Network inference

Part 1

Tiago P. Peixoto

University of Bath

Tehran, August 2018

Introductory text:

"Bayesian stochastic blockmodeling", arXiv: 1705.10225

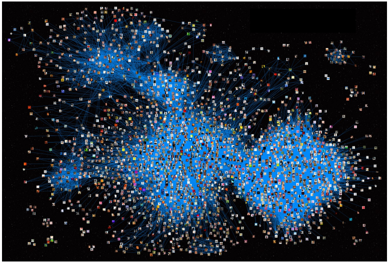
For code, see:

`https://graph-tool.skewed.de`

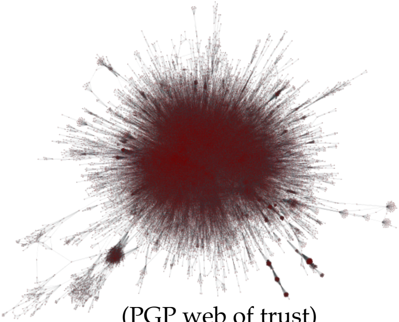
(See also HOWTO at: `https://graph-tool.skewed.de/static/doc/demos/inference/inference.html`)

More infos at: `https://skewed.de/tiago`

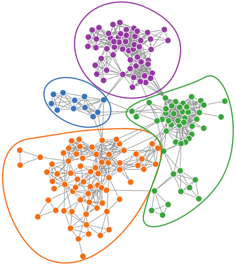
How to characterize large-scale structures?



(Flickr social network)



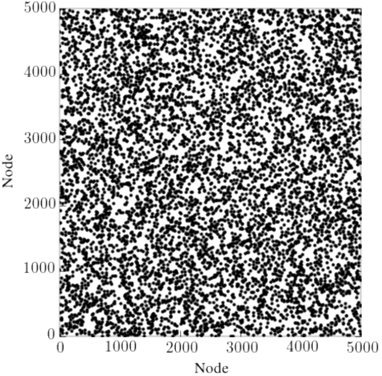
(PGP web of trust)



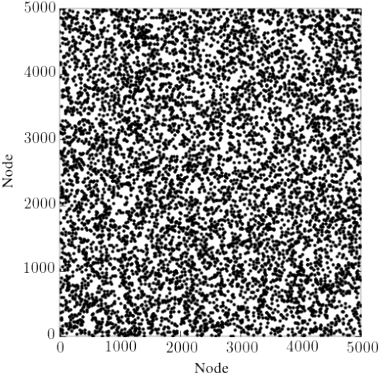
Find a meaningful network partition that provides:

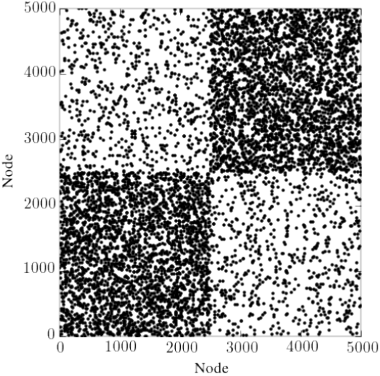
- A division of nodes into groups that share similar properties.
- An understandable summary of the large-scale structure.
- An insight on function and evolution.

Structure vs. Randomness

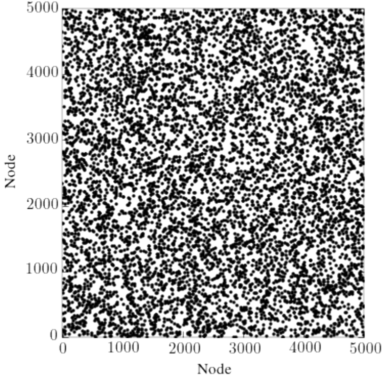


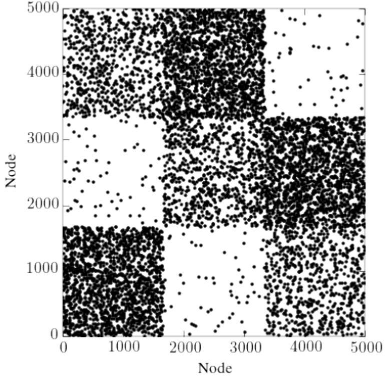
Structure vs. Randomness



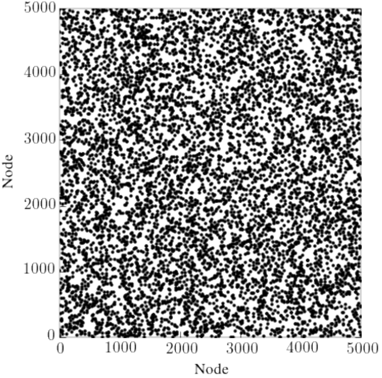


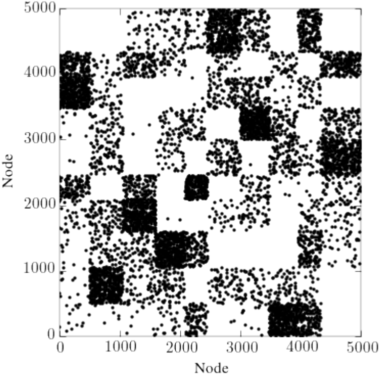
Structure vs. Randomness



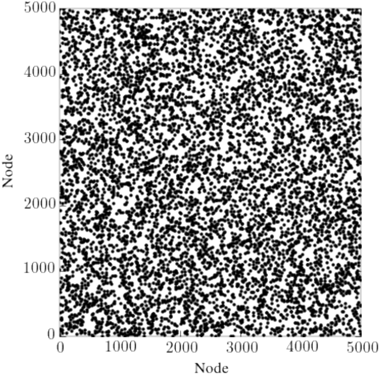


Structure vs. Randomness

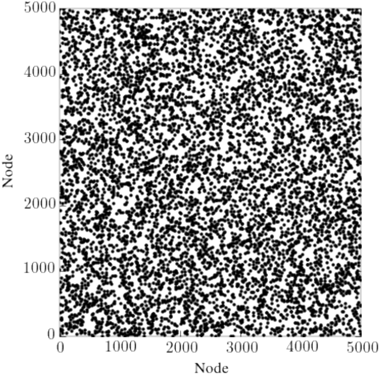




Structure vs. Randomness

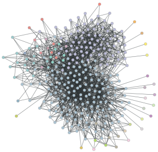


Structure vs. Randomness

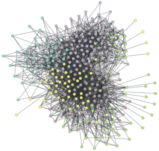


We need *well defined*, **principled** methodology, grounded on robust concepts of **statistical inference**.

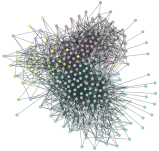
Different methods, different results...



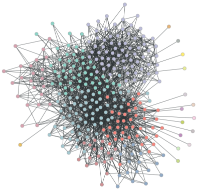
Betweenness



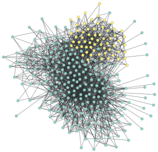
Modularity matrix



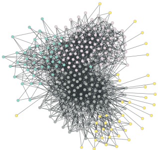
Infomap



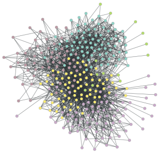
Walk Trap



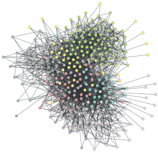
Label propagation



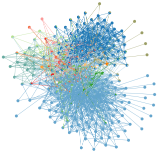
Modularity



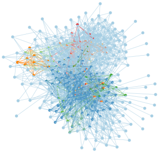
Modularity
(Blondel)



Spin glass



Infomap
(overlapping)



Clique percolation

A principled alternative: Statistical inference

What we have: $\boldsymbol{A} = \{A_{ij}\} \to$ Network

What we want: $\boldsymbol{b} = \{b_i\}, b_i \in \{1, \dots, B\} \to$ Partition into groups

Generative model

$P(\boldsymbol{A}|\boldsymbol{\theta}, \boldsymbol{b}) \to$ Model likelihood

$\boldsymbol{\theta} \to$ Extra model parameters

Bayesian posterior

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{A})} \quad \text{(Bayes' rule)}$$

$$P(\boldsymbol{A}|\boldsymbol{b}) = \int P(\boldsymbol{A}|\boldsymbol{\theta}, \boldsymbol{b})P(\boldsymbol{\theta}|\boldsymbol{b})\,\mathrm{d}\boldsymbol{\theta} \quad \text{(Marginal likelihood)}$$

$$P(\boldsymbol{\theta}, \boldsymbol{b}) = P(\boldsymbol{\theta}|\boldsymbol{b})P(\boldsymbol{b}) \quad \text{(Prior probability)}$$

$$P(\boldsymbol{A}) = \sum_{\boldsymbol{b}} P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b}) \quad \text{(Evidence)}$$

Why statistical inference?

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{A})} \qquad \text{(Bayes' rule)}$$

$$P(\boldsymbol{A}|\boldsymbol{b}) = \int P(\boldsymbol{A}|\boldsymbol{\theta}, \boldsymbol{b})P(\boldsymbol{\theta}|\boldsymbol{b})\,\mathrm{d}\boldsymbol{\theta}$$

Nonparametric

- Dimension of the model (i.e. number of groups B) can be inferred from data.
- Implements Occam's razor and prevents overfitting.

Model selection

Different model classes, $\mathcal{C}_1, \mathcal{C}_3, \mathcal{C}_3, \ldots$

Posterior odds ratio:

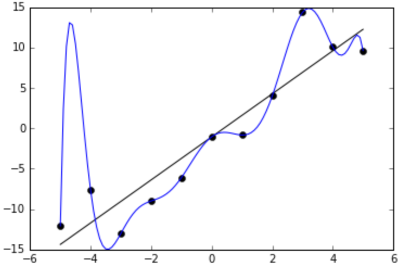
$$\Lambda = \frac{P(\boldsymbol{b}_1, \mathcal{C}_1|\boldsymbol{A})}{P(\boldsymbol{b}_2, \mathcal{C}_2|\boldsymbol{A})} = \frac{P(\boldsymbol{A}|\boldsymbol{b}_1, \mathcal{C}_1)P(\boldsymbol{b}_1)P(\mathcal{C}_1)}{P(\boldsymbol{A}|\boldsymbol{b}_2, \mathcal{C}_2)P(\boldsymbol{b}_2)P(\mathcal{C}_2)}$$

If $\Lambda > 1$, $(\boldsymbol{b}_1, \mathcal{C}_1)$ should be preferred over $(\boldsymbol{b}_2, \mathcal{C}_2)$.

Overfitting, regularization and Occam's razor

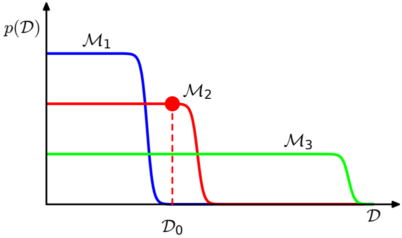
$$P(\boldsymbol{\theta}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathcal{D})}$$

$$\mathcal{D} \to \text{Data}, \boldsymbol{\theta} \to \text{Parameters}$$



Evidence:

$$P(\mathcal{D}) = \int P(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$$



Two approaches: Parametric vs. non-parametric

Parametric

$$P(\boldsymbol{b}|\boldsymbol{A},\boldsymbol{\theta}) = \frac{P(\boldsymbol{A}|\boldsymbol{b},\boldsymbol{\theta})P(\boldsymbol{b})}{P(\boldsymbol{A}|\boldsymbol{\theta})}$$

$\boldsymbol{\theta} \to$ Remaining parameters are assumed to be known.

- Dimension of the model cannot be determined from data.
- Parameters $\boldsymbol{\theta}$ are not typically known in practice.
- Problem has a simpler form, and becomes analytically tractable.
- Yields deeper understanding of the problem, and uncovers fundamental limits.

(Lenka Zdeborová's lectures.)

Non-parametric

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{A})}$$

$$P(\boldsymbol{A}|\boldsymbol{b}) = \int P(\boldsymbol{A}|\boldsymbol{\theta},\boldsymbol{b})P(\boldsymbol{\theta}|\boldsymbol{b})\,\mathrm{d}\boldsymbol{\theta}$$

$\boldsymbol{\theta} \to$ Remaining parameters are unknown.

