

Statistical inference of network structure

Part 1

Tiago P. Peixoto

University of Bath

Berlin, August 2017

Introductory text:

“Bayesian stochastic blockmodeling”, arXiv: 1705.10225

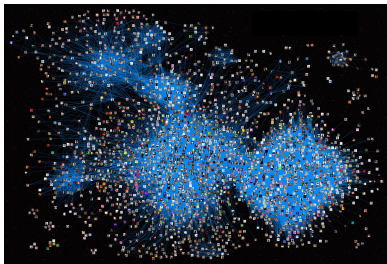
For code, see:

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

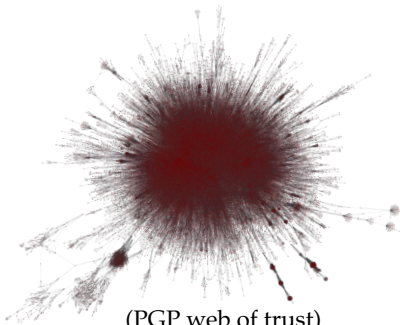
(See also HOWTO at: `https://graph-tool.skewed.de/static/doc/demos/inference/inference.html`)

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

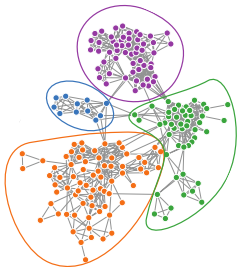
HOW TO CHARACTERIZE LARGE-SCALE STRUCTURES?



(Flickr social network)



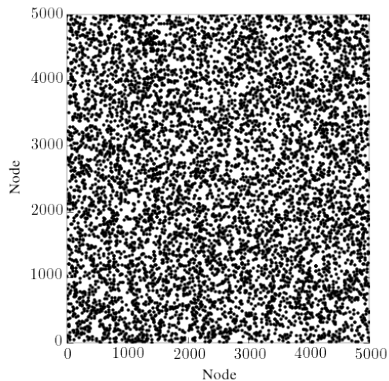
(PGP web of trust)



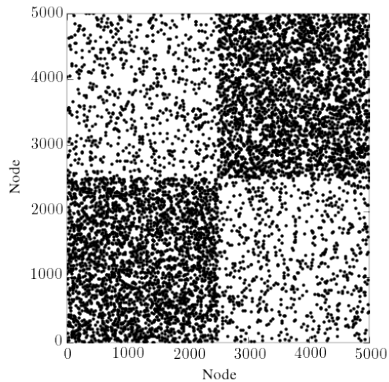
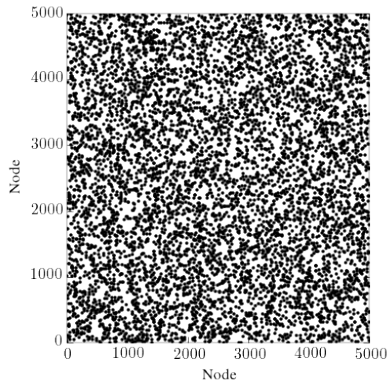
Find a meaningful network partition that provides:

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

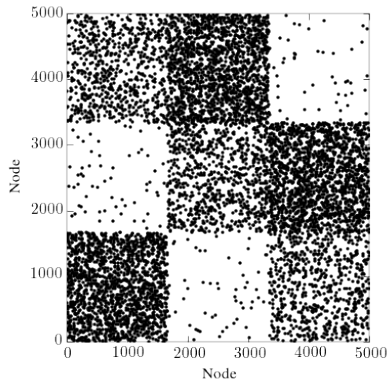
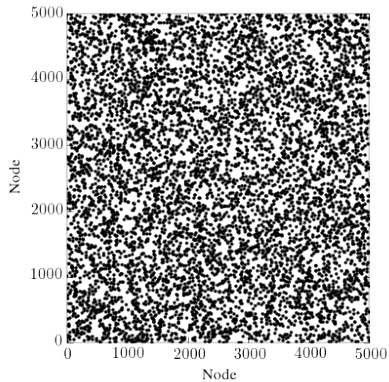
STRUCTURE VS. RANDOMNESS



