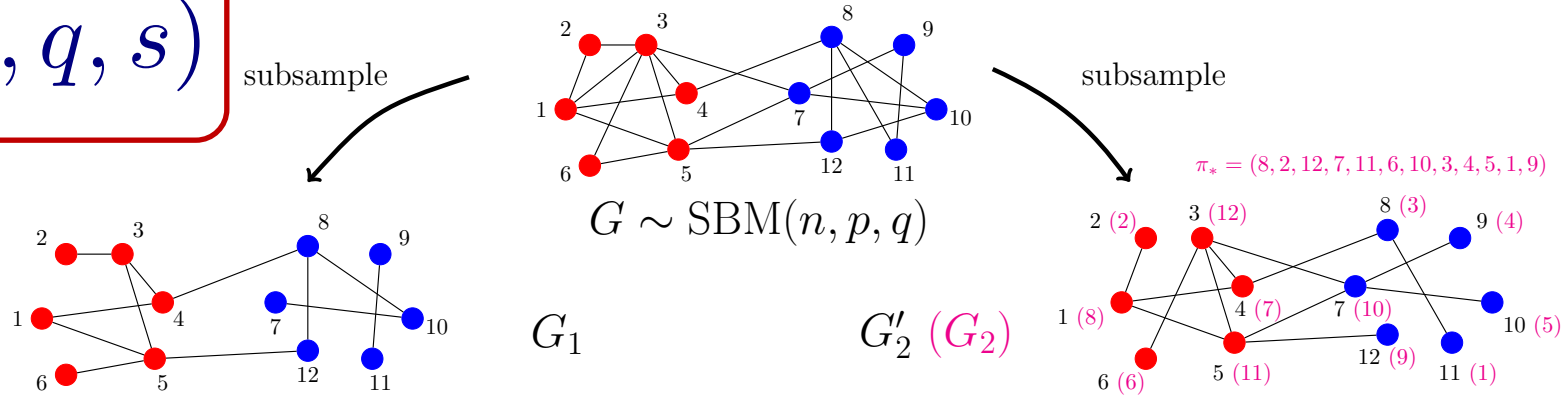
$$(G_1, G_2) \sim \text{CSBM}(n, p, q, s)$$

(Onaran, Garg, Erkip, 2016)

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

Correlated stochastic block model

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



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 \Rightarrow logarithmic degree regime: $p = a \log(n) / n$ and $q = b \log(n) / n$

Exact community recovery in the SBM

Need no isolated vertices \Rightarrow 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: $G \sim \text{SBM} \left(n, \frac{a \log n}{n}, \frac{b \log n}{n} \right)$

Exact recovery is **possible** (in polynomial time) if

$$|\sqrt{a} - \sqrt{b}| > \sqrt{2}$$

Exact recovery is **impossible** if

$$|\sqrt{a} - \sqrt{b}| < \sqrt{2}$$

Exact community recovery in the SBM

Need no isolated vertices \Rightarrow 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: $G \sim \text{SBM} \left(n, \frac{a \log n}{n}, \frac{b \log n}{n} \right)$

Exact recovery is **possible** (in polynomial time) if

$$|\sqrt{a} - \sqrt{b}| > \sqrt{2}$$

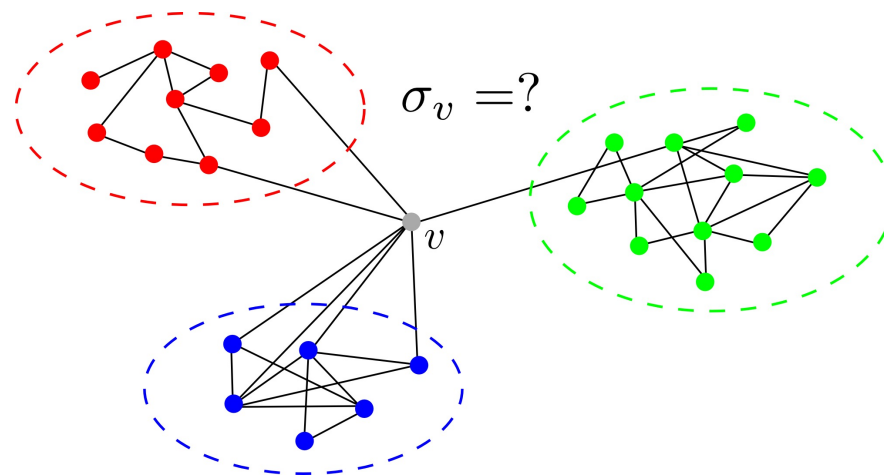
Exact recovery is **impossible** if

$$|\sqrt{a} - \sqrt{b}| < \sqrt{2}$$

Abbé, Sandon (2015): threshold for general SBMs

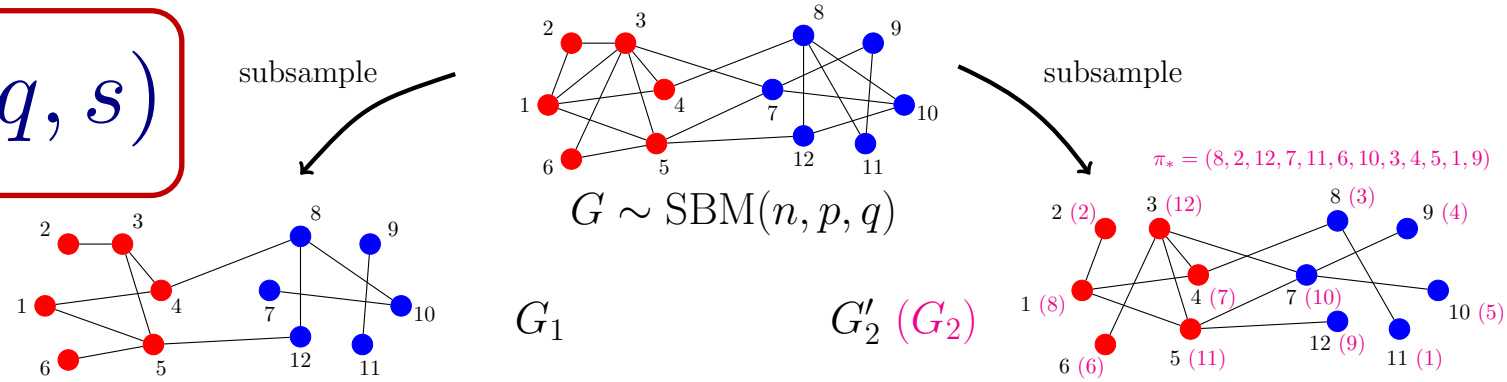
Intuition:

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



Exact community recovery in correlated SBMs

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

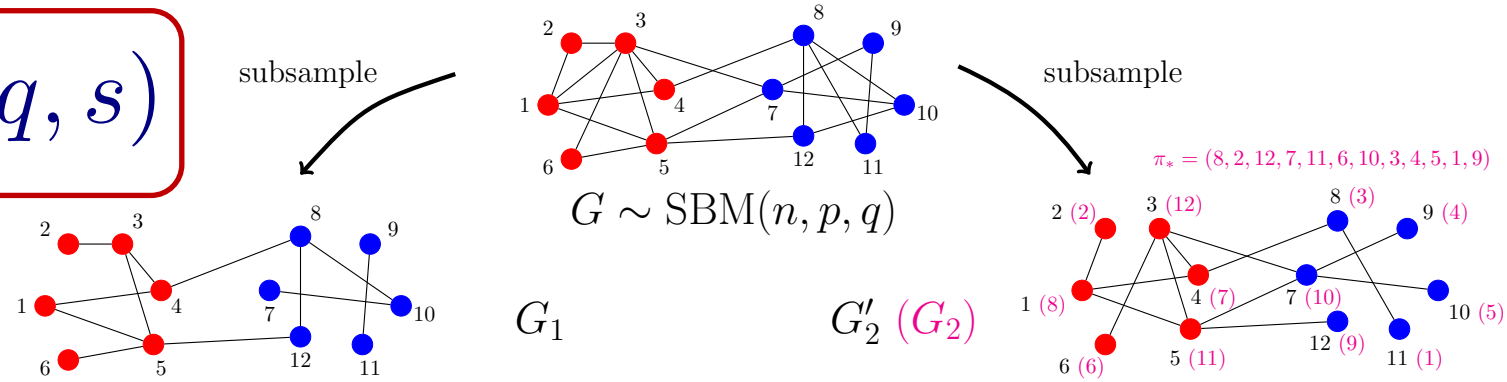


Since $G_1 \sim \text{SBM}(n, ps, qs)$, exact community recovery is possible from G_1 iff

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

Exact community recovery in correlated SBMs

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



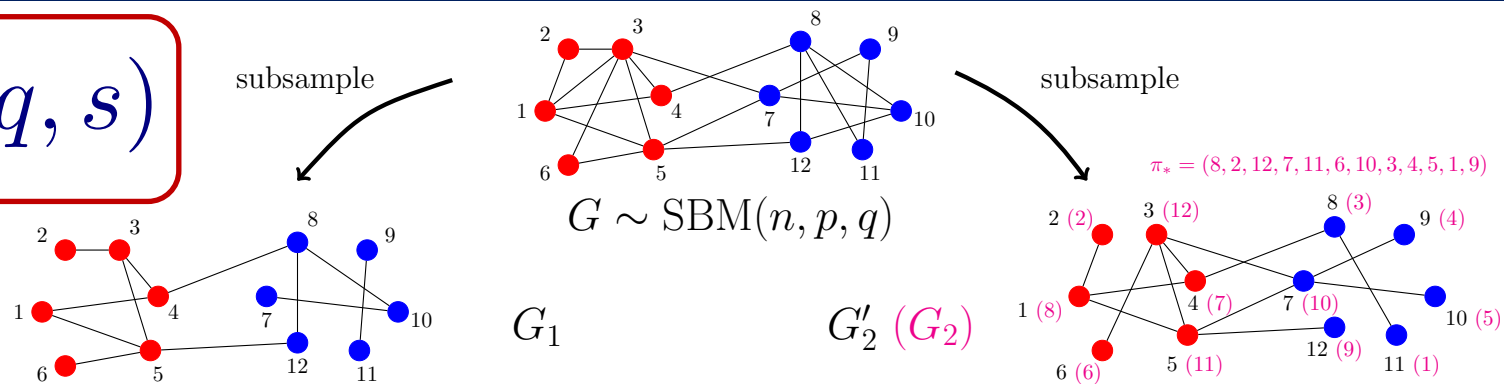
Since $G_1 \sim \text{SBM}(n, ps, qs)$, exact community recovery is possible from G_1 iff