- Dimension of the model can be determined from data.
- Does not require knowledge of $\boldsymbol{\theta}$.
- Enables model selection, and prevents overfitting.
- Problem has a more complicated form, not amenable to the same tools used for the parametric case.

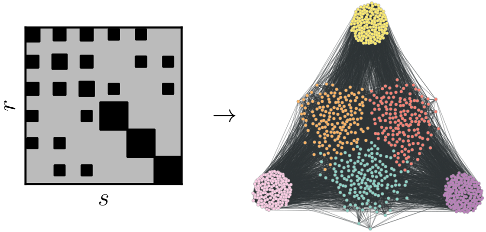
(Our focus.)

THE STOCHASTIC BLOCK MODEL (SBM)

Planted partition: N nodes divided into B groups (blocks).

Parameters: $b_i \rightarrow$ group membership of node i

$\lambda_{rs} \rightarrow$ edge probability from group r to s.



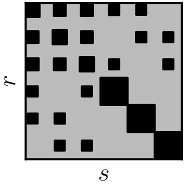
- Not restricted to assortative structures ("communities").
- Easily generalizable (edge direction, overlapping groups, etc.)

Bayesian statistical inference

Network probability: $P(\boldsymbol{A}|\boldsymbol{\lambda}, \boldsymbol{b})$

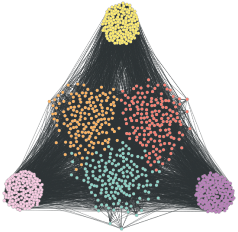
$\boldsymbol{A} \to$ Observed network

$\boldsymbol{\lambda}, \boldsymbol{b} \to$ Model parameters



$\boldsymbol{b}, \boldsymbol{\lambda}$

$\xrightarrow{}$ $P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{b})$



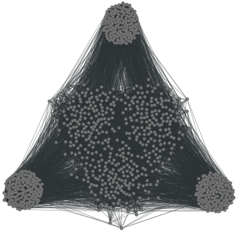
$\boldsymbol{A}$

Bayesian statistical inference

Network probability: $P(\boldsymbol{A}|\boldsymbol{\lambda}, \boldsymbol{b})$

$\boldsymbol{A} \to$ Observed network

$\boldsymbol{\lambda}, \boldsymbol{b} \to$ Model parameters



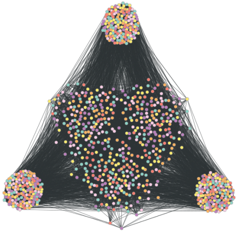
$\boldsymbol{A}$

Bayesian statistical inference

Network probability: $P(\boldsymbol{A}|\boldsymbol{\lambda}, \boldsymbol{b})$

$\boldsymbol{A} \to$ Observed network

$\boldsymbol{\lambda}, \boldsymbol{b} \to$ Model parameters



$\boldsymbol{A}$

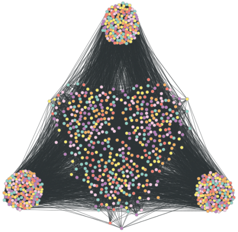
Bayesian statistical inference

Network probability: $P(\boldsymbol{A}|\boldsymbol{\lambda}, \boldsymbol{b})$

$\boldsymbol{A} \rightarrow$ Observed network

$\boldsymbol{\lambda}, \boldsymbol{b} \rightarrow$ Model parameters

$\underset{P(\boldsymbol{b}|\boldsymbol{A})}{\longleftarrow}$



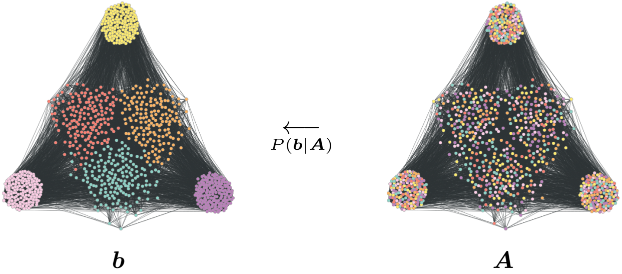
$\boldsymbol{A}$

Bayesian statistical inference

Network probability: $P(\boldsymbol{A}|\boldsymbol{\lambda}, \boldsymbol{b})$

$\boldsymbol{A} \rightarrow$ Observed network

$\boldsymbol{\lambda}, \boldsymbol{b} \rightarrow$ Model parameters

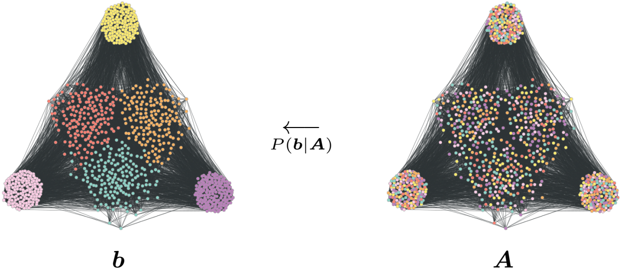


Bayesian statistical inference

Network probability: $P(\boldsymbol{A}|\boldsymbol{\lambda}, \boldsymbol{b})$

$\boldsymbol{A} \rightarrow$ Observed network

$\boldsymbol{\lambda}, \boldsymbol{b} \rightarrow$ Model parameters



$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{A})} \qquad \text{(Bayes' rule)}$$

$$P(\boldsymbol{A}|\boldsymbol{b}) = \int P(\boldsymbol{A}|\boldsymbol{\lambda}, \boldsymbol{b})P(\boldsymbol{\lambda}|\boldsymbol{b})\,\mathrm{d}\boldsymbol{\lambda}$$

The Poisson SBM

Model likelihood:

$$P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{b}) = \prod_{i<j} \frac{\lambda_{b_i b_j}^{A_{ij}} e^{-\lambda_{b_i b_j}}}{A_{ij}!} \times \prod_i \frac{(\lambda_{b_i b_i}/2)^{A_{ii}/2} e^{-\lambda_{b_i b_i}/2}}{(A_{ii}/2)!}$$

The Poisson SBM

Model likelihood:

$$P(\boldsymbol{A}|\boldsymbol{\lambda}, \boldsymbol{b}) = \prod_{i<j} \frac{\lambda_{b_i b_j}^{A_{ij}} e^{-\lambda_{b_i b_j}}}{A_{ij}!} \times \prod_{i} \frac{(\lambda_{b_i b_i}/2)^{A_{ii}/2} e^{-\lambda_{b_i b_i}/2}}{(A_{ii}/2)!}$$

Noninformative prior for λ:

$$P(\boldsymbol{\lambda}|\boldsymbol{b}) = \prod_{r<s} \frac{n_r n_s}{\bar{\lambda}} e^{-n_r n_s \lambda_{rs}/\bar{\lambda}} \times \prod_{r} \frac{n_r^2}{2\bar{\lambda}} e^{-n_r^2 \lambda_{rs}/2\bar{\lambda}}$$

The Poisson SBM

Model likelihood:

$$P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{b}) = \prod_{i<j}\frac{\lambda_{b_ib_j}^{A_{ij}}e^{-\lambda_{b_ib_j}}}{A_{ij}!}\times\prod_i\frac{(\lambda_{b_ib_i}/2)^{A_{ii}/2}e^{-\lambda_{b_ib_i}/2}}{(A_{ii}/2)!}$$

Noninformative prior for λ:

$$P(\boldsymbol{\lambda}|\boldsymbol{b}) = \prod_{r<s}\frac{n_rn_s}{\bar{\lambda}}e^{-n_rn_s\lambda_{rs}/\bar{\lambda}}\times\prod_r\frac{n_r^2}{2\bar{\lambda}}e^{-n_r^2\lambda_{rs}/2\bar{\lambda}}$$

Marginal likelihood:

$$\begin{aligned}P(\boldsymbol{A}|\boldsymbol{b}) &= \int P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{b})P(\boldsymbol{\lambda}|\boldsymbol{b})\,\mathrm{d}\boldsymbol{\lambda}\\ &= \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}}\times\frac{\prod_{r<s}e_{rs}!\prod_r e_{rr}!!}{\prod_r n_r^{e_r}\prod_{i<j}A_{ij}!\prod_i A_{ii}!!}\end{aligned}$$

The Poisson SBM

Model likelihood:

$$P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{b}) = \prod_{i<j} \frac{\lambda_{b_i b_j}^{A_{ij}} e^{-\lambda_{b_i b_j}}}{A_{ij}!} \times \prod_i \frac{(\lambda_{b_i b_i}/2)^{A_{ii}/2} e^{-\lambda_{b_i b_i}/2}}{(A_{ii}/2)!}$$

Noninformative prior for λ:

$$P(\boldsymbol{\lambda}|\boldsymbol{b}) = \prod_{r<s} \frac{n_r n_s}{\bar{\lambda}} e^{-n_r n_s \lambda_{rs}/\bar{\lambda}} \times \prod_r \frac{n_r^2}{2\bar{\lambda}} e^{-n_r^2 \lambda_{rs}/2\bar{\lambda}}$$

Marginal likelihood:

$$P(\boldsymbol{A}|\boldsymbol{b}) = \int P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{b}) P(\boldsymbol{\lambda}|\boldsymbol{b})\, \mathrm{d}\boldsymbol{\lambda}$$

$$= \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}} \times \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!!}{\prod_r n_r^{e_r} \prod_{i<j} A_{ij}! \prod_i A_{ii}!!}$$

Prior for $\boldsymbol{b}$ (tentative):

$$P(\boldsymbol{b}|B) = \frac{1}{B^N}$$

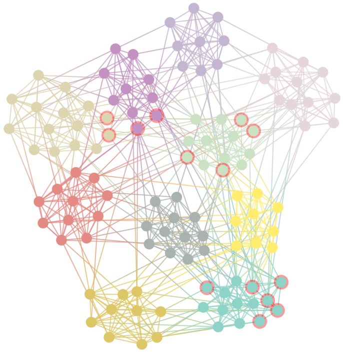
$$P(B) = 1/N$$

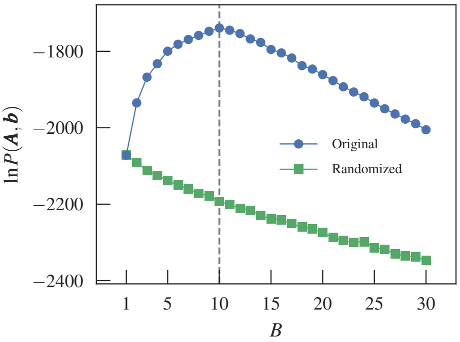
Posterior distribution:

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{A})}$$

$$P(\boldsymbol{A}) = \sum_{\boldsymbol{b}} P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})$$

Example: American college football teams





Example: The Internet Movie Database (IMDB)

Bipartite network of actors and films.

Fairly large: $N = 372,787$, $E = 1,812,657$

Example: The Internet Movie Database (IMDB)

Bipartite network of actors and films.

Fairly large: $N = 372,787$, $E = 1,812,657$

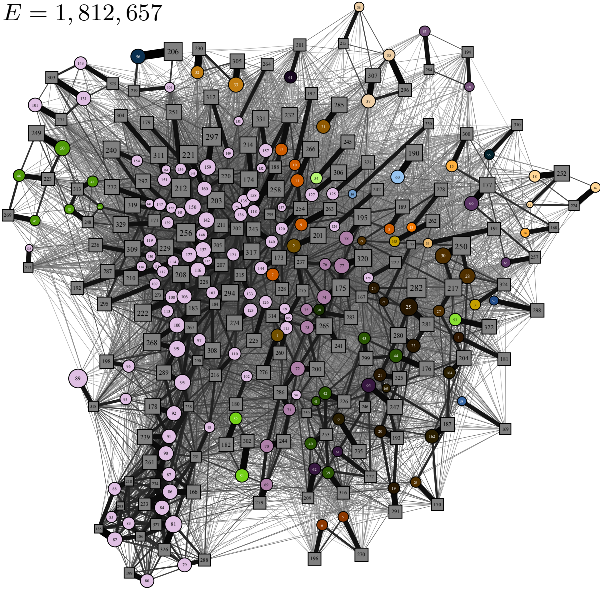
MDL selects: $B = 332$

Example: The Internet Movie Database (IMDB)

Bipartite network of actors and films.