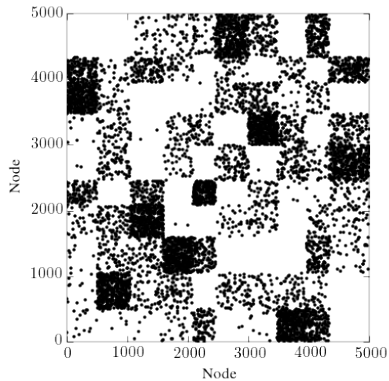
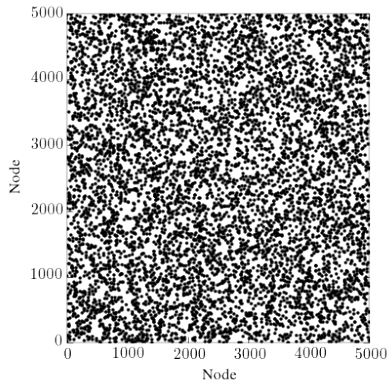
STRUCTURE VS. RANDOMNESS



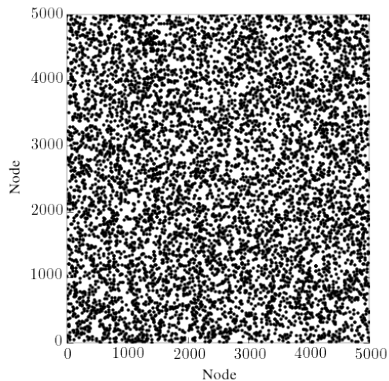
STRUCTURE VS. RANDOMNESS



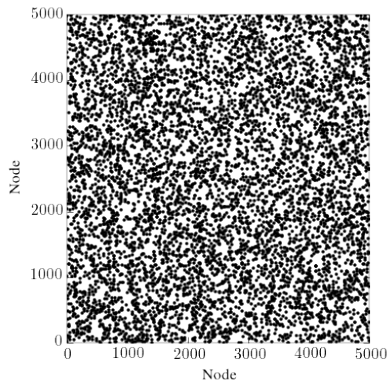
STRUCTURE VS. RANDOMNESS



STRUCTURE VS. RANDOMNESS

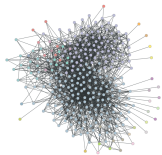


STRUCTURE VS. RANDOMNESS

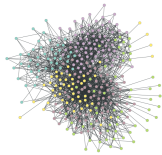


We need *well defined*, **principled** methodology, grounded on robust concepts of **statistical inference**.

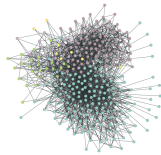
DIFFERENT METHODS, DIFFERENT RESULTS...



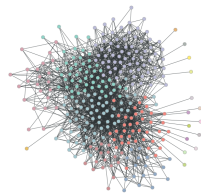
Betweenness



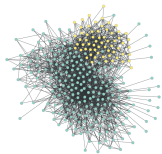
Modularity matrix



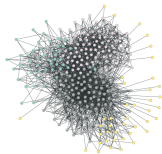
Infomap



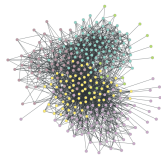
Walk Trap



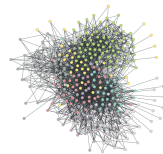
Label propagation



Modularity



Modularity
(Blondel)



Spin glass



Infomap
(overlapping)



Clique percolation

A PRINCIPLED ALTERNATIVE: STATISTICAL INFERENCE

What we have: $\mathbf{A} = \{A_{ij}\} \rightarrow$ Network

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

Generative model

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

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

Bayesian posterior

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

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

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

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

WHY STATISTICAL INFERENCE?