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

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

Exact community recovery in correlated SBMs

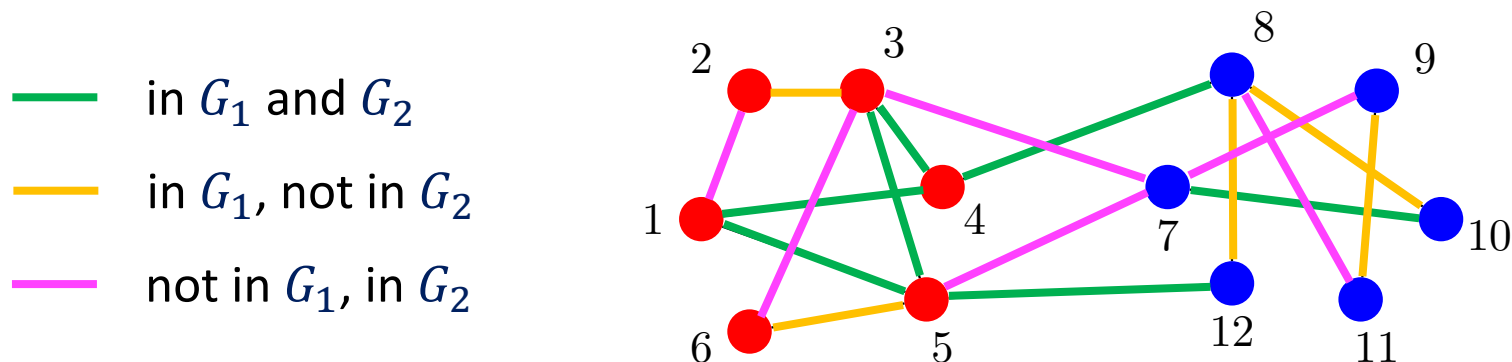
$$(G_1, G_2) \sim \text{CSBM}(n, p, q, s)$$



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

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

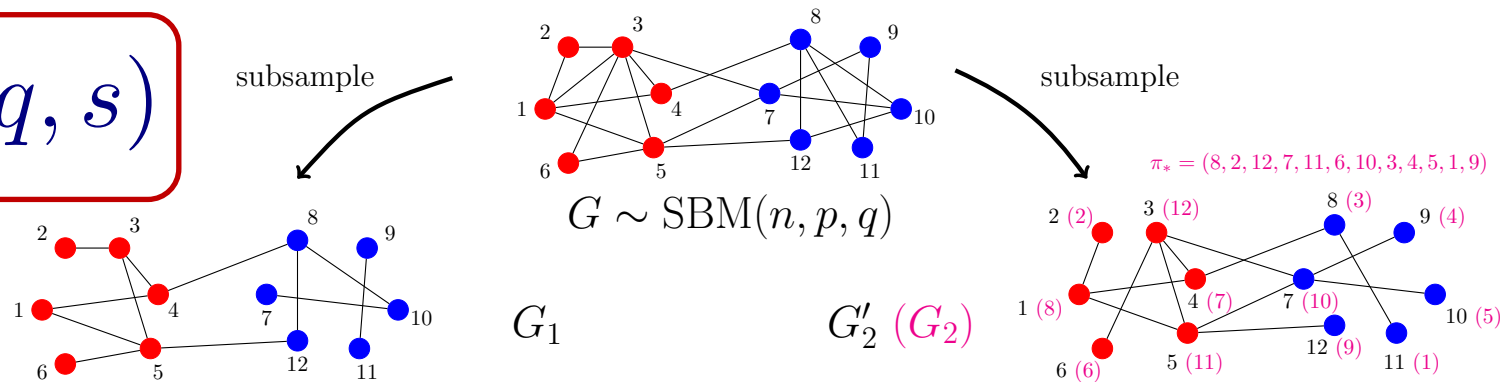
How can we use both G_1 and G_2 ? **Suppose that π_* is known.** Then:



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

Exact community recovery in correlated SBMs

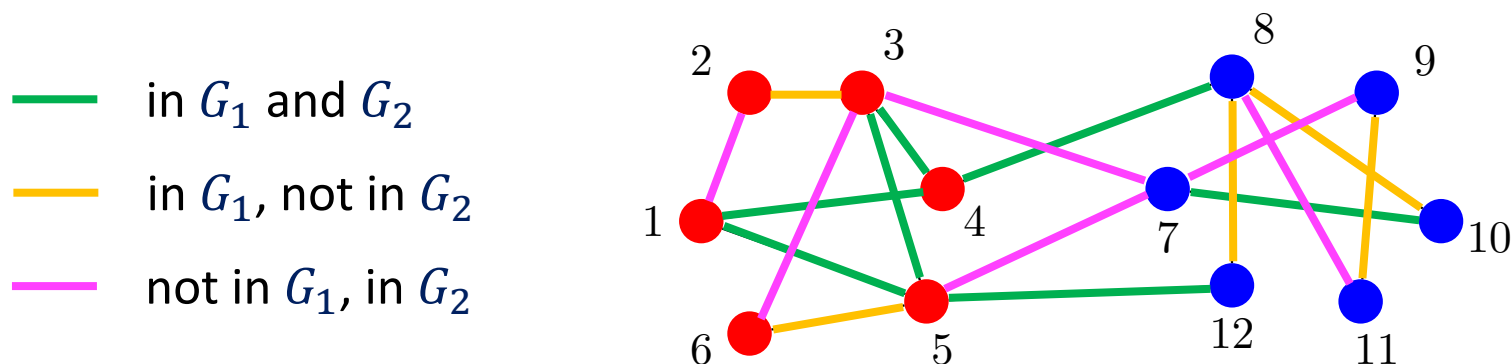
$$(G_1, G_2) \sim \text{CSBM}(n, p, q, s)$$



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

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

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



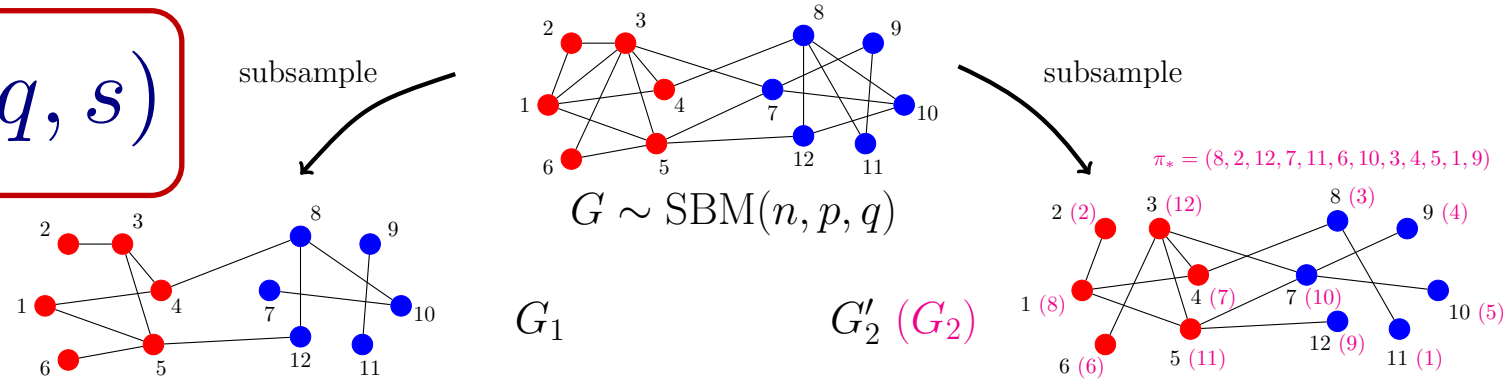
Thus exact community recovery is possible iff

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

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

Exact community recovery in correlated SBMs

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



Sin

H

In particular, if π_* is known and

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

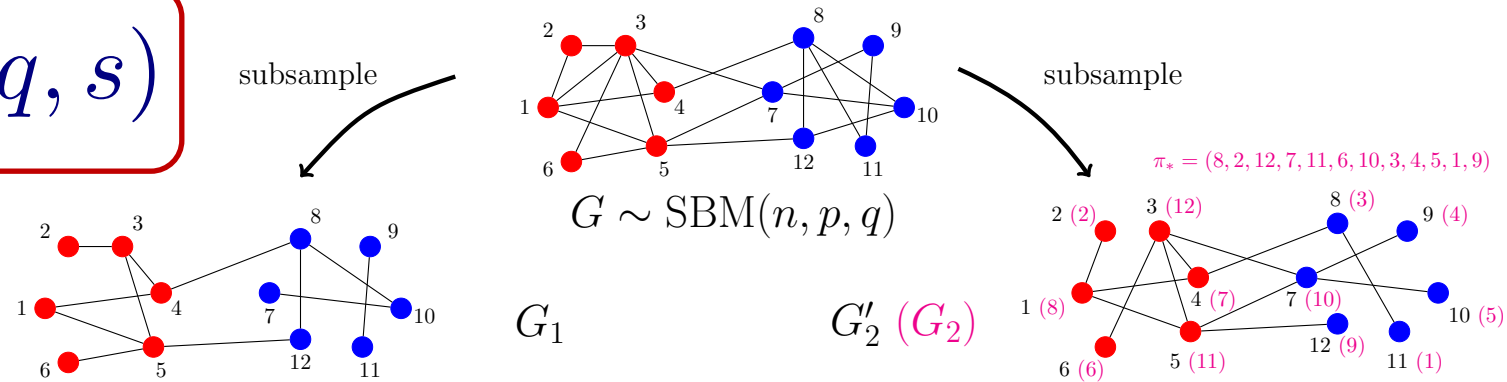
then exact community recovery is possible from G_1 and G_2 ,
even though it is impossible from G_1 alone