Fairly large: $N = 372,787$, $E = 1,812,657$

MDL selects: $B = 332$

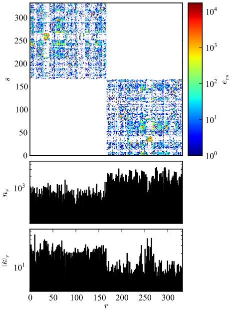


Example: The Internet Movie Database (IMDB)

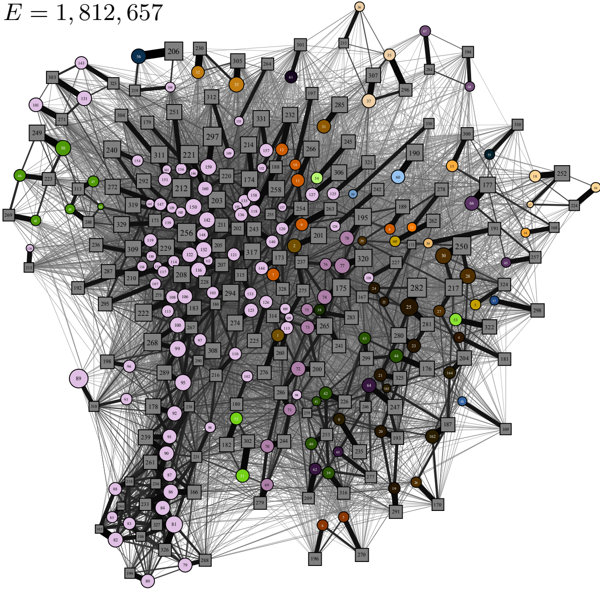
Bipartite network of actors and films.

Fairly large: $N = 372,787$, $E = 1,812,657$

MDL selects: $B = 332$



Bipartiteness is fully uncovered!

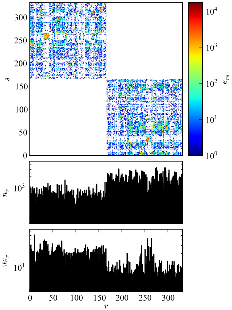


Example: The Internet Movie Database (IMDB)

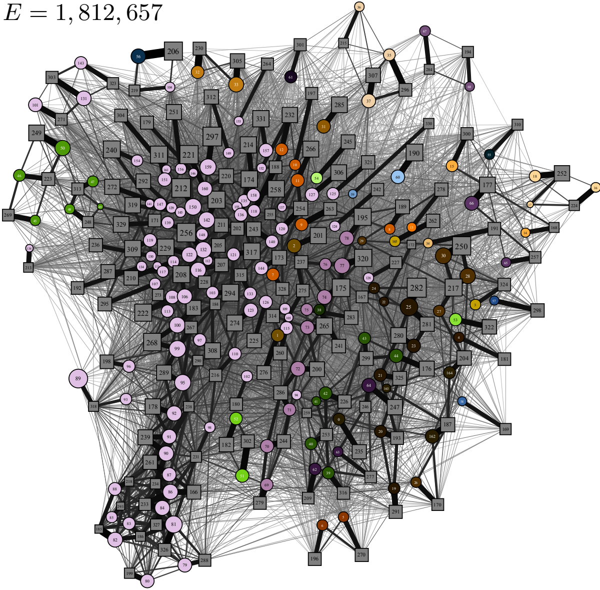
Bipartite network of actors and films.

Fairly large: $N = 372,787$, $E = 1,812,657$

MDL selects: $B = 332$



Bipartiteness is fully uncovered!



Meaningful features:

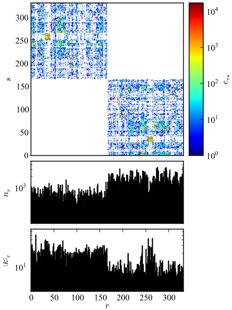
- Temporal
- Spatial (Country)
- Type/Genre

Example: The Internet Movie Database (IMDB)

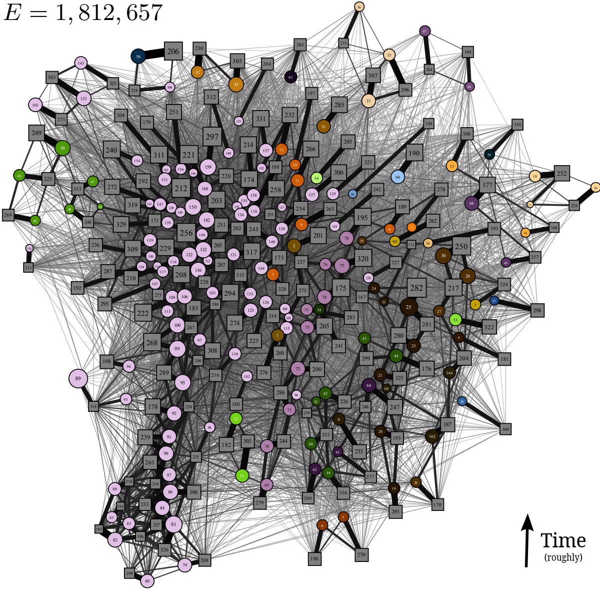
Bipartite network of actors and films.

Fairly large: $N = 372,787$, $E = 1,812,657$

MDL selects: $B = 332$



Bipartiteness is fully uncovered!



Meaningful features:

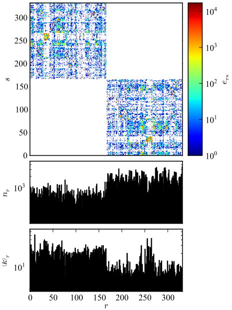
- Temporal
- Spatial (Country)
- Type/Genre

Example: The Internet Movie Database (IMDB)

Bipartite network of actors and films.

Fairly large: $N = 372,787$, $E = 1,812,657$

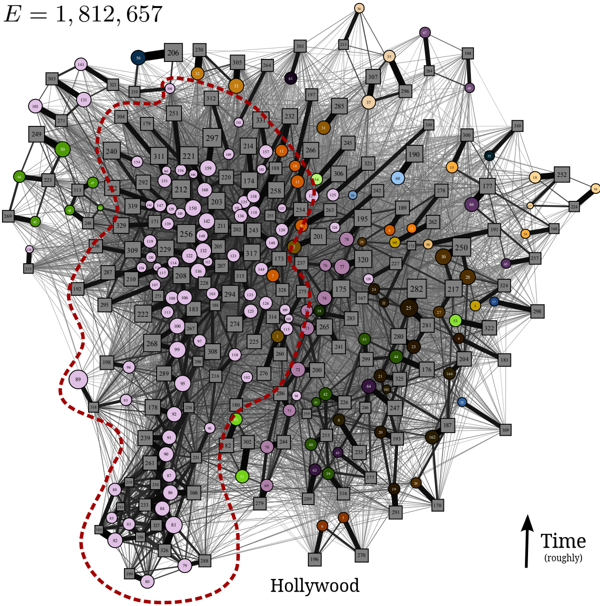
MDL selects: $B = 332$



Bipartiteness is fully uncovered!

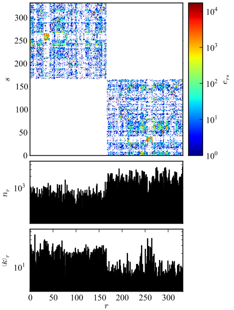
Meaningful features:

- Temporal
- Spatial (Country)
- Type/Genre



Example: The Internet Movie Database (IMDB)

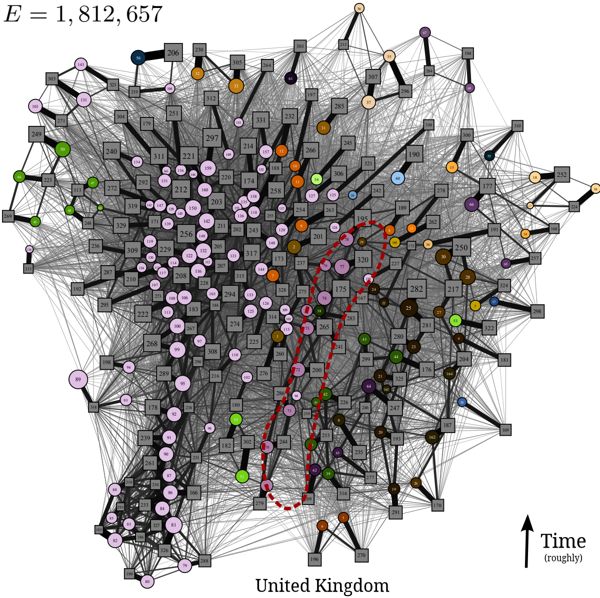
Bipartite network of actors and films.
Fairly large: $N = 372,787$, $E = 1,812,657$
MDL selects: $B = 332$



Bipartiteness is fully uncovered!

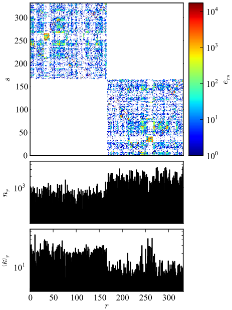
Meaningful features:

- Temporal
- Spatial (Country)
- Type/Genre



Example: The Internet Movie Database (IMDB)

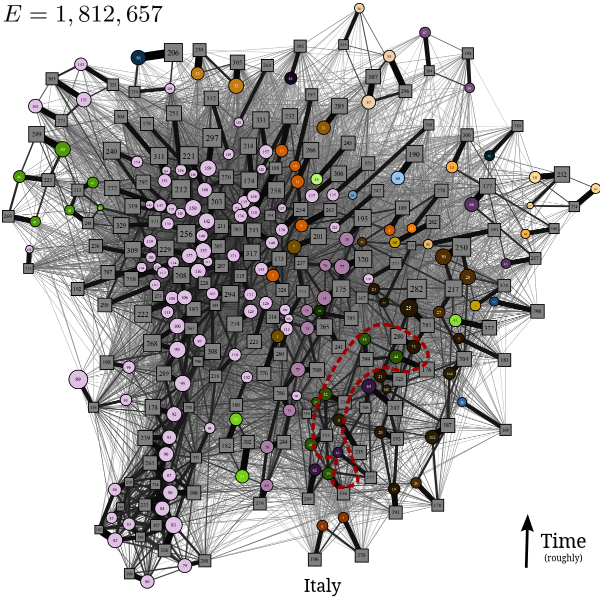
Bipartite network of actors and films.
Fairly large: $N = 372,787$, $E = 1,812,657$
MDL selects: $B = 332$



Bipartiteness is fully uncovered!

Meaningful features:

- Temporal
- Spatial (Country)
- Type/Genre

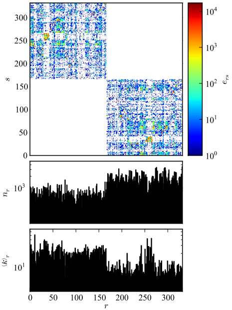


Example: The Internet Movie Database (IMDB)

Bipartite network of actors and films.

Fairly large: $N = 372,787$, $E = 1,812,657$

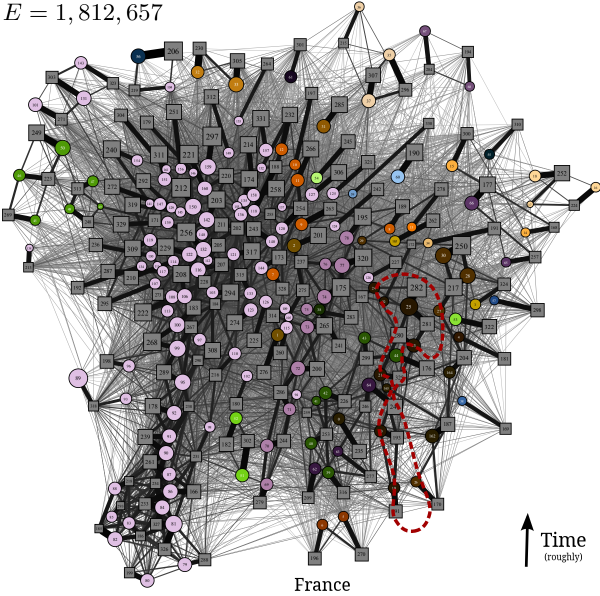
MDL selects: $B = 332$



Bipartiteness is fully uncovered!