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

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

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, $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \dots$

Posterior odds ratio:

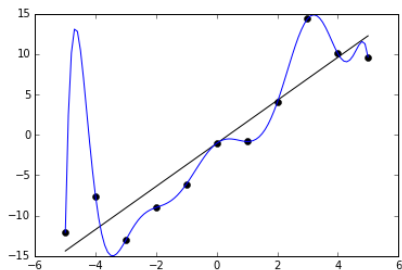
$$\Lambda = \frac{P(\mathbf{b}_1, \mathcal{C}_1|\mathbf{A})}{P(\mathbf{b}_2, \mathcal{C}_2|\mathbf{A})} = \frac{P(\mathbf{A}|\mathbf{b}_1, \mathcal{C}_1)P(\mathbf{b}_1)P(\mathcal{C}_1)}{P(\mathbf{A}|\mathbf{b}_2, \mathcal{C}_2)P(\mathbf{b}_2)P(\mathcal{C}_2)}$$

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

OVERFITTING, REGULARIZATION AND OCCAM'S RAZOR

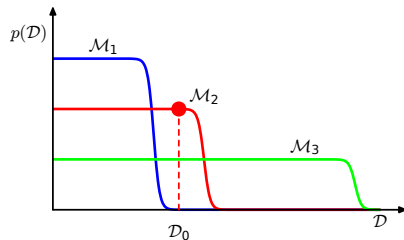
$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

$\mathcal{D} \rightarrow$ Data, $\theta \rightarrow$ Parameters



Evidence:

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



TWO APPROACHES: PARAMETRIC VS. NON-PARAMETRIC

Parametric

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

$\boldsymbol{\theta} \rightarrow$ 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

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

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

$\boldsymbol{\theta} \rightarrow$ 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.

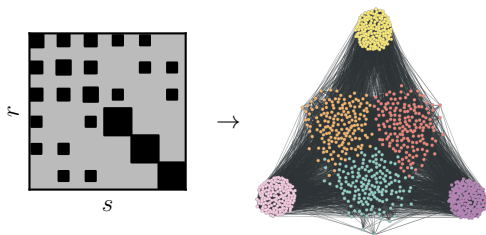
(Our focus.)

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 .

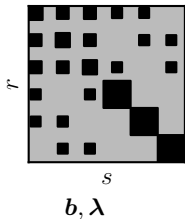


-
- ▶ Not restricted to assortative structures (“communities”).
 - ▶ Easily generalizable (edge direction, overlapping groups, etc.)

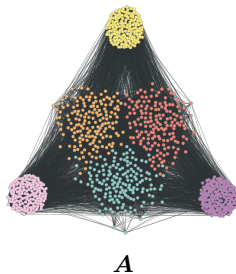
BAYESIAN STATISTICAL INFERENCE

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

$\mathbf{A} \rightarrow$ Observed network
 $\boldsymbol{\lambda}, \mathbf{b} \rightarrow$ Model parameters



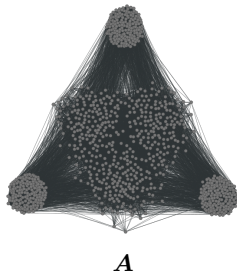
$\xrightarrow{P(\mathbf{A}|\boldsymbol{\lambda}, \mathbf{b})}$



BAYESIAN STATISTICAL INFERENCE

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

$\mathbf{A} \rightarrow$ Observed network
 $\boldsymbol{\lambda}, \mathbf{b} \rightarrow$ Model parameters

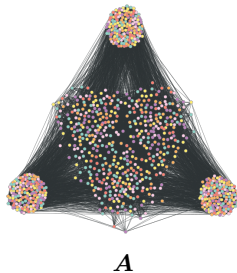


BAYESIAN STATISTICAL INFERENCE

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

$\mathbf{A} \rightarrow$ Observed network

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

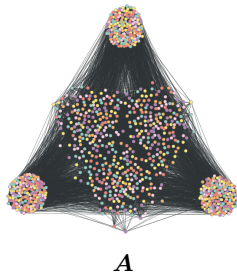


BAYESIAN STATISTICAL INFERENCE

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

$\mathbf{A} \rightarrow$ Observed network
 $\boldsymbol{\lambda}, \mathbf{b} \rightarrow$ Model parameters