G_1

(2)

Graph matching

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

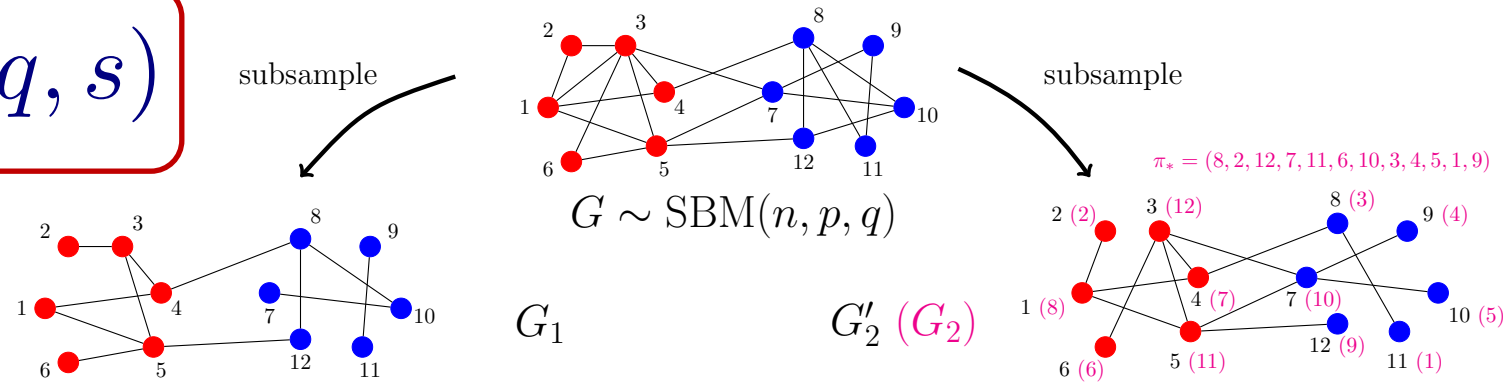


Main Q:

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

Graph matching

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

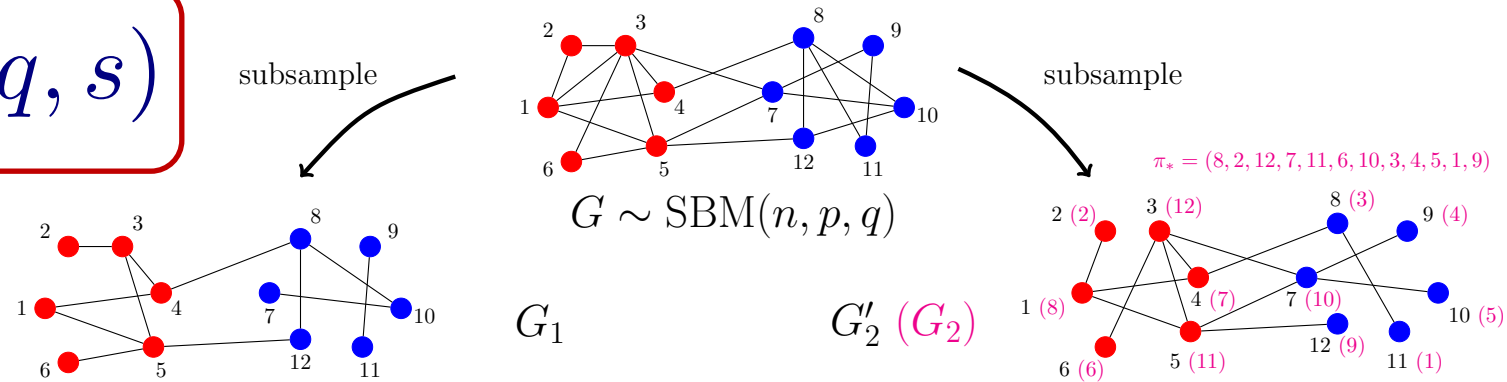


Main Q:

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

Graph matching

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

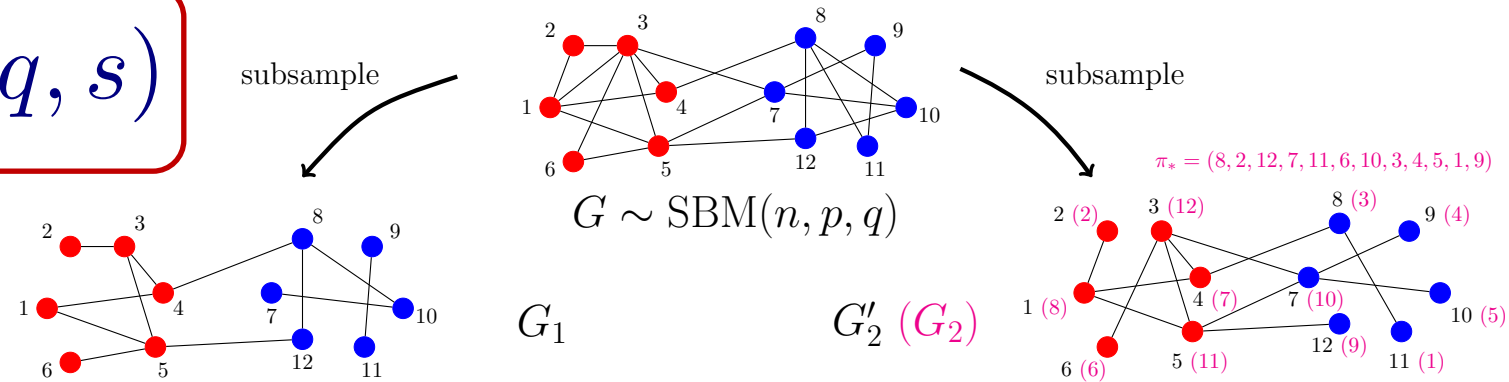


Main Q:

- given (G_1, G_2) , when can we (exactly) recover the latent permutation π_* ?
- Of significant independent interest
- Correlated Erdős-Rényi random graphs:
Pedersani, Grossglauser (2011)

Graph matching

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



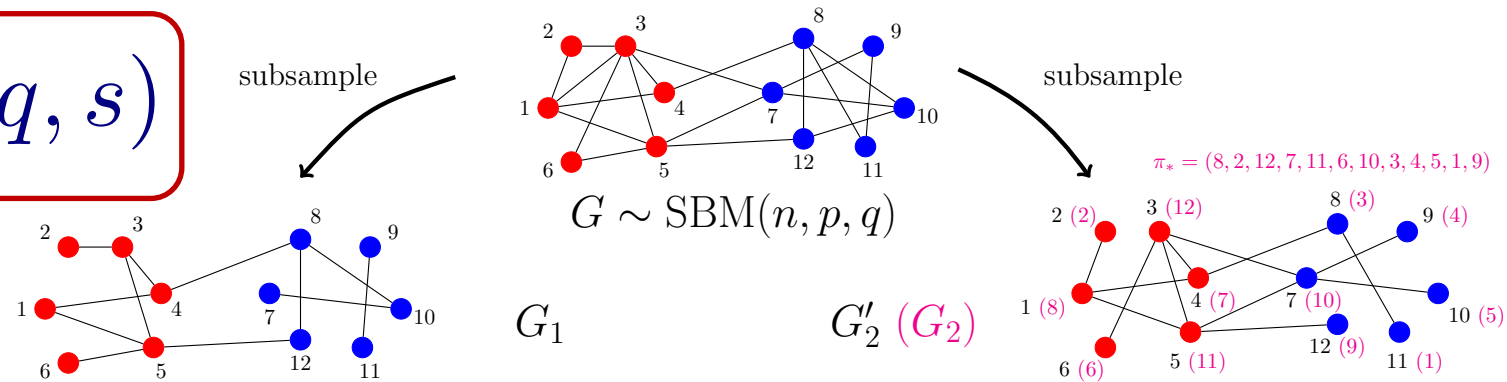
Main Q:

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

- Of significant independent interest
- Correlated Erdős-Rényi random graphs: Pedersani, Grossglauser (2011)
- Many works in probability/statistics/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)
 - Cullina, Kiyavash, Mittal, Poor (2020)
 - Wu, Xu, Yu (2020, 2021)
 - Mao, Rudelson, Tikhomirov (2021)
 - Ganassali, Lelarge, Massoulié (2021)

Correlated SBMs: graph matching and community recovery

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



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 .

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

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



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 .

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

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



- $\hat{\pi}$ 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 .

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

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



- $\hat{\pi}$ 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)

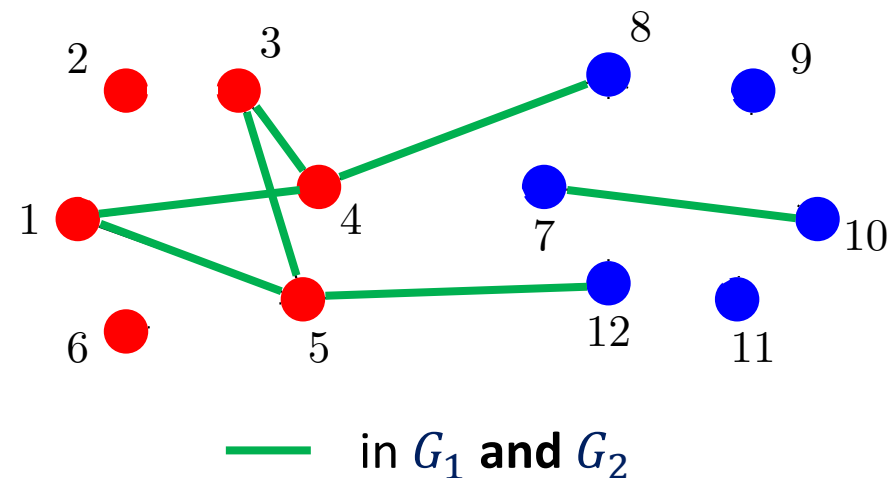
Let $\hat{\pi}(G_1, G_2)$ be a vertex mapping that maximizes the number of agreeing edges between G_1 and G_2 .