Meaningful features:

- Temporal
- Spatial (Country)
- Type/Genre

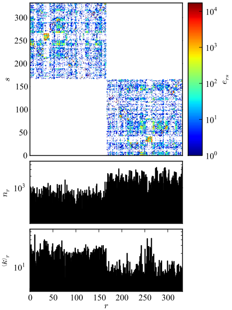


Example: The Internet Movie Database (IMDB)

Bipartite network of actors and films.

Fairly large: $N = 372,787$, $E = 1,812,657$

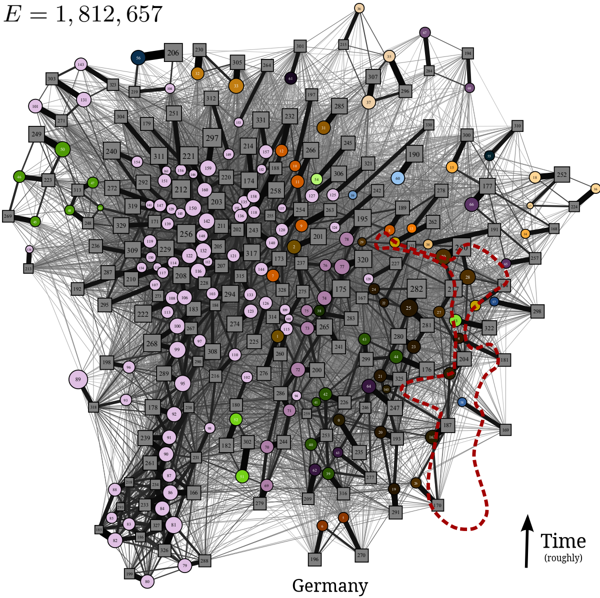
MDL selects: $B = 332$



Bipartiteness is fully uncovered!

Meaningful features:

- Temporal
- Spatial (Country)
- Type/Genre

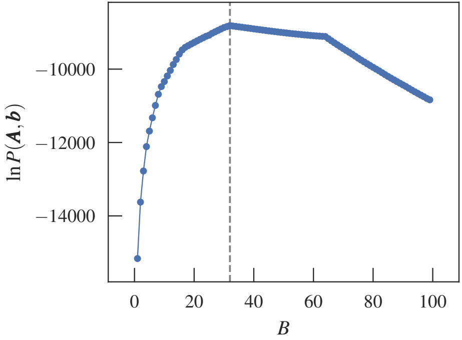


Underfitting: The “resolution limit” problem

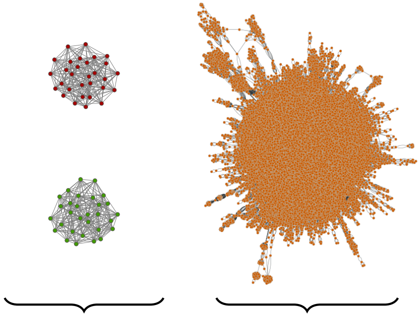
64 cliques of size 10

Underfitting: The "resolution limit" problem

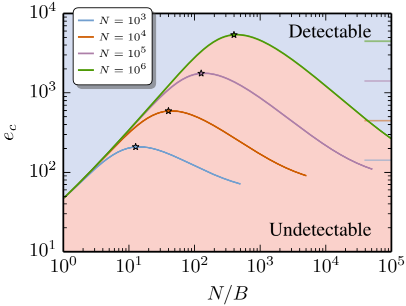
64 cliques of size 10



Underfitting: The "resolution limit" problem



Merge candidates | Remaining network



Minimum detectable block size $\sim \sqrt{N}$.

Why?

Microcanonical equivalence

Marginal likelihood:

$$
\begin{aligned}
P(\boldsymbol{A}|\boldsymbol{b}) &= \int P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{b})P(\boldsymbol{\lambda}|\boldsymbol{b})\,\mathrm{d}\boldsymbol{\lambda}\\
&= \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}} \times \frac{\prod_{r<s} e_{rs}!\prod_r e_{rr}!!}{\prod_r n_r^{e_r}\prod_{i<j} A_{ij}!\prod_i A_{ii}!!}
\end{aligned}
$$

Microcanonical equivalence

Marginal likelihood:

$$P(\boldsymbol{A}|\boldsymbol{b}) = \int P(\boldsymbol{A}|\boldsymbol{\lambda}, \boldsymbol{b}) P(\boldsymbol{\lambda}|\boldsymbol{b})\, \mathrm{d}\boldsymbol{\lambda}$$
$$= \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}} \times \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!!}{\prod_r n_r^{e_r} \prod_{i<j} A_{ij}! \prod_i A_{ii}!!}$$

Equal to the joint probability of a **microcanonical** SBM:

$$P(\boldsymbol{A}|\boldsymbol{b}) = P(\boldsymbol{A}|\boldsymbol{e}, \boldsymbol{b}) P(\boldsymbol{e}|\boldsymbol{b})$$

Microcanonical equivalence

Marginal likelihood:

$$P(\boldsymbol{A}|\boldsymbol{b}) = \int P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{b})P(\boldsymbol{\lambda}|\boldsymbol{b})\,\mathrm{d}\boldsymbol{\lambda}$$

$$= \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}} \times \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!!}{\prod_r n_r^{e_r} \prod_{i<j} A_{ij}! \prod_i A_{ii}!!}$$

Equal to the joint probability of a **microcanonical** SBM:

$$P(\boldsymbol{A}|\boldsymbol{b}) = P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{e}|\boldsymbol{b})$$

$$P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{b}) = \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!!}{\prod_r n_r^{e_r} \prod_{i<j} A_{ij}! \prod_i A_{ii}!!}$$

$$P(\boldsymbol{e}|\boldsymbol{b}) = \prod_{r<s} \frac{\bar{\lambda}^{e_{rs}}}{(\bar{\lambda}+1)^{e_{rs}+1}} \prod_r \frac{\bar{\lambda}^{e_{rs}/2}}{(\bar{\lambda}+1)^{e_{rs}/2+1}}$$

Microcanonical equivalence

Marginal likelihood:

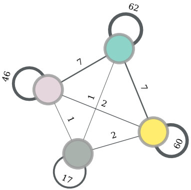
$$P(\boldsymbol{A}|\boldsymbol{b}) = \int P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{b})P(\boldsymbol{\lambda}|\boldsymbol{b})\,\mathrm{d}\boldsymbol{\lambda}$$
$$= \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}} \times \frac{\prod_{r<s} e_{rs}!\prod_r e_{rr}!!}{\prod_r n_r^{e_r}\prod_{i<j} A_{ij}!\prod_i A_{ii}!!}$$

Equal to the joint probability of a **microcanonical** SBM:

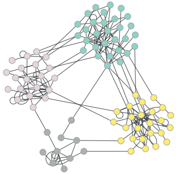
$$P(\boldsymbol{A}|\boldsymbol{b}) = P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{e}|\boldsymbol{b})$$

$$P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{b}) = \frac{\prod_{r<s} e_{rs}!\prod_r e_{rr}!!}{\prod_r n_r^{e_r}\prod_{i<j} A_{ij}!\prod_i A_{ii}!!}$$

$$P(\boldsymbol{e}|\boldsymbol{b}) = \prod_{r<s}\frac{\bar{\lambda}^{e_{rs}}}{(\bar{\lambda}+1)^{e_{rs}+1}}\prod_r\frac{\bar{\lambda}^{e_{rs}/2}}{(\bar{\lambda}+1)^{e_{rs}/2+1}}$$



Edge counts, $\boldsymbol{e}$


Network, $\boldsymbol{A}$

Minimum description length (MDL)

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

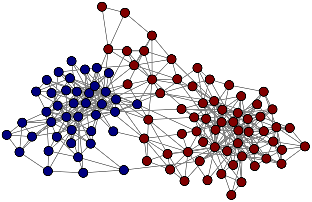
$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{b})}_{\text{data}|\text{model}, \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{e},\boldsymbol{b})}_{\text{model}, \mathcal{L}}$$

Minimum description length (MDL)

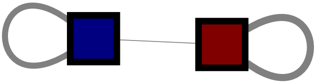
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}, \boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A}, \boldsymbol{e}, \boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A}, \boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{e}, \boldsymbol{b})}_{\text{data|model}, \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{e}, \boldsymbol{b})}_{\text{model}, \mathcal{L}}$$



$B = 2, \mathcal{S} \approx 1805.3$ bits



Model, $\mathcal{L} \approx 122.6$ bits

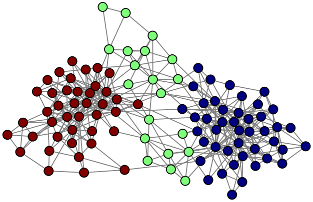
$\Sigma \approx 1927.9$ bits

Minimum description length (MDL)

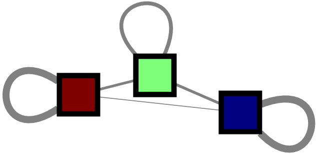
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{- \log_2 P(\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$



$B = 3$, $\mathcal{S} \approx 1688.1$ bits



Model, $\mathcal{L} \approx 203.4$ bits

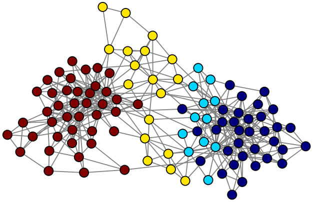
$\Sigma \approx 1891.5$ bits

Minimum description length (MDL)

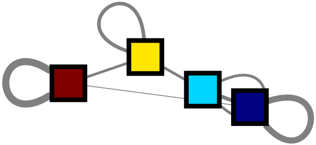
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}, \boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A}, \boldsymbol{e}, \boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A}, \boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{e}, \boldsymbol{b})}_{\text{data|model}, \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{e}, \boldsymbol{b})}_{\text{model}, \mathcal{L}}$$



$B = 4$, $\mathcal{S} \approx 1640.8$ bits



Model, $\mathcal{L} \approx 270.7$ bits

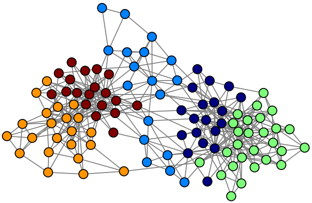
$\Sigma \approx 1911.5$ bits

Minimum description length (MDL)

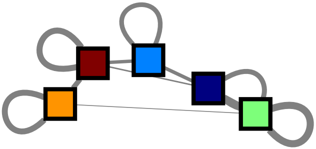
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}, \boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A}, \boldsymbol{e}, \boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A}, \boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{e}, \boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{e}, \boldsymbol{b})}_{\text{model, } \mathcal{L}}$$



$B = 5$, $\mathcal{S} \approx 1590.5$ bits



Model, $\mathcal{L} \approx 330.8$ bits

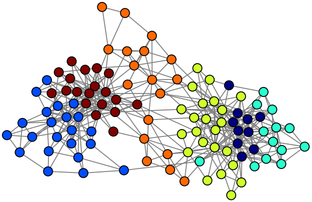
$\Sigma \approx 1921.3$ bits

Minimum description length (MDL)

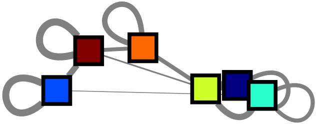
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$