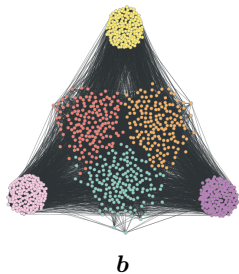
\leftarrow
 $P(\mathbf{b}|\mathbf{A})$



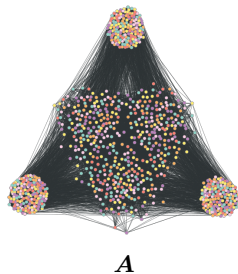
BAYESIAN STATISTICAL INFERENCE

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

$\mathbf{A} \rightarrow$ Observed network
 $\boldsymbol{\lambda}, \mathbf{b} \rightarrow$ Model parameters



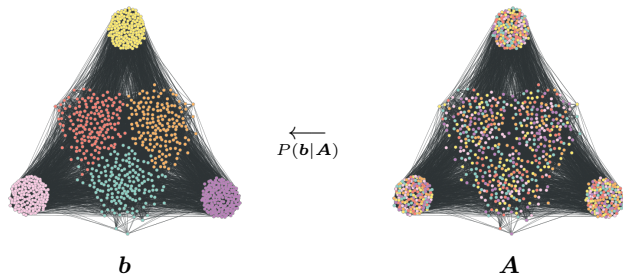
\leftarrow
 $P(\mathbf{b}|\mathbf{A})$



BAYESIAN STATISTICAL INFERENCE

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

$\mathbf{A} \rightarrow$ Observed network
 $\boldsymbol{\lambda}, \mathbf{b} \rightarrow$ Model parameters



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

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

THE POISSON SBM

Model likelihood:

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

THE POISSON SBM

Model likelihood:

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

Noninformative prior for λ :

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

THE POISSON SBM

Model likelihood:

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

Noninformative prior for λ :

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

Marginal likelihood:

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

THE POISSON SBM

Model likelihood:

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

Noninformative prior for λ :

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

Marginal likelihood:

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

Prior for \mathbf{b} (tentative):

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

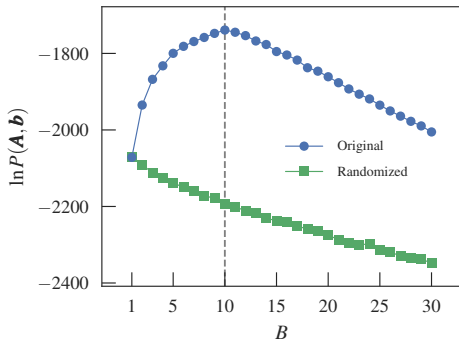
$$P(B) = 1/N$$

Posterior distribution:

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

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

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

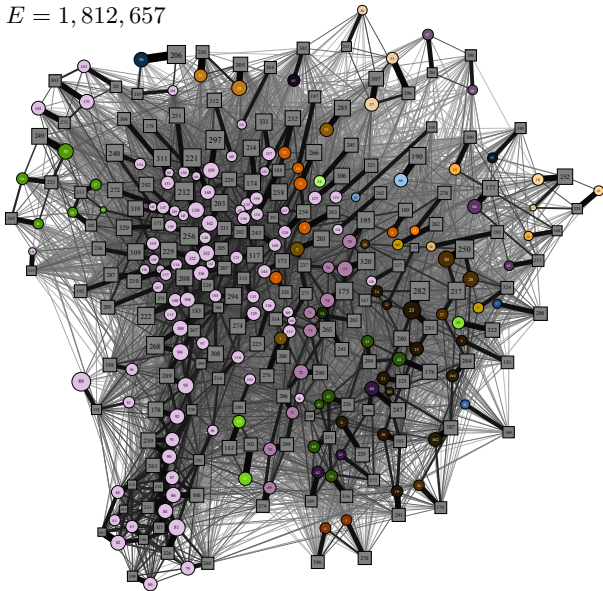
MDL selects: $B = 332$

EXAMPLE: THE INTERNET MOVIE DATABASE (IMDB)

Bipartite network of actors and films.