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

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



- $\hat{\pi}$ 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)

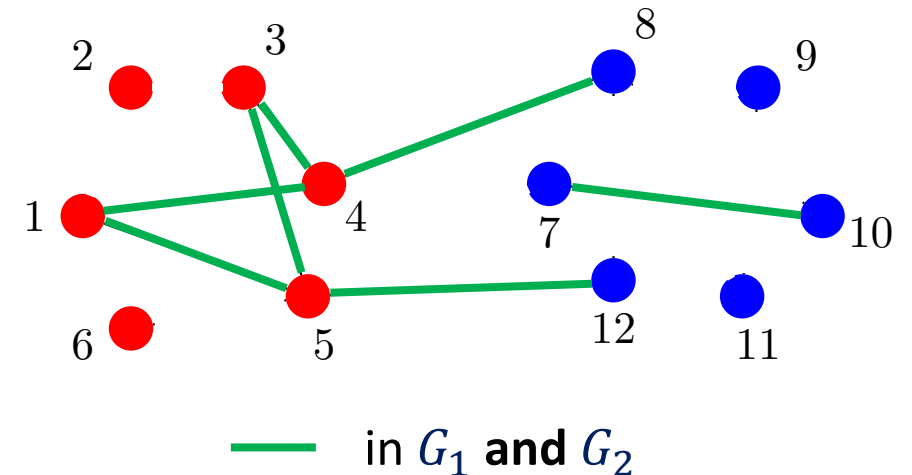
Let $\hat{\pi}(G_1, G_2)$ be a vertex mapping that maximizes the number of agreeing edges between G_1 and G_2 .

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

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



- $\hat{\pi}$ 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)

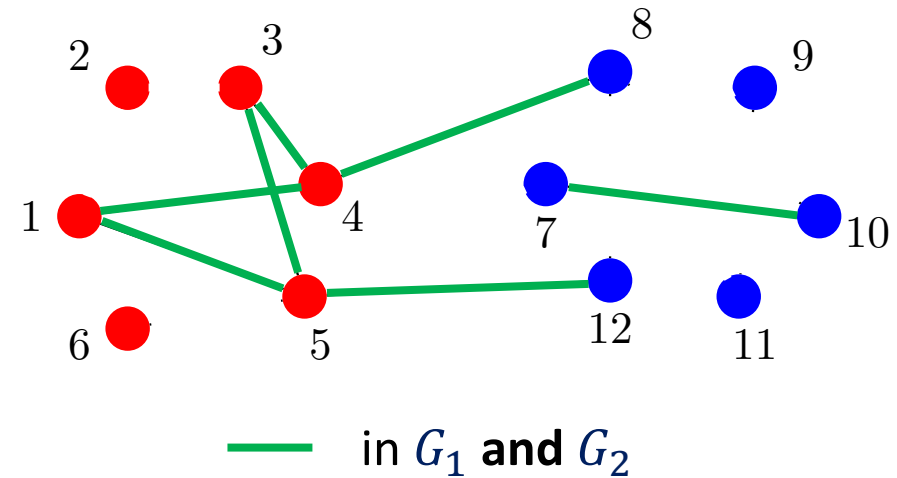
If $s^2 \left(\frac{a+b}{2} \right) < 1$ then $\lim_{n \rightarrow \infty} \mathbb{P}(\tilde{\pi}(G_1, G_2) = \pi_*) = 0$ for every estimator $\tilde{\pi}$

Exact graph matching – converse

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

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

- Condition: the intersection graph is disconnected (whp)

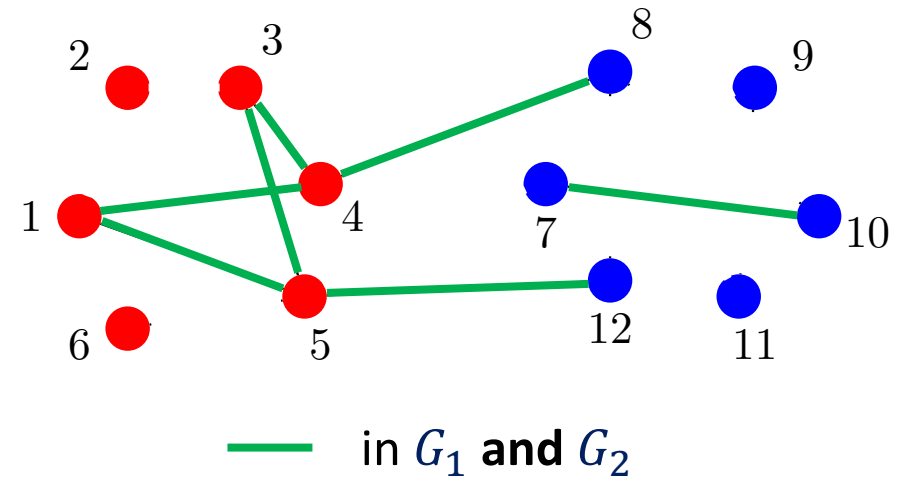


Exact graph matching – converse

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

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

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

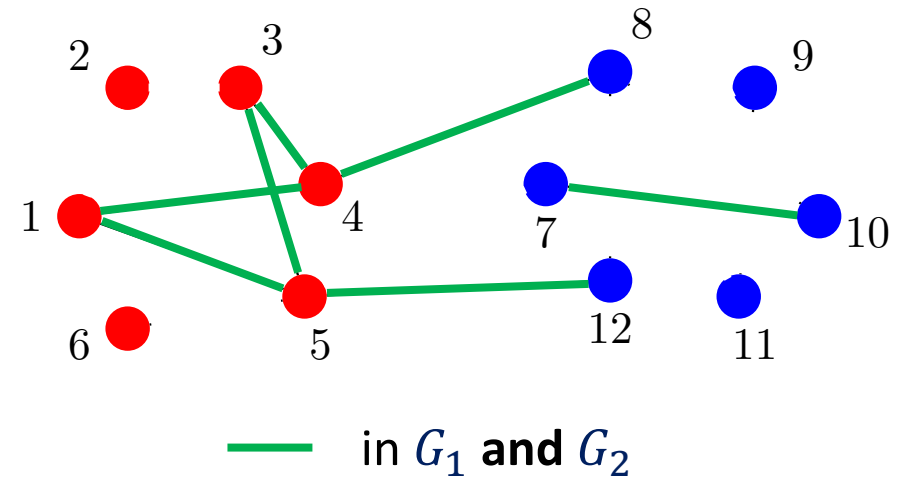


Exact graph matching – converse

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

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

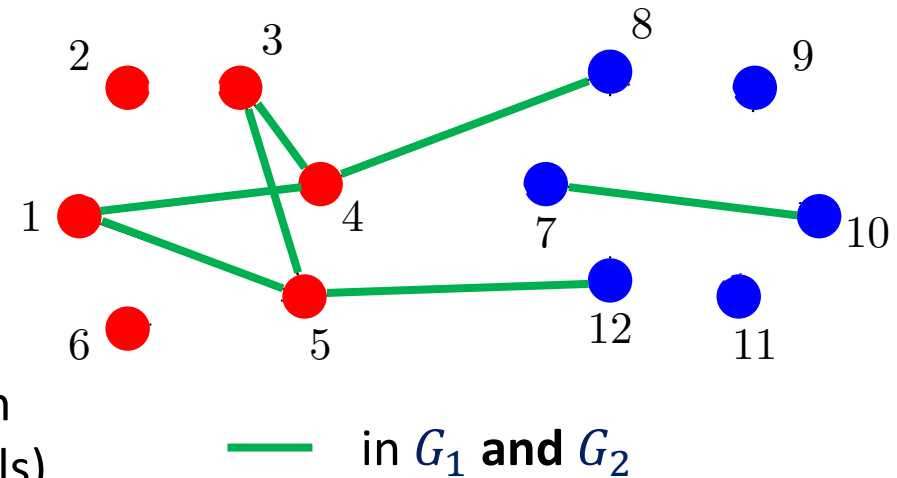


Exact graph matching – converse

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

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

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

then there is an estimator $\hat{\sigma} = \hat{\sigma}(G_1, G_2)$ such that

$$\lim_{n \rightarrow \infty} \mathbb{P}(\text{ov}(\hat{\sigma}, \sigma) = 1) = 1$$

Exact community recovery

Theorem (R., Sridhar, 2021)

Exact community recovery is **possible**

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

then there is an estimator $\hat{\sigma} = \hat{\sigma}(G_1, G_2)$ such that

$$\lim_{n \rightarrow \infty} \mathbb{P}(\text{ov}(\hat{\sigma}, \sigma) = 1) = 1$$

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**

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

then there is an estimator $\hat{\sigma} = \hat{\sigma}(G_1, G_2)$ such that

$$\lim_{n \rightarrow \infty} \mathbb{P}(\text{ov}(\hat{\sigma}, \sigma) = 1) = 1$$

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**

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

then for any estimator $\tilde{\sigma} = \tilde{\sigma}(G_1, G_2)$ we have that

$$\lim_{n \rightarrow \infty} \mathbb{P}(\text{ov}(\tilde{\sigma}, \sigma) = 1) = 0$$

Exact community recovery

Theorem (R., Sridhar, 2021)

Exact community recovery is **possible**

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

then there is an estimator $\hat{\sigma} = \hat{\sigma}(G_1, G_2)$ such that

$$\lim_{n \rightarrow \infty} \mathbb{P}(\text{ov}(\hat{\sigma}, \sigma) = 1) = 1$$

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**

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

then for any estimator $\tilde{\sigma} = \tilde{\sigma}(G_1, G_2)$ we have that

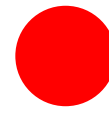
$$\lim_{n \rightarrow \infty} \mathbb{P}(\text{ov}(\tilde{\sigma}, \sigma) = 1) = 0$$

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

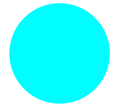
Phase diagrams



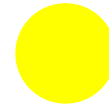
Exact community recovery possible from G_1