$B = 6$, $\mathcal{S} \approx 1554.2$ bits



Model, $\mathcal{L} \approx 386.7$ bits

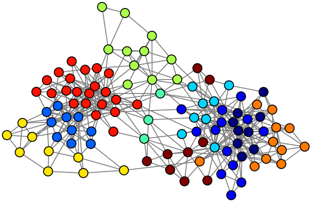
$\Sigma \approx 1940.9$ bits

Minimum description length (MDL)

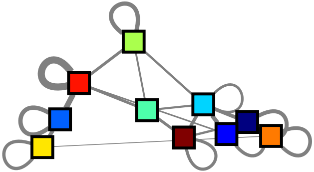
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, }\mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{e},\boldsymbol{b})}_{\text{model, }\mathcal{L}}$$



$B = 10$, $\mathcal{S} \approx 1451.0$ bits



Model, $\mathcal{L} \approx 590.8$ bits

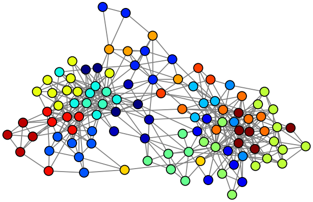
$\Sigma \approx 2041.8$ bits

Minimum description length (MDL)

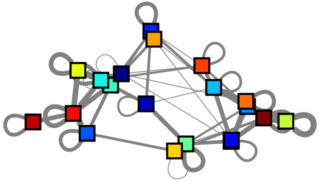
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}, \boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A}, \boldsymbol{e}, \boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A}, \boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{e}, \boldsymbol{b})}_{\text{data|model}, \, \mathcal{S}} \underbrace{- \log_2 P(\boldsymbol{e}, \boldsymbol{b})}_{\text{model}, \, \mathcal{L}}$$



$B = 20$, $\mathcal{S} \approx 1300.7$ bits



Model, $\mathcal{L} \approx 1037.8$ bits

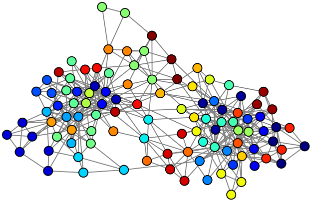
$\Sigma \approx 2338.6$ bits

Minimum description length (MDL)

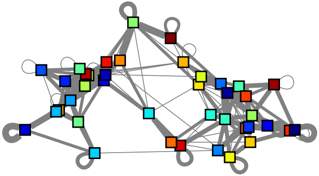
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$



$B = 40$, $\mathcal{S} \approx 1092.8$ bits



Model, $\mathcal{L} \approx 1730.3$ bits

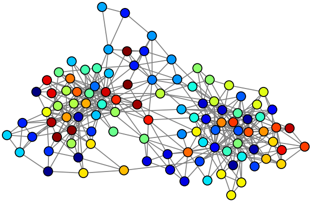
$\Sigma \approx 2823.1$ bits

Minimum description length (MDL)

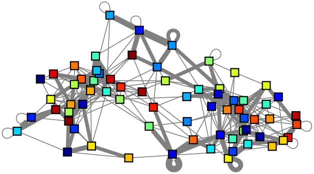
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

<u>Description length:</u>

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$



$B = 70$, $\mathcal{S} \approx 881.3$ bits



Model, $\mathcal{L} \approx 2427.3$ bits

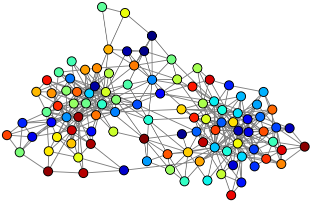
$\Sigma \approx 3308.6$ bits

Minimum description length (MDL)

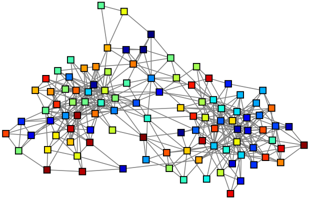
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}, \boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A}, \boldsymbol{e}, \boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A}, \boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{e}, \boldsymbol{b})}_{\text{data|model}, \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{e}, \boldsymbol{b})}_{\text{model}, \mathcal{L}}$$



$B = N$, $\mathcal{S} = 0$ bits



Model, $\mathcal{L} \approx 3714.9$ bits

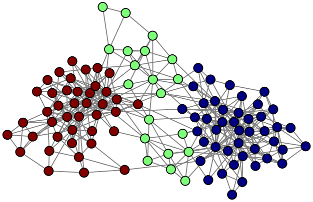
$\Sigma \approx 3714.9$ bits

Minimum description length (MDL)

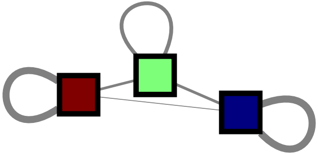
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$



$B = 3$, $\mathcal{S} \approx 1688.1$ bits



Model, $\mathcal{L} \approx 203.4$ bits

$\Sigma \approx 1891.5$ bits

Noninformative priors and underfitting

$$\Sigma \approx -(E-N)\log_2 B + \frac{B(B+1)}{2}\log_2 E$$

Noninformative priors and underfitting

$$\Sigma \approx -(E-N)\log_2 B + \frac{B(B+1)}{2}\log_2 E$$

$$B_{\max} \propto \sqrt{N}$$

“resolution limit”

Noninformative priors and underfitting

$$\Sigma \approx -(E-N)\log_2 B + \frac{B(B+1)}{2}\log_2 E$$

$$B_{\max} \propto \sqrt{N}$$

"resolution limit"

Q: How to do better *without* prior information?

Noninformative priors and underfitting

$$\Sigma \approx -(E-N)\log_2 B + \frac{B(B+1)}{2}\log_2 E$$

$$B_{\max} \propto \sqrt{N}$$

“resolution limit”

Q: How to do better *without* prior information?
A: Construct a model for the model!

$$P(\boldsymbol{A}|\boldsymbol{b}) = \int P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{b})P(\boldsymbol{\lambda}|\boldsymbol{\phi})P(\boldsymbol{\phi})\,\mathrm{d}\boldsymbol{\lambda}\,\mathrm{d}\boldsymbol{\phi}$$

Noninformative priors and underfitting

$$\Sigma \approx -(E-N)\log_2 B + \frac{B(B+1)}{2}\log_2 E$$

$$B_{\max} \propto \sqrt{N}$$

"resolution limit"

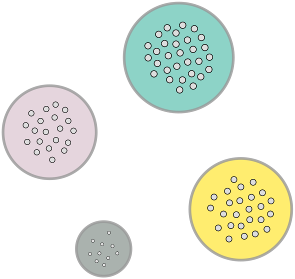
Q: How to do better *without* prior information?
A: Construct a model for the model!

$$P(\boldsymbol{A}|\boldsymbol{b}) = \int P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{b})P(\boldsymbol{\lambda}|\boldsymbol{\phi})P(\boldsymbol{\phi})\,\mathrm{d}\boldsymbol{\lambda}\,\mathrm{d}\boldsymbol{\phi}$$

Microcanonical:

$$P(\boldsymbol{A}|\boldsymbol{b}) = P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{e}|\boldsymbol{\psi})P(\boldsymbol{\psi})$$

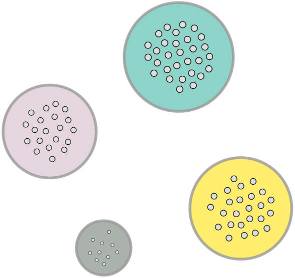
Prior for the partition



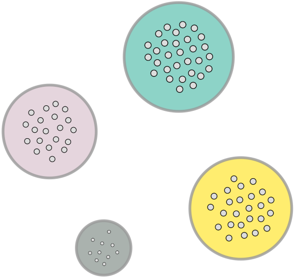
Prior for the partition

Option 1: Noninformative

$$P(\boldsymbol{b}) = \frac{1}{\sum_{B=1}^{N} \left\{ {N \atop B} \right\} B!}$$



Prior for the partition



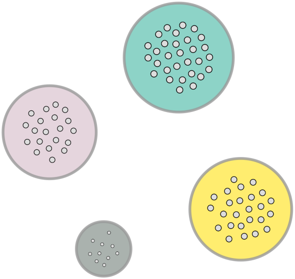
Option 1: Noninformative

$$P(\boldsymbol{b}) = \frac{1}{\sum_{B=1}^{N} \left\{ {N \atop B} \right\} B!}$$

Option 2: Slightly less noninformative

$$P(\boldsymbol{b}|B) = \frac{1}{\left\{ {N \atop B} \right\} B!} \quad P(B) = \frac{1}{N}$$

Prior for the partition



Option 1: Noninformative

$$P(\boldsymbol{b}) = \frac{1}{\sum_{B=1}^{N} \left\{ {N \atop B} \right\} B!}$$

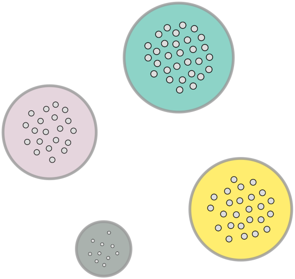
Option 2: Slightly less noninformative

$$P(\boldsymbol{b}|B) = \frac{1}{\left\{ {N \atop B} \right\} B!} \quad P(B) = \frac{1}{N}$$

Option 3: Conditioned on group sizes

$$\begin{aligned} P(\boldsymbol{b}|B) &= P(\boldsymbol{b}|\boldsymbol{n})P(\boldsymbol{n}) \\ &= \frac{N!}{\prod_r n_r!} \times \binom{N-1}{B-1}^{-1} \end{aligned}$$

Prior for the partition



Option 1: Noninformative

$$P(\boldsymbol{b}) = \frac{1}{\sum_{B=1}^{N} \left\{ {N \atop B} \right\} B!}$$

Option 2: Slightly less noninformative

$$P(\boldsymbol{b}|B) = \frac{1}{\left\{ {N \atop B} \right\} B!} \quad P(B) = \frac{1}{N}$$

Option 3: Conditioned on group sizes

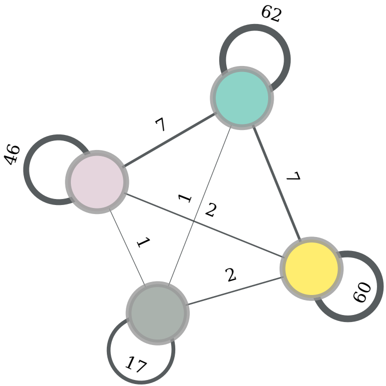
$$\begin{aligned} P(\boldsymbol{b}|B) &= P(\boldsymbol{b}|\boldsymbol{n})P(\boldsymbol{n}) \\ &= \frac{N!}{\prod_r n_r!} \times \binom{N-1}{B-1}^{-1} \end{aligned}$$

For $N \gg B$:

$$-\ln P(\boldsymbol{b}|B) \approx NH(\boldsymbol{n}) + O(B \ln N)$$

$$H(\boldsymbol{n}) = -\sum_r \frac{n_r}{N} \ln \frac{n_r}{N}$$

Prior for the edge counts

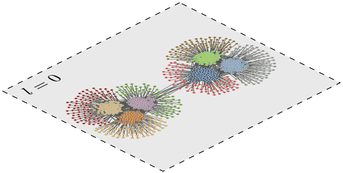


Option 1: Noninformative

$$P(\boldsymbol{e}) = \left(\!\!\left(\begin{matrix}\left(\!\binom{B}{2}\!\right) \\ E\end{matrix}\right)\!\!\right)^{-1}$$

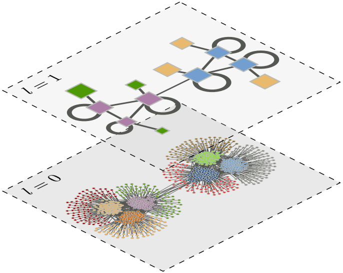
$$\left(\!\binom{n}{m}\!\right) = \binom{n+m-1}{m}$$

Nested SBM: Group hierarchies



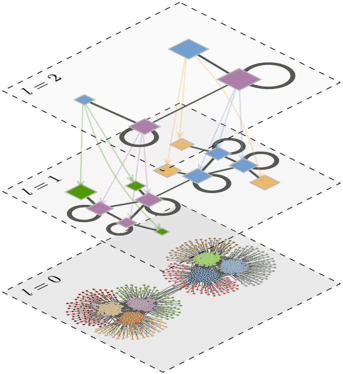
Condition $P(\boldsymbol{e})$ on... another SBM!