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

MDL selects: $B = 332$

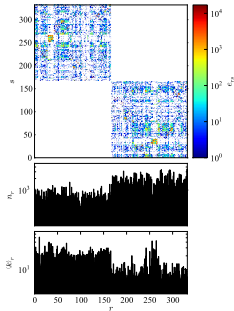


EXAMPLE: THE INTERNET MOVIE DATABASE (IMDB)

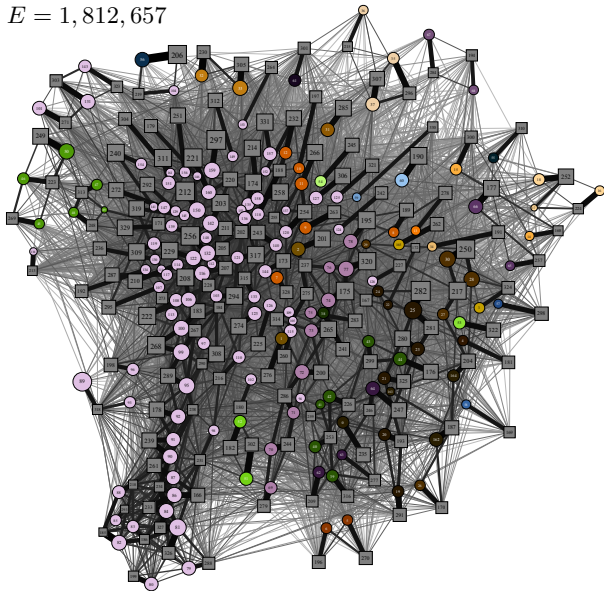
Bipartite network of actors and films.

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

MDL selects: $B = 332$



Bipartiteness is fully uncovered!

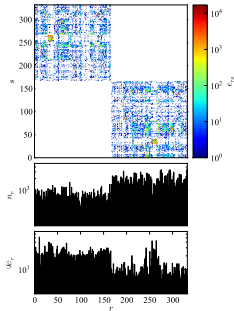


EXAMPLE: THE INTERNET MOVIE DATABASE (IMDB)

Bipartite network of actors and films.

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

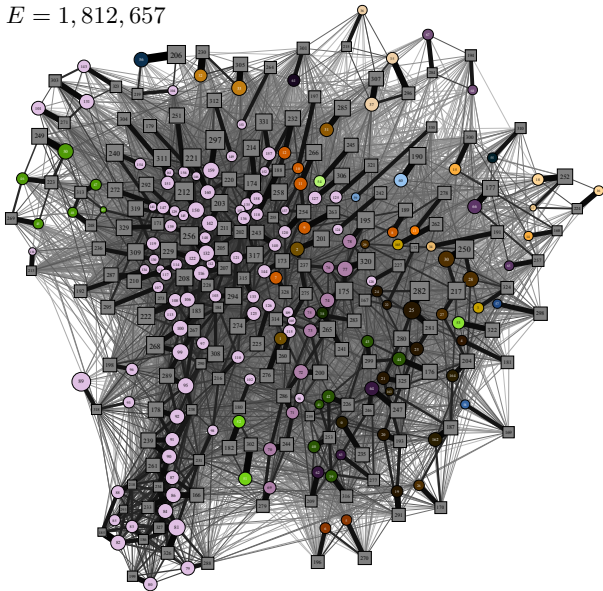
MDL selects: $B = 332$



Bipartiteness is fully uncovered!