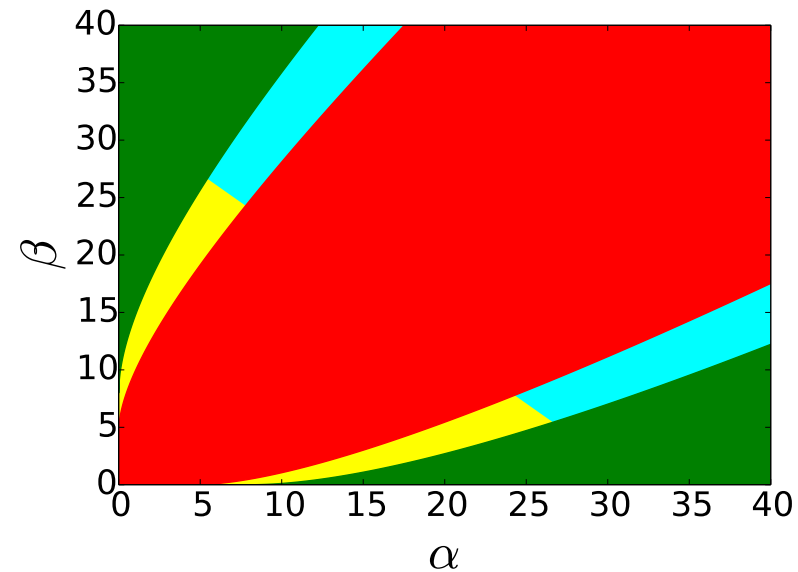
Exact community recovery impossible from (G_1, G_2)



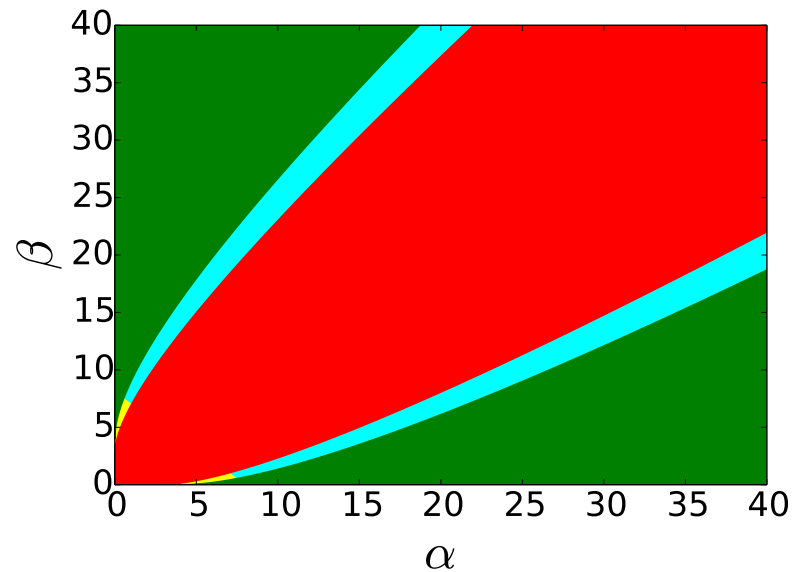
Exact community recovery impossible from G_1 , possible from (G_1, G_2)



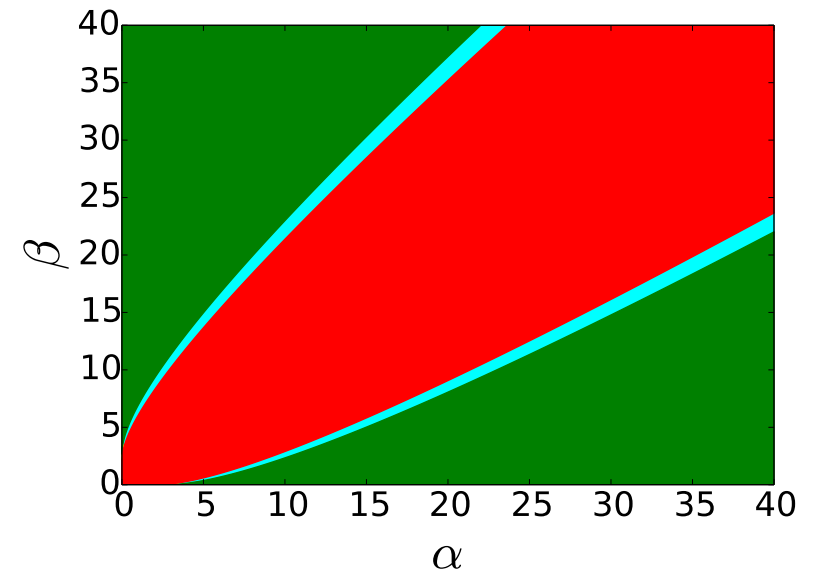
Exact community recovery impossible from G_1 , unknown from (G_1, G_2)



$s = 0.25$



$s = 0.5$

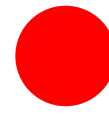


$s = 0.75$

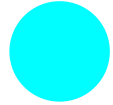
Phase diagrams



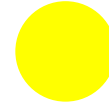
Exact community recovery possible from G_1



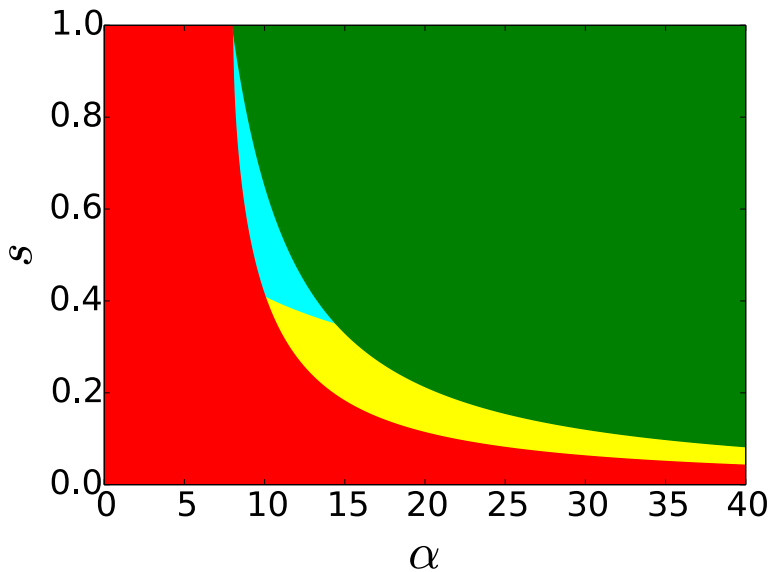
Exact community recovery impossible from (G_1, G_2)



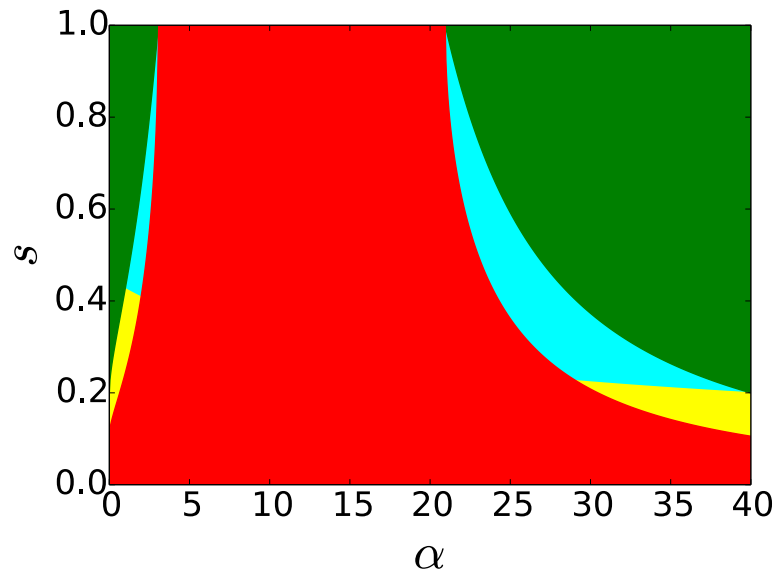
Exact community recovery impossible from G_1 , possible from (G_1, G_2)



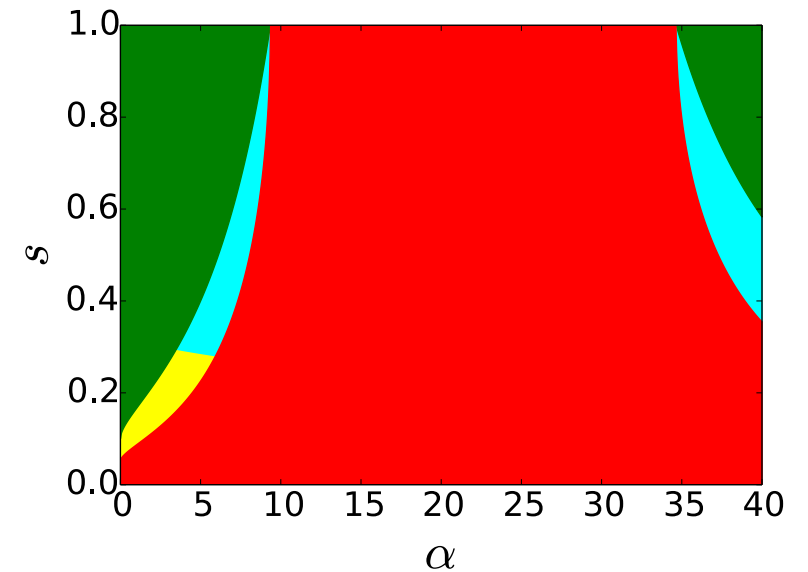
Exact community recovery impossible from G_1 , unknown from (G_1, G_2)



$b = 2$



$b = 10$

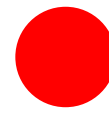


$b = 20$

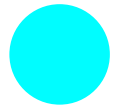
Phase diagrams



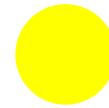
Exact community recovery possible from G_1