Nested SBM: Group hierarchies



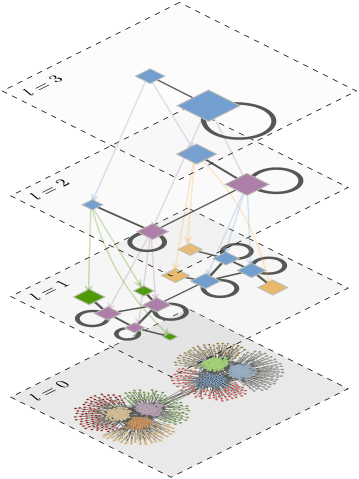
Condition $P(\boldsymbol{e})$ on... another SBM!

Nested SBM: Group hierarchies



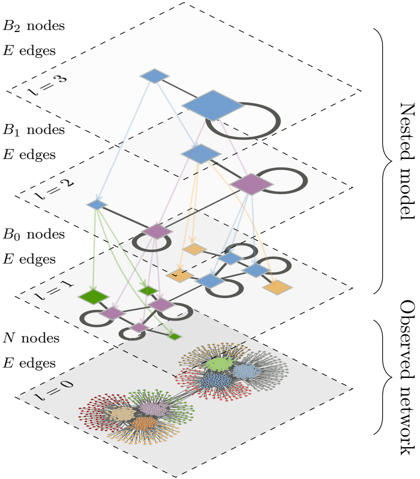
Condition $P(\boldsymbol{e})$ on... another SBM!

Nested SBM: Group hierarchies



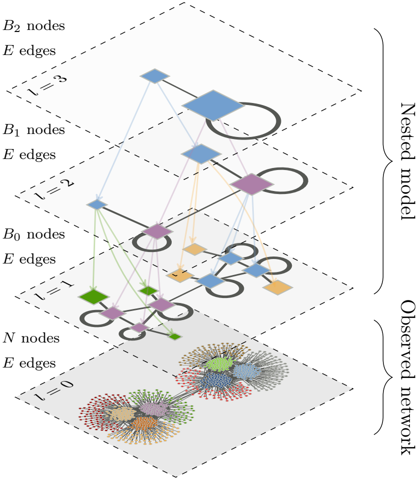
Condition $P(\boldsymbol{e})$ on... another SBM!

NESTED SBM: GROUP HIERARCHIES



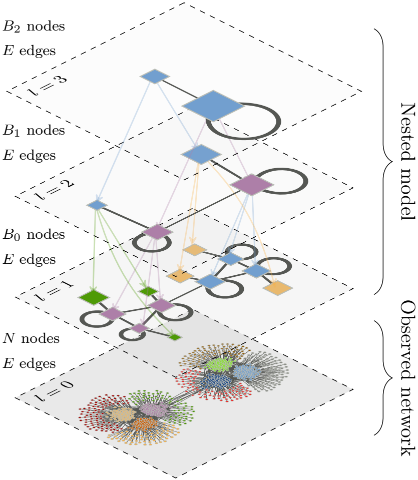
Condition $P(\boldsymbol{e})$ on... another SBM!

Nested SBM: Group hierarchies

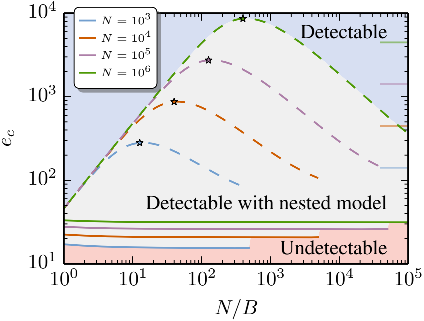


Condition $P(\boldsymbol{e})$ on... another SBM!

Nested SBM: Group hierarchies



Condition $P(\boldsymbol{e})$ on... another SBM!

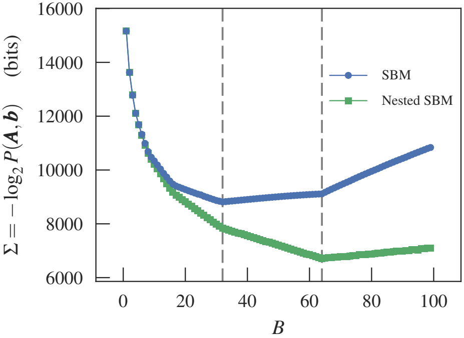


For planted partition:

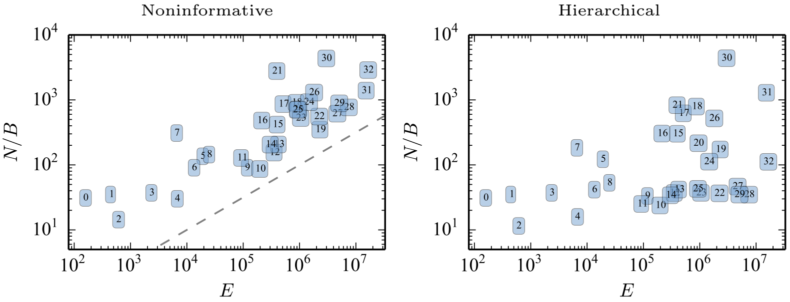
$$B_{\max} = O\left(\frac{N}{\log N}\right) \quad \gg \sqrt{N}$$

Nested SBM prevents underfitting

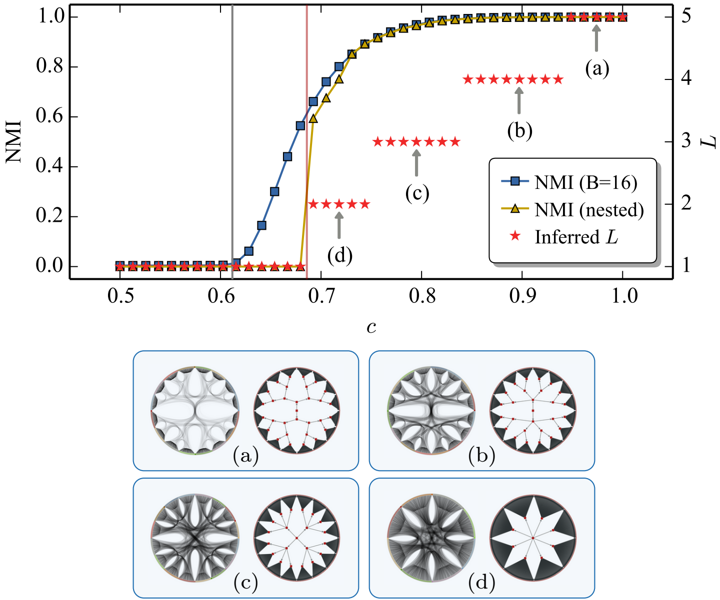
64 cliques of size 10



Resolution problem in the wild



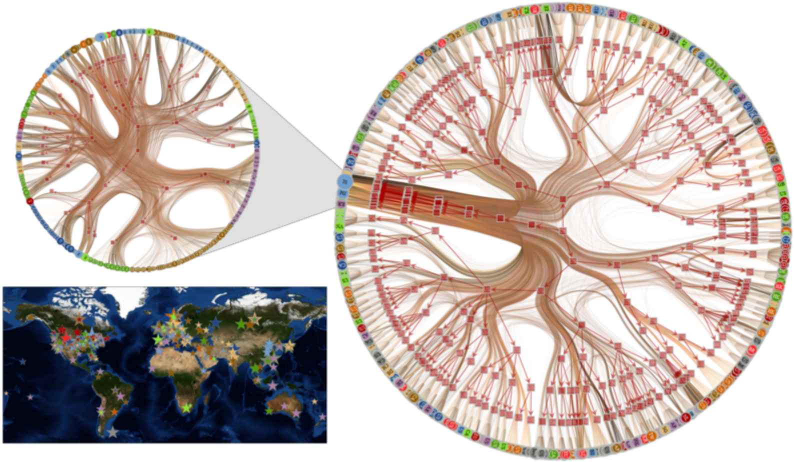
Hierarchical model: Built-in validation



Nested SBM: Modeling at multiple scales

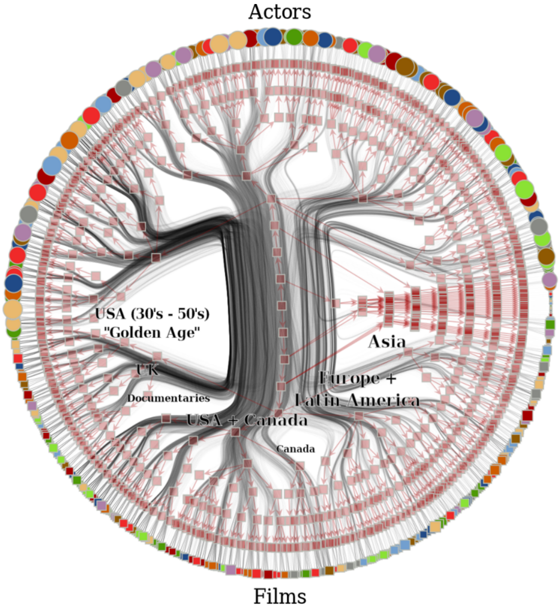
Internet (AS)

$(N = 52,104, E = 399,625)$



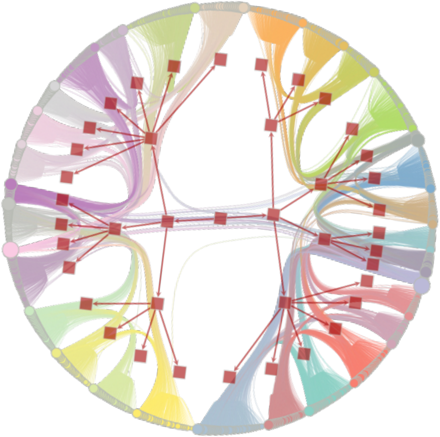
Nested SBM: Modeling at multiple scales

IMDB Film-Actor Network ($N = 372,447, E = 1,812,312, B = 717$)



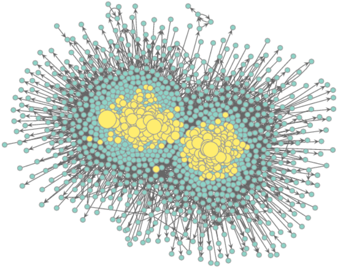
Nested SBM: Modeling at multiple scales

Human connectome



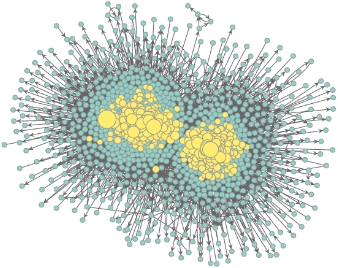
Degree-correction

Traditional SBM



Degree-correction

Traditional SBM



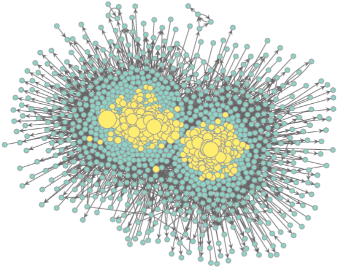
Degree-corrected SBM

$$P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{b}) = \prod_{i<j} \frac{(\theta_i\theta_j\lambda_{b_ib_j})^{A_{ij}} e^{-\theta_i\theta_j\lambda_{b_ib_j}}}{A_{ij}!} \times \prod_i \frac{(\theta_i^2\lambda_{b_ib_i}/2)^{A_{ii}/2} e^{-\theta_i^2\lambda_{b_ib_i}/2}}{(A_{ii}/2)!}$$

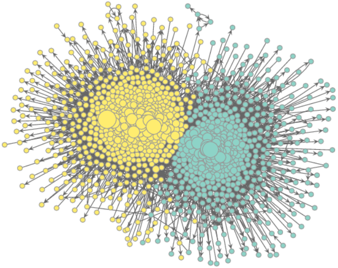
$\theta_i \to$ average degree of node i

(Karrer and Newman 2011)