Meaningful features:

- ▶ Temporal
- ▶ Spatial (Country)
- ▶ Type/Genre

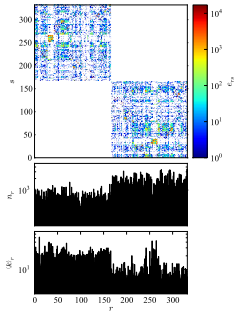


EXAMPLE: THE INTERNET MOVIE DATABASE (IMDB)

Bipartite network of actors and films.

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

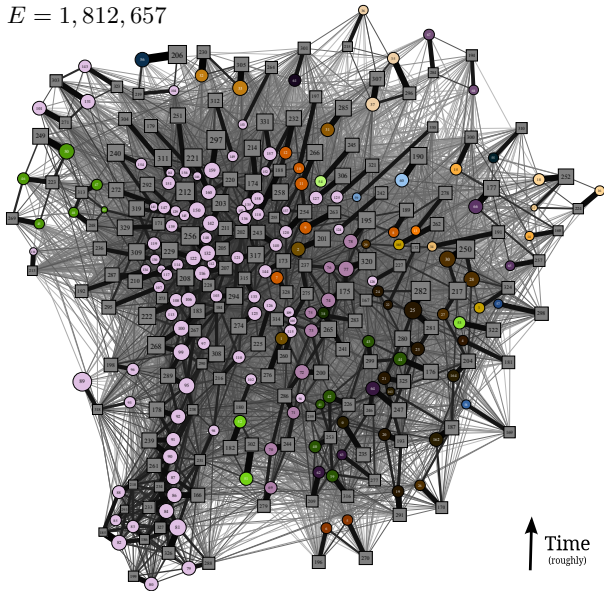
MDL selects: $B = 332$



Bipartiteness is fully uncovered!

Meaningful features:

- ▶ Temporal
- ▶ Spatial (Country)
- ▶ Type/Genre

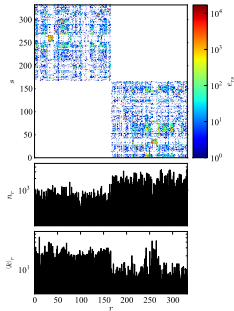


EXAMPLE: THE INTERNET MOVIE DATABASE (IMDB)

Bipartite network of actors and films.

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

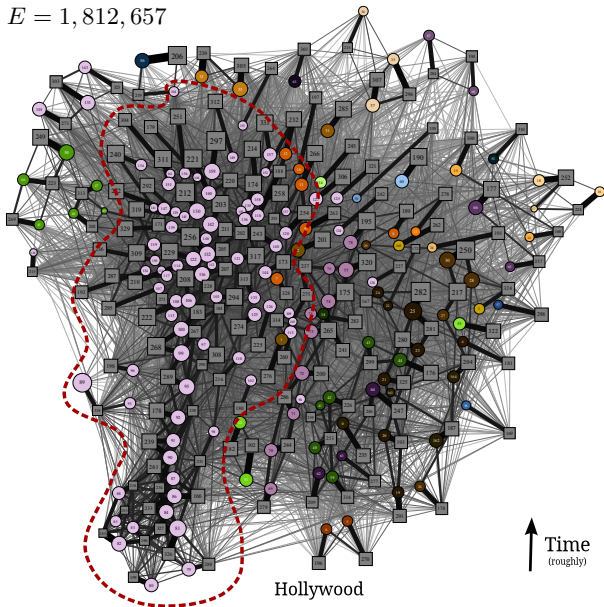
MDL selects: $B = 332$



Bipartiteness is fully uncovered!