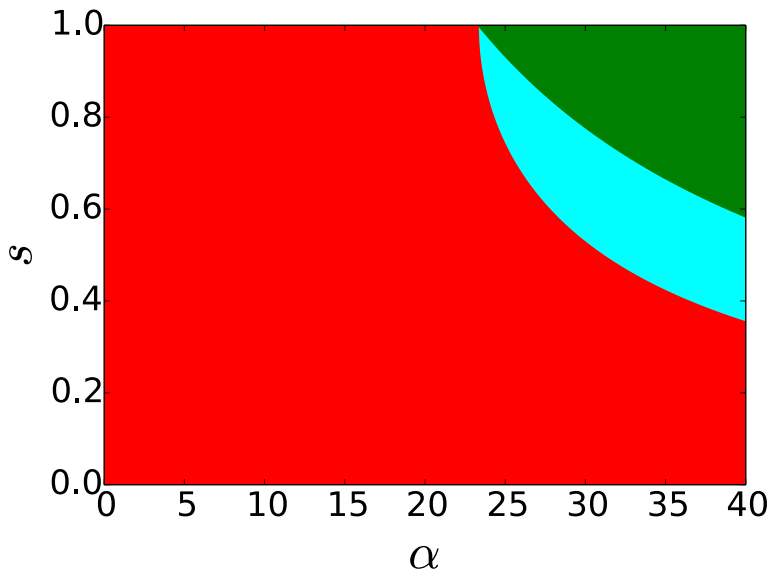
Exact community recovery impossible from (G_1, G_2)



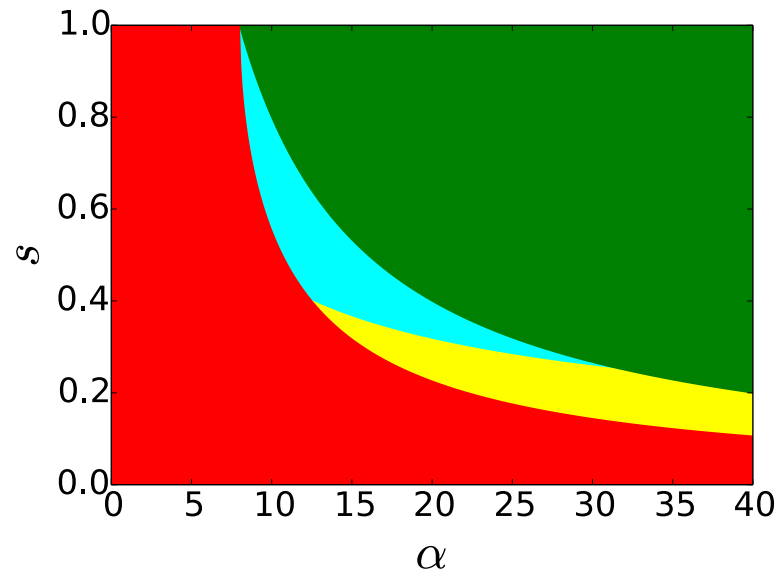
Exact community recovery impossible from G_1 , possible from (G_1, G_2)



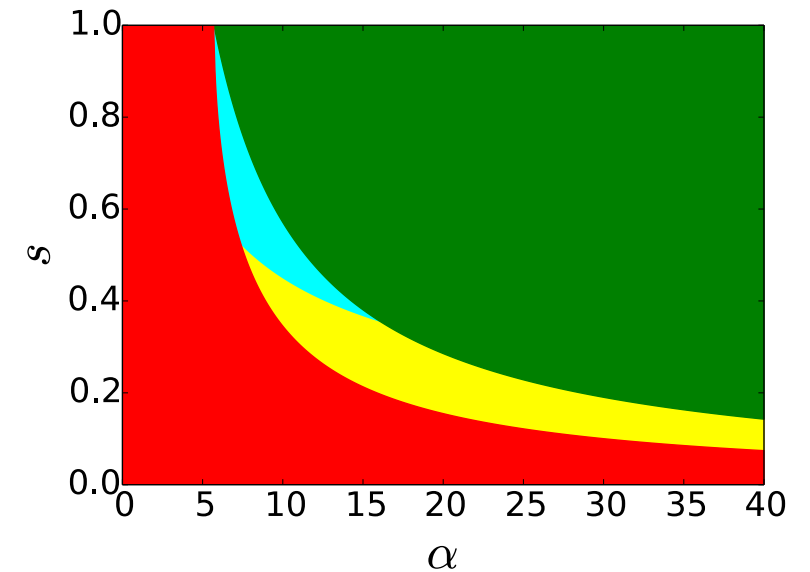
Exact community recovery impossible from G_1 , unknown from (G_1, G_2)



$a/b = 2$



$a/b = 4$



$a/b = 6$

Proof (graph matching)

Estimator

A, B : adjacency matrices of G_1, G_2

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

Estimator

A, B : adjacency matrices of G_1, G_2

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

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

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



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

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

Estimator

A, B : adjacency matrices of G_1, G_2

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

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



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

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

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

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

Estimator

A, B : adjacency matrices of G_1, G_2

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

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



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

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

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

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

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_+ (relative to π_*)
- k_2 vertices are mismatched in V_-

Permutations

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

- k_1 vertices are mismatched in V_+ (relative to π_*)
- k_2 vertices are mismatched in V_-

From vertex mismatches to edge mismatches:

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

Permutations

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

- k_1 vertices are mismatched in V_+ (relative to π_*)
- k_2 vertices are mismatched in V_-

From vertex mismatches to edge mismatches: $M^+(\tau) := |\{e \in \mathcal{E}^+(\sigma) : \tau(e) \neq \tau_*(e)\}|$
 $M^-(\tau) := |\{e \in \mathcal{E}^-(\sigma) : \tau(e) \neq \tau_*(e)\}|$

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_+ (relative to π_*)
- k_2 vertices are mismatched in V_-

From vertex mismatches to edge mismatches: $M^+(\tau) := |\{e \in \mathcal{E}^+(\sigma) : \tau(e) \neq \tau_*(e)\}|$
 $M^-(\tau) := |\{e \in \mathcal{E}^-(\sigma) : \tau(e) \neq \tau_*(e)\}|$

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

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

Analysis

Claim

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

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

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

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

Proof sketch:

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

$$\begin{aligned} \mathbb{P}(\hat{\tau} = \tau \mid \boldsymbol{\sigma}, \tau_*) &\leq \mathbb{P}(X(\tau) \leq 0 \mid \boldsymbol{\sigma}, \tau_*) = \mathbb{P}\left(n^{-X(\tau)/2} \geq 1 \mid \boldsymbol{\sigma}, \tau_*\right) \\ &\leq \mathbb{E}\left[\left(1/\sqrt{n}\right)^{X(\tau)} \mid \boldsymbol{\sigma}, \tau_*\right] \end{aligned}$$

Analysis

Claim

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

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

Proof sketch:

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

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

Generating function

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

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

$$Y^+(\tau) := \sum_{e \in \mathcal{E}^+(\sigma) : \tau(e) \neq \tau_*(e)} A_e B_{\tau_*(e)},$$

$$Y^-(\tau) := \sum_{e \in \mathcal{E}^-(\sigma) : \tau(e) \neq \tau_*(e)} A_e B_{\tau_*(e)}.$$

Joint generating function

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

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

Generating function

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

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

$$Y^+(\tau) := \sum_{e \in \mathcal{E}^+(\sigma) : \tau(e) \neq \tau_*(e)} A_e B_{\tau_*(e)},$$

$$Y^-(\tau) := \sum_{e \in \mathcal{E}^-(\sigma) : \tau(e) \neq \tau_*(e)} A_e B_{\tau_*(e)}.$$

Joint generating function

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

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:

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

Generating function

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

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

$$Y^+(\tau) := \sum_{e \in \mathcal{E}^+(\sigma) : \tau(e) \neq \tau_*(e)} A_e B_{\tau_*(e)},$$

$$Y^-(\tau) := \sum_{e \in \mathcal{E}^-(\sigma) : \tau(e) \neq \tau_*(e)} A_e B_{\tau_*(e)}.$$

Joint generating function

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

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:

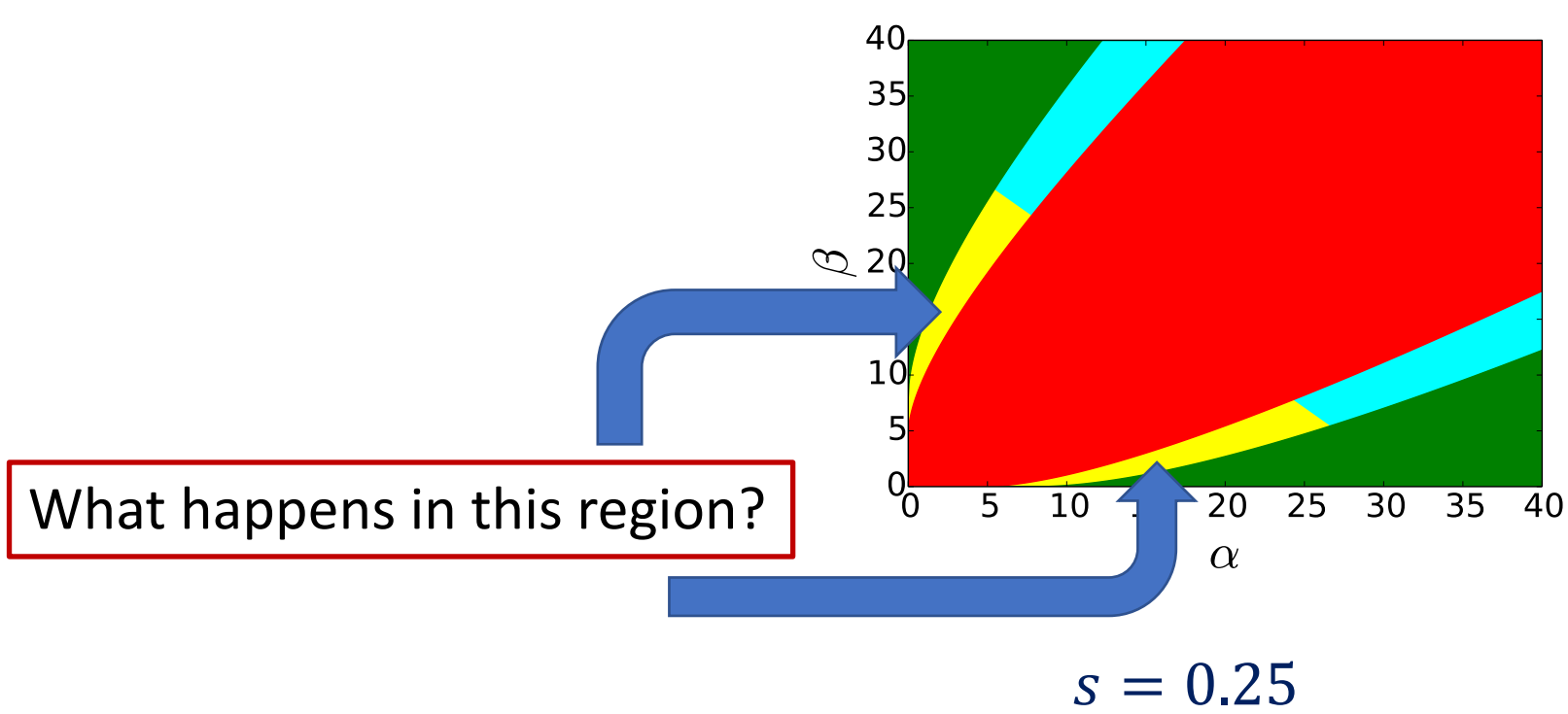
$$\Phi^\tau(1/\sqrt{n}, \omega, \zeta) \leq \exp \left(-(1-\varepsilon)s^2 (\alpha M^+(\tau) + \beta M^-(\tau)) \frac{\log n}{n} \right)$$