Degree-correction



Traditional SBM



Degree-corrected SBM

$$P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{b}) = \prod_{i<j} \frac{(\theta_i\theta_j\lambda_{b_ib_j})^{A_{ij}} e^{-\theta_i\theta_j\lambda_{b_ib_j}}}{A_{ij}!} \times \prod_i \frac{(\theta_i^2\lambda_{b_ib_i}/2)^{A_{ii}/2} e^{-\theta_i^2\lambda_{b_ib_i}/2}}{(A_{ii}/2)!}$$

$\theta_i \rightarrow$ average degree of node i

(Karrer and Newman 2011)

Degree-corrected microcanonical SBM

Noninformative priors:

$$P(\boldsymbol{\lambda}|\boldsymbol{b}) = \prod_{r\le s} e^{-\lambda_{rs}/(1+\delta_{rs})\bar{\lambda}}/(1+\delta_{rs})\bar{\lambda}$$

$$P(\boldsymbol{\theta}|\boldsymbol{b}) = \prod_{r} (n_r - 1)!\delta(\textstyle\sum_i \theta_i \delta_{b_i,r} - 1)$$

Degree-corrected microcanonical SBM

Noninformative priors:

$$P(\boldsymbol{\lambda}|\boldsymbol{b}) = \prod_{r\le s} e^{-\lambda_{rs}/(1+\delta_{rs})\bar{\lambda}}/(1+\delta_{rs})\bar{\lambda}$$

$$P(\boldsymbol{\theta}|\boldsymbol{b}) = \prod_{r}(n_r-1)!\delta(\textstyle\sum_i \theta_i\delta_{b_i,r}-1)$$

Marginal likelihood:

$$\begin{aligned}P(\boldsymbol{A}|\boldsymbol{b}) &= \int P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{b})P(\boldsymbol{\lambda}|\boldsymbol{b})P(\boldsymbol{\theta}|\boldsymbol{b})\;\mathrm{d}\boldsymbol{\lambda}\mathrm{d}\boldsymbol{\theta}\\ &= \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}}\times\frac{\prod_{r<s}e_{rs}!\prod_r e_{rr}!!}{\prod_{i<j}A_{ij}!\prod_i A_{ii}!!}\times\prod_r\frac{(n_r-1)!}{(e_r+n_r-1)!}\times\prod_i k_i!\end{aligned}$$

Degree-corrected microcanonical SBM

Noninformative priors:

$$P(\boldsymbol{\lambda}|\boldsymbol{b}) = \prod_{r\le s} e^{-\lambda_{rs}/(1+\delta_{rs})\bar{\lambda}}/(1+\delta_{rs})\bar{\lambda}$$

$$P(\boldsymbol{\theta}|\boldsymbol{b}) = \prod_{r}(n_r-1)!\,\delta(\textstyle\sum_i \theta_i\delta_{b_i,r}-1)$$

Marginal likelihood:

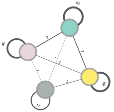
$$\begin{aligned}P(\boldsymbol{A}|\boldsymbol{b}) &= \int P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{b})P(\boldsymbol{\lambda}|\boldsymbol{b})P(\boldsymbol{\theta}|\boldsymbol{b})\;\mathrm{d}\boldsymbol{\lambda}\mathrm{d}\boldsymbol{\theta}\\ &= \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}}\times\frac{\prod_{r<s}e_{rs}!\prod_r e_{rr}!!}{\prod_{i<j}A_{ij}!\prod_i A_{ii}!!}\times\prod_r\frac{(n_r-1)!}{(e_r+n_r-1)!}\times\prod_i k_i!\end{aligned}$$

Microcanonical equivalence:

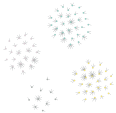
$$P(\boldsymbol{A}|\boldsymbol{b}) = P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{e})$$

$$P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \frac{\prod_{r<s}e_{rs}!\prod_r e_{rr}!!\prod_i k_i!}{\prod_{i<j}A_{ij}!\prod_i A_{ii}!!\prod_r e_r!!}$$

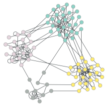
$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) = \prod_r \left(\!\!\binom{n_r}{e_r}\!\!\right)^{-1}$$



Edge counts $\boldsymbol{e}$.



Degrees, $\boldsymbol{k}$.

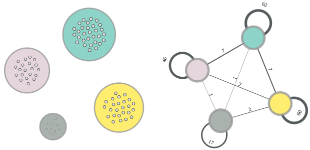


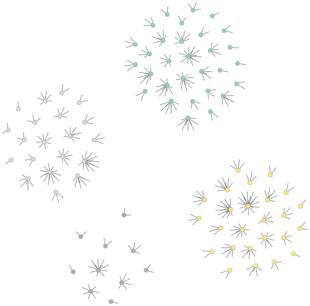
Network, $\boldsymbol{A}$.

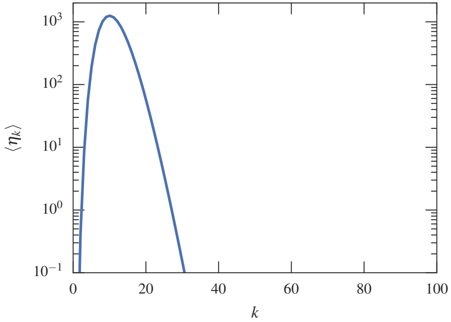
Prior for the degrees

Option 0: Random half-edges
(Non-degree-corrected)

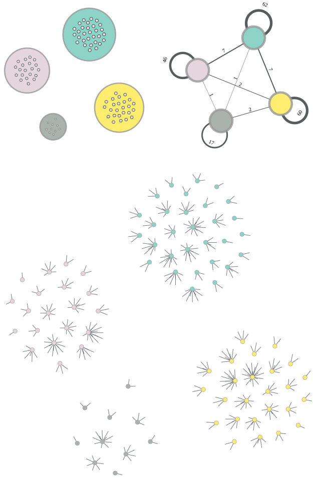
$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) = \prod_r \frac{e_r!}{n_r^{e_r}\prod_{i\in r} k_i!}$$







Prior for the degrees

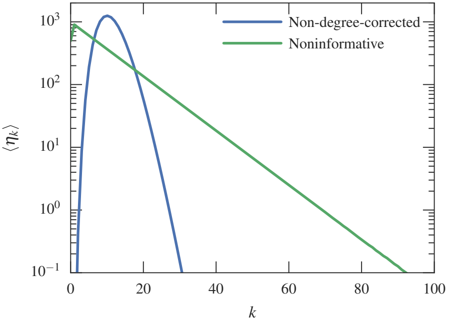


Option 0: Random half-edges
(Non-degree-corrected)

$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) = \prod_r \frac{e_r!}{n_r^{e_r}\prod_{i\in r} k_i!}$$

Option 1: Noninformative

$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) = \prod_r \left(\!\!\binom{n_r}{e_r}\!\!\right)^{-1}$$



Prior for the degrees

Option 2: Conditioned on degree distribution

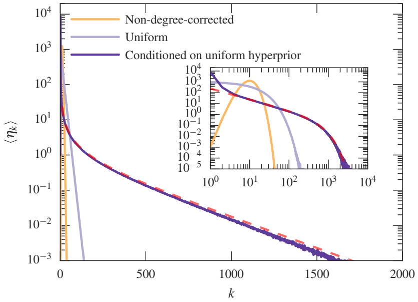
$$
\begin{aligned}
P(\boldsymbol{k}|\boldsymbol{b},\boldsymbol{e}) &= P(\boldsymbol{k}|\boldsymbol{\eta})P(\boldsymbol{\eta}) \\
&= \prod_r \frac{n_r!}{\prod_r \eta_k^r!} \times \prod_r q(e_r, n_r)^{-1}
\end{aligned}
$$

Prior for the degrees

Option 2: Conditioned on degree distribution

$$P(\boldsymbol{k}|\boldsymbol{b},\boldsymbol{e}) = P(\boldsymbol{k}|\boldsymbol{\eta})P(\boldsymbol{\eta})$$

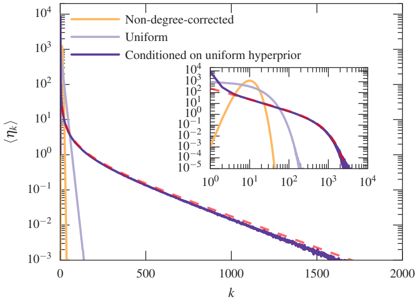
$$= \prod_r \frac{n_r!}{\prod_r \eta_k^r!} \times \prod_r q(e_r, n_r)^{-1}$$



Prior for the degrees

Option 2: Conditioned on degree distribution

$$P(\boldsymbol{k}|\boldsymbol{b},\boldsymbol{e}) = P(\boldsymbol{k}|\boldsymbol{\eta})P(\boldsymbol{\eta})$$
$$= \prod_r \frac{n_r!}{\prod_r \eta_k^r!} \times \prod_r q(e_r, n_r)^{-1}$$



<u>Bose-Einstein statistics</u>

N "bosons" (nodes)

$k_i \rightarrow$ "energy level" (degree)

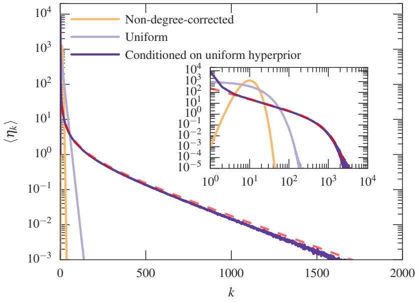
$$\langle \eta_k \rangle \approx \frac{1}{\exp\left(k\sqrt{\zeta(2)/E}\right) - 1}.$$

$$\langle \eta_k \rangle \approx \begin{cases} \sqrt{E/\zeta(2)}/k & \text{for } k \ll \sqrt{E}, \\ \exp(-k\sqrt{\zeta(2)/E}) & \text{for } k \gg \sqrt{E}. \end{cases}$$

Prior for the degrees

Option 2: Conditioned on degree distribution

$$P(\boldsymbol{k}|\boldsymbol{b},\boldsymbol{e}) = P(\boldsymbol{k}|\boldsymbol{\eta})P(\boldsymbol{\eta})$$
$$= \prod_r \frac{n_r!}{\prod_r \eta_k^r!} \times \prod_r q(e_r, n_r)^{-1}$$



Bose-Einstein statistics

N "bosons" (nodes)

$k_i \rightarrow$ "energy level" (degree)

$$\langle \eta_k \rangle \approx \frac{1}{\exp\left(k\sqrt{\zeta(2)/E}\right) - 1}.$$

$$\langle \eta_k \rangle \approx \begin{cases} \sqrt{E/\zeta(2)}/k & \text{for } k \ll \sqrt{E}, \\ \exp(-k\sqrt{\zeta(2)/E}) & \text{for } k \gg \sqrt{E}. \end{cases}$$

$$-\ln P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) \approx \sum_r n_r H(\boldsymbol{\eta}_r) + O(\sqrt{n_r})$$

$$H(\boldsymbol{\eta}_r) = -\sum_k (\eta_k^r/n_r)\ln(\eta_k^r/n_r)$$

Prior for the degrees: Integer partitions

$q(m, n) \to$ number of partitions of integer m into at most n parts

Exact computation

$$q(m, n) = q(m, n-1) + q(m-n, n)$$

Full table for $n \leq m \leq M$ in time $O(M^2)$.

Prior for the degrees: Integer partitions

$q(m,n) \rightarrow$ number of partitions of integer m into at most n parts

Exact computation

$$q(m,n) = q(m,n-1) + q(m-n,n)$$

Full table for $n \leq m \leq M$ in time $O(M^2)$.