Analysis:

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

Open problems / future directions

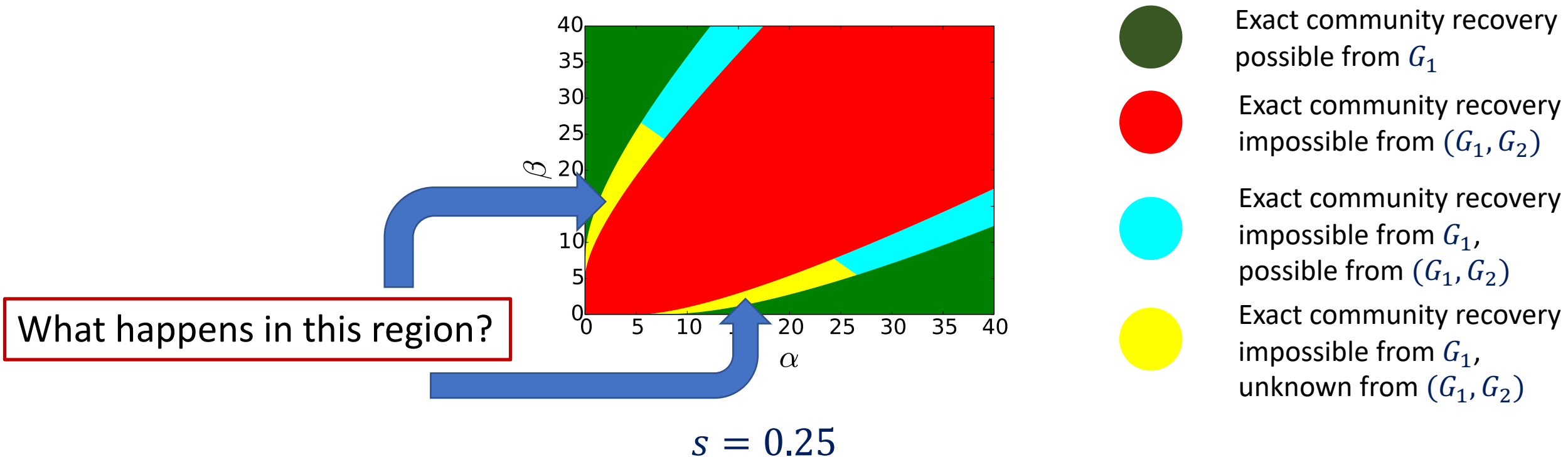
Closing the gap for exact community recovery



- 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 , unknown from (G_1, G_2)

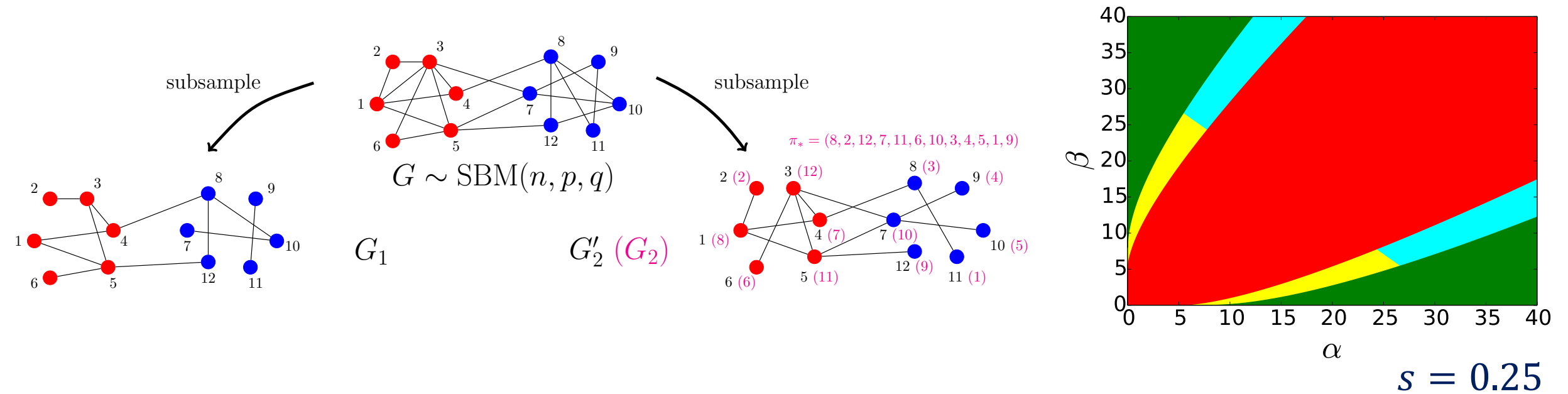
- Exact graph matching is impossible
- Q:** is exact community recovery from (G_1, G_2) possible?

Closing the gap for exact community recovery



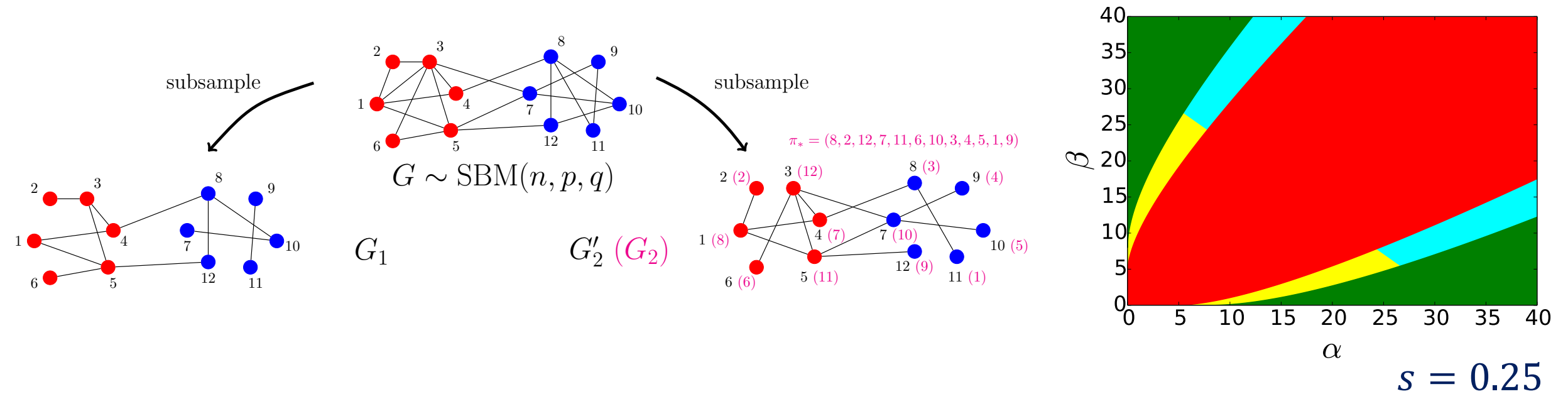
- Exact graph matching is impossible
- **Q:** is exact community recovery from (G_1, G_2) possible?
- **Conjecture:** in part of the region **yes**, in part of it **no**

Efficient algorithms



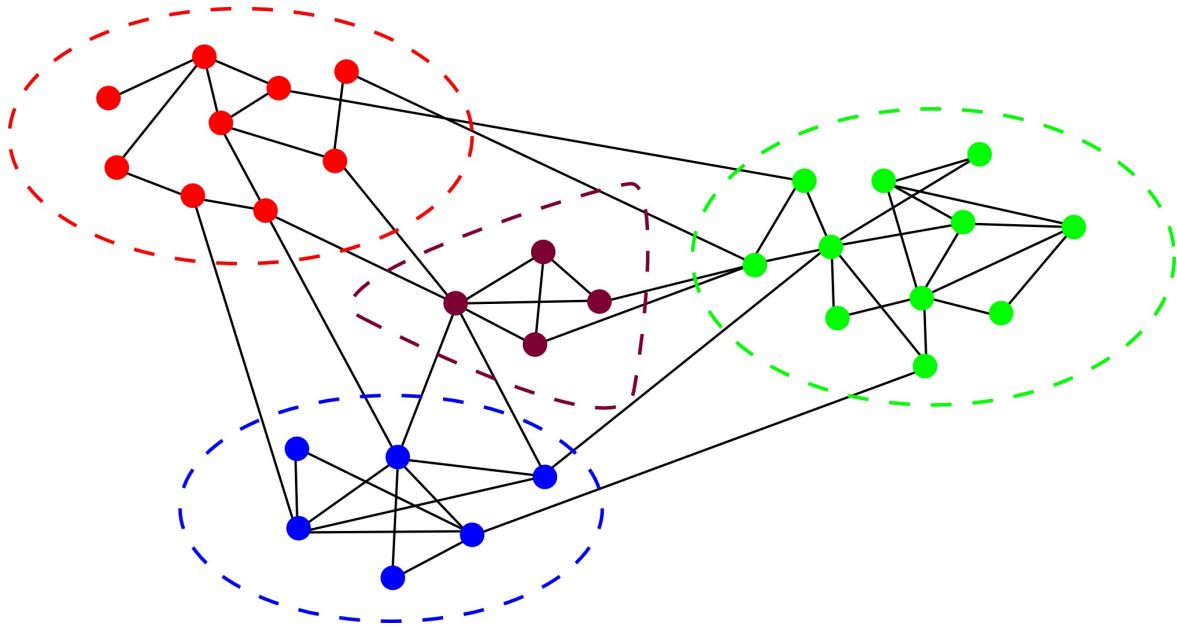
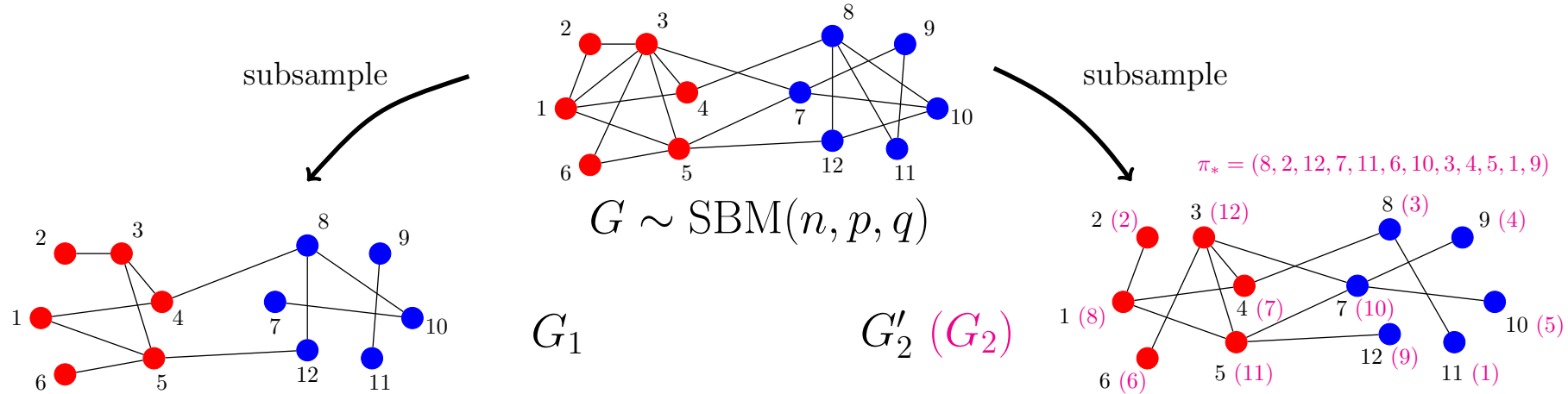
- Current algorithm for exact graph matching is not efficient
- Do there exist efficient algorithms for graph matching?

Efficient algorithms



- Current algorithm for exact graph matching is not efficient
- Do there exist efficient algorithms for graph matching?
- If not:
 - Is it possible to recover communities exactly using a polynomial time relaxation of the graph matching subroutine?

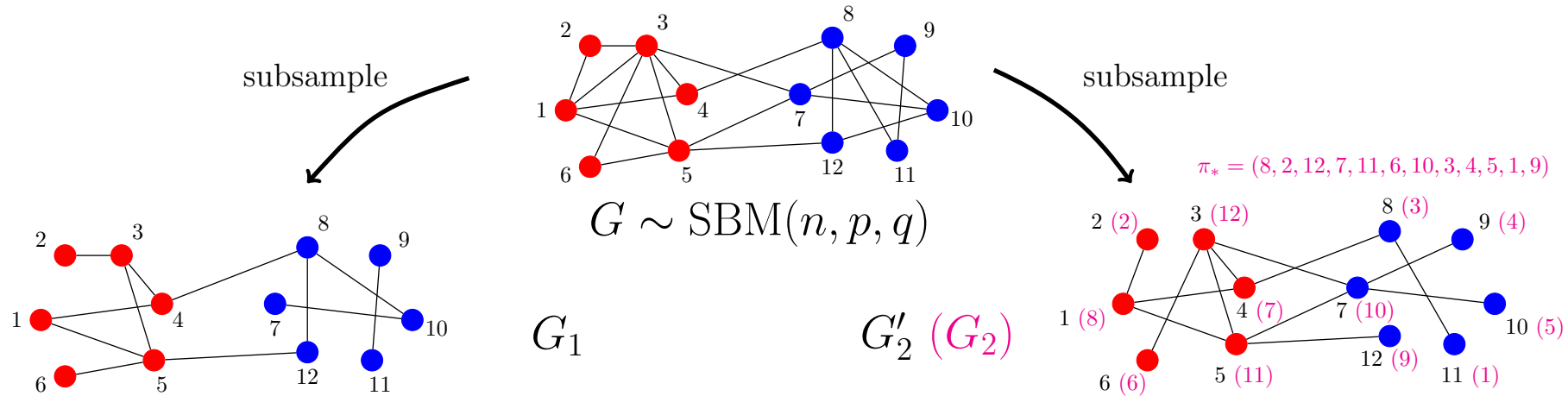
General correlated SBMs



k communities, general parameters:

- graph matching?
- (exact) community recovery?

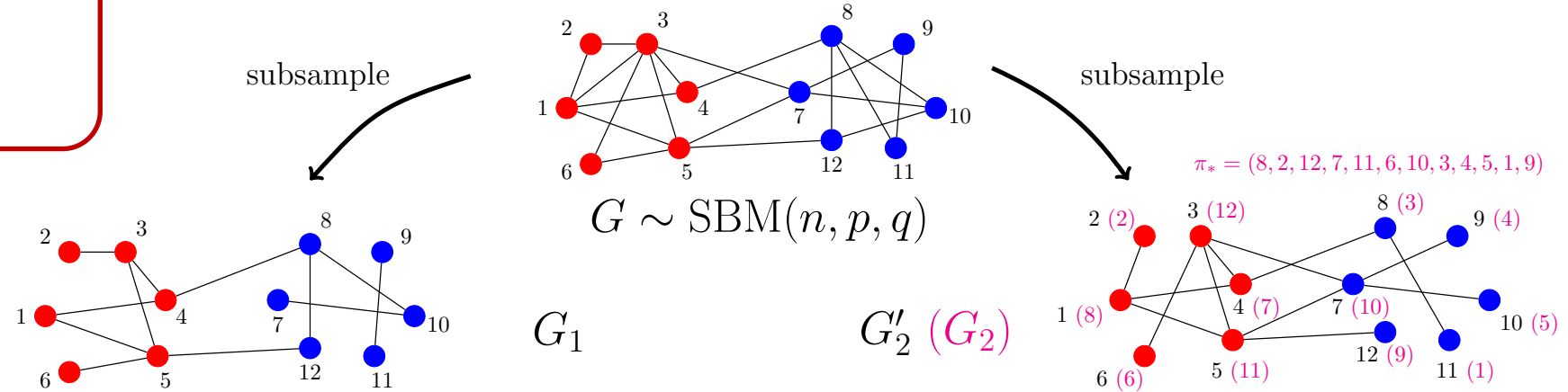
Beyond exact community recovery



- Partial recovery?
- Community detection?

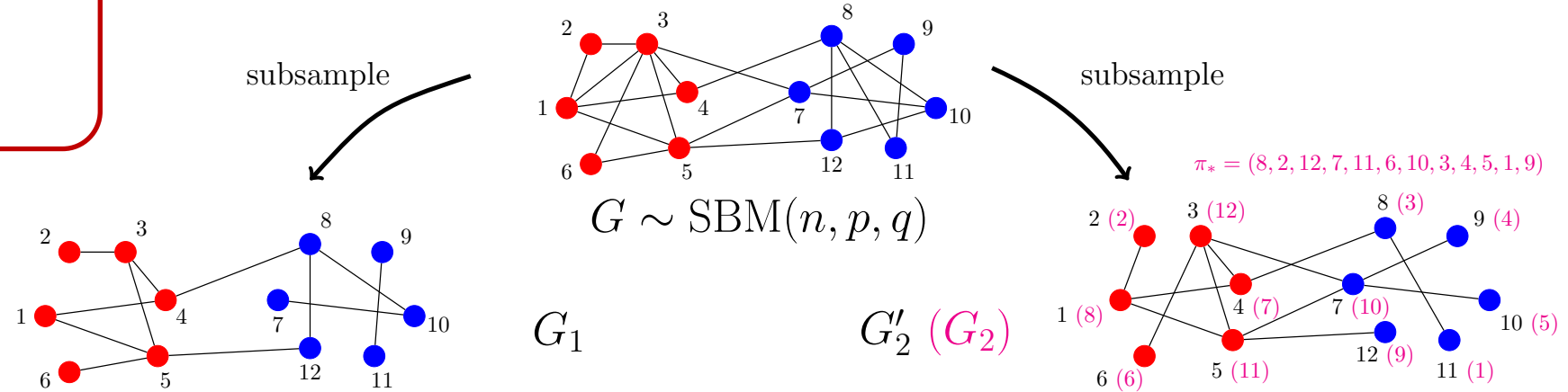
Challenge: in these regimes exact graph matching is impossible; cannot use as a black box

Summary



- **Correlated SBMs:** determined the fundamental limits of **exact graph matching**
- **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**
- **Exact community recovery** possible in regimes where it is not possible from G_1 alone
- **Correlated random graphs:** many challenges and applications

Thank you!