Approximation

$$q(m,n) \approx \frac{f(u)}{m} \exp(\sqrt{m}\, g(u)),$$

$$u = n/\sqrt{m}$$

$$f(u) = \frac{v(u)}{2^{3/2}\pi u}\left[1 - (1+u^2/2)e^{-v(u)}\right]^{-1/2}$$

$$g(u) = \frac{2v(u)}{u} - u\ln(1 - e^{-v(u)})$$

$$v = u\sqrt{-v^2/2 - \mathrm{Li}_2(1 - e^v)}$$

(Szekeres, 1951)

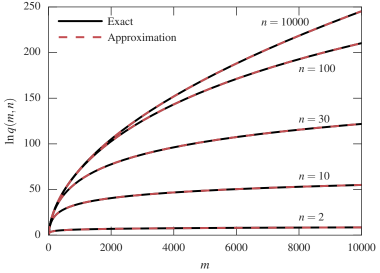
Prior for the degrees: Integer partitions

$q(m,n) \to$ number of partitions of integer m into at most n parts

Exact computation

$$q(m,n) = q(m,n-1) + q(m-n,n)$$

Full table for $n \le m \le M$ in time $O(M^2)$.



Approximation

$$q(m,n) \approx \frac{f(u)}{m} \exp(\sqrt{m}g(u)),$$

$$u = n/\sqrt{m}$$

$$f(u) = \frac{v(u)}{2^{3/2}\pi u}\left[1-(1+u^2/2)e^{-v(u)}\right]^{-1/2}$$

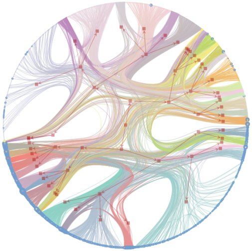
$$g(u) = \frac{2v(u)}{u} - u\ln(1-e^{-v(u)})$$

$$v = u\sqrt{-v^2/2 - \mathrm{Li}_2(1-e^v)}$$

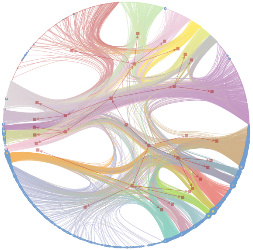
(Szekeres, 1951)

Example: Political blogs redux

$$N = 1,222, E = 19,027$$



(a) NDC-SBM, $B_1 = 42$, $\Sigma \approx 89938$ bits



(b) DC-SBM, uniform prior, $B_1 = 23$, $\Sigma \approx 87162$ bits

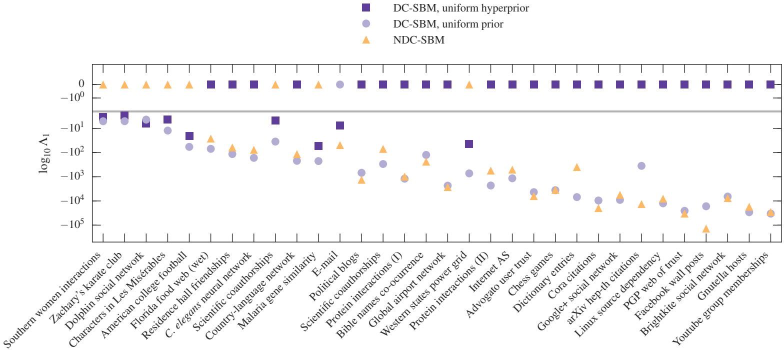
(c) DC-SBM, uniform hyperprior, $B_1 = 20$, $\Sigma \approx 84890$ bits

Model selection

$$\Lambda = \frac{P(\boldsymbol{A}, \{\boldsymbol{b}_l\} | \mathcal{C}_1) P(\mathcal{C}_1)}{P(\boldsymbol{A}, \{\boldsymbol{b}'_l\} | \mathcal{C}_2) P(\mathcal{C}_2)} = 2^{-\Delta\Sigma}$$

Model selection

$$\Lambda = \frac{P(\boldsymbol{A}, \{\boldsymbol{b}_l\}|\mathcal{C}_1)P(\mathcal{C}_1)}{P(\boldsymbol{A}, \{\boldsymbol{b}'_l\}|\mathcal{C}_2)P(\mathcal{C}_2)} = 2^{-\Delta\Sigma}$$



Inferring group assignments: Sampling vs. Maximization

$\hat{\boldsymbol{b}} \to$ estimator, $\boldsymbol{b}^* \to$ true partition

Maximum *a posteriori* (MAP)

Indicator loss function:

$$\Delta(\hat{\boldsymbol{b}}, \boldsymbol{b}^*) = \prod_i \delta_{\hat{b}_i, b_i^*}$$

Maximization over the posterior

$$\sum_{\boldsymbol{b}} \Delta(\hat{\boldsymbol{b}}, \boldsymbol{b}) P(\boldsymbol{b}|\boldsymbol{A})$$

yields the MAP estimator

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmax}}\, P(\boldsymbol{b}|\boldsymbol{A}).$$

(Identical to MDL.)

Node marginals

Overlap loss function:

$$d(\hat{\boldsymbol{b}}, \boldsymbol{b}^*) = \frac{1}{N} \sum_i \delta_{\hat{b}_i, b_i^*}$$

Maximization over the posterior

$$\sum_{\boldsymbol{b}} d(\hat{\boldsymbol{b}}, \boldsymbol{b}) P(\boldsymbol{b}|\boldsymbol{A})$$

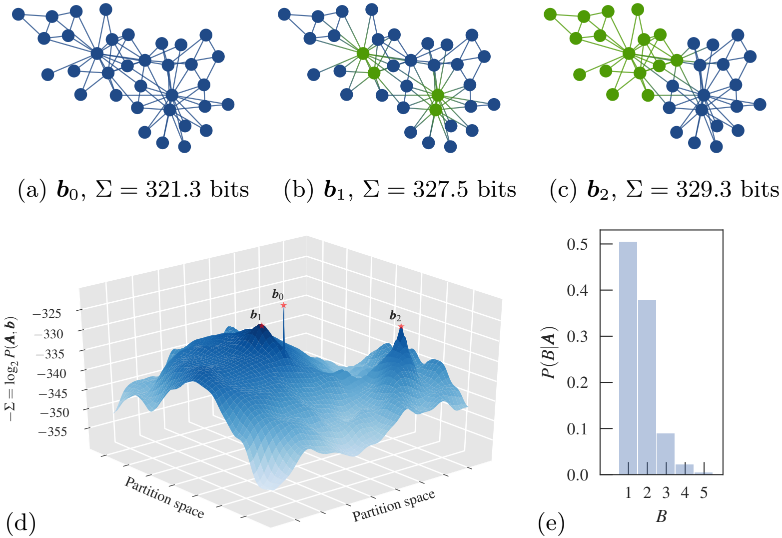
yields the marginal estimator

$$\hat{b}_i = \underset{r}{\operatorname{argmax}}\, \pi_i(r),$$

$$\pi_i(r) = \sum_{\boldsymbol{b} \setminus b_i} P(b_i = r, \boldsymbol{b} \setminus b_i | \boldsymbol{A}).$$

Sampling vs. maximization

Zachary's karate club



Sampling vs. Maximization

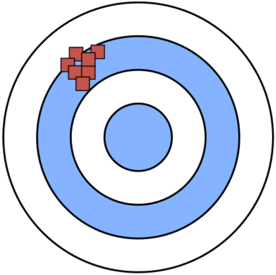
Bias-variance trade-off

<u>Maximization (MDL)</u>

$$\{\hat{\boldsymbol{b}}_l\} = \underset{\{\boldsymbol{b}_l\}}{\operatorname{argmax}}\, P(\{\boldsymbol{b}_l\}|\boldsymbol{A})$$

Finds the best partition, and number of groups $\{B_l\}$.

Lower variance, higher bias



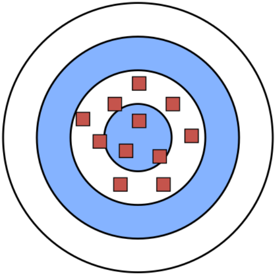
*Tendency to **underfit**.*

<u>Sampling</u>

$$\{\boldsymbol{b}_l\} \sim P(\{\boldsymbol{b}_l\}|\boldsymbol{A})$$

Finds the best posterior ensemble; marginal posterior $P(B_l|\boldsymbol{A})$.

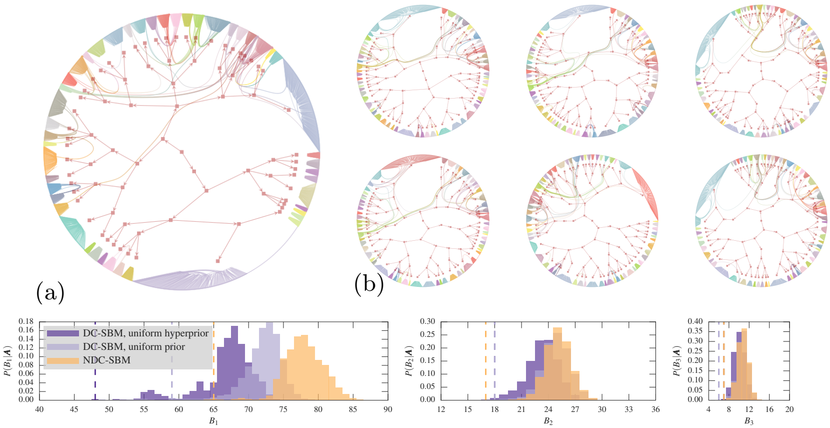
Higher variance, lower bias



*Tendency to **overfit**.*

Sampling vs. maximization

Collaborations between scientists ($N = 1,589$, $E = 2,742$)



Inference algorithm: Metropolis-Hastings

- Move proposal, $b_i = r \to b_i = s$
- Accept with probability

$$a = \min\left(1, \frac{P(\boldsymbol{b}'|\boldsymbol{A})P(\boldsymbol{b} \to \boldsymbol{b}')}{P(\boldsymbol{b}|\boldsymbol{A})P(\boldsymbol{b}' \to \boldsymbol{b})}\right).$$

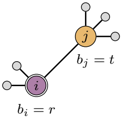
Move proposal of node i requires $O(k_i)$ operations. A whole MCMC sweep can be performed in $O(E)$ time, **independent of the number of groups,** B.

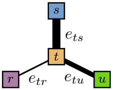
In contrast to:
1. EM + BP with Bernoulli SBM: $O(EB^2)$ (Semiparametric) [Decelle et al, 2011]
2. Variational Bayes with (overlapping) Bernoulli SBM: $O(EB)$ (Semiparametric) [Gopalan and Blei, 2011]
3. Bernoulli SBM with noninformative priors: $O(EB^2)$ (Greedy) [Côme and Latouche, 2015]
4. Poisson SBM with noninformative priors: $O(EB^2)$ (heat bath) [Newman and Reinert, 2016]

Efficient inference: other improvements

Smart move proposals

- Choose a random vertex i (happens to belong to group r).
- Move it to a random group $s \in [1, B]$, chosen with a probability $p(r \to s|t)$ proportional to $e_{ts} + \epsilon$, where t is the group membership of a randomly chosen neighbour of i.

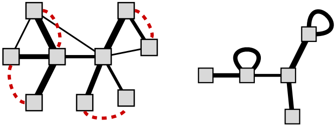




Fast mixing times.

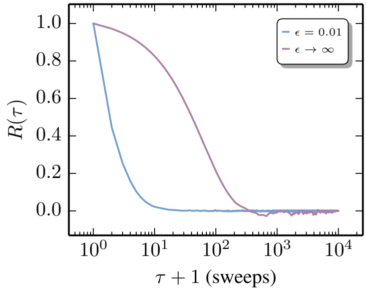
Agglomerative initialization

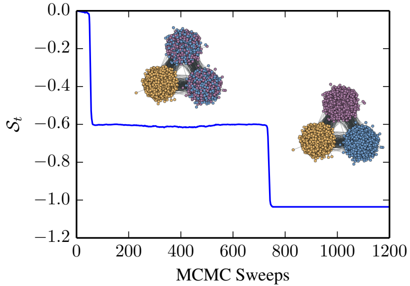
- Start with $B = N$.
- Progressively merge groups.



Avoids metastable states.

Efficient inference: other improvements





Greedy heuristic: $P(\boldsymbol{b}|\boldsymbol{A})^{\beta}$, $\beta \to \infty$