Meaningful features:

- ▶ Temporal
- ▶ Spatial (Country)
- ▶ Type/Genre

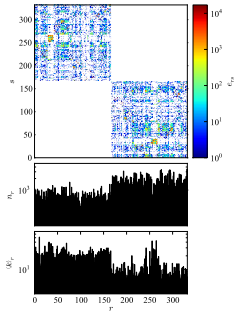


EXAMPLE: THE INTERNET MOVIE DATABASE (IMDB)

Bipartite network of actors and films.

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

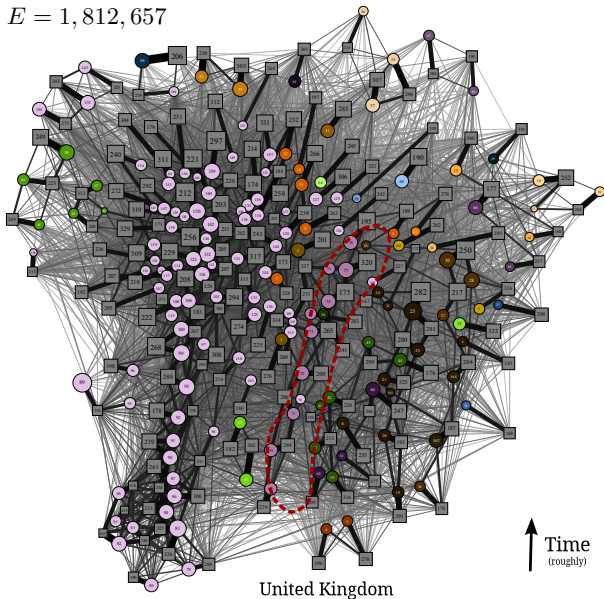
MDL selects: $B = 332$



Bipartiteness is fully uncovered!

Meaningful features:

- ▶ Temporal
- ▶ Spatial (Country)
- ▶ Type/Genre

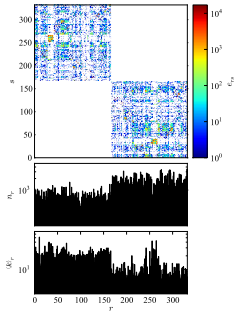


EXAMPLE: THE INTERNET MOVIE DATABASE (IMDB)

Bipartite network of actors and films.

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

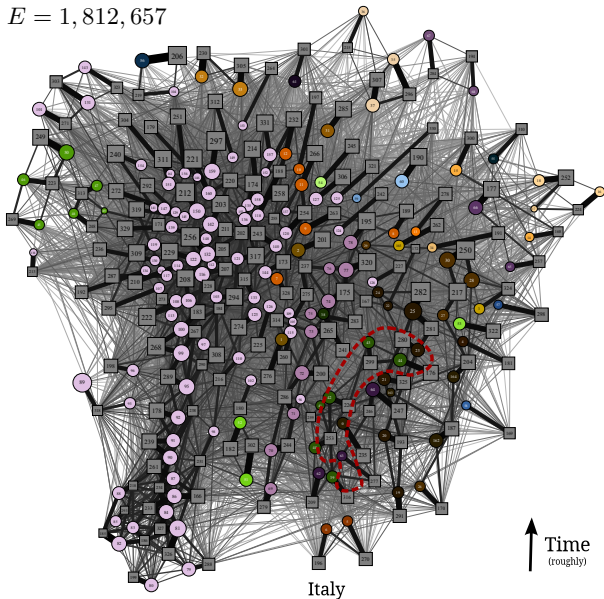
MDL selects: $B = 332$



Bipartiteness is fully uncovered!

Meaningful features:

- ▶ Temporal
- ▶ Spatial (Country)
- ▶ Type/Genre

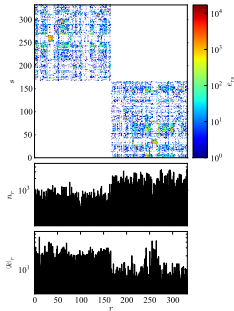


EXAMPLE: THE INTERNET MOVIE DATABASE (IMDB)

Bipartite network of actors and films.

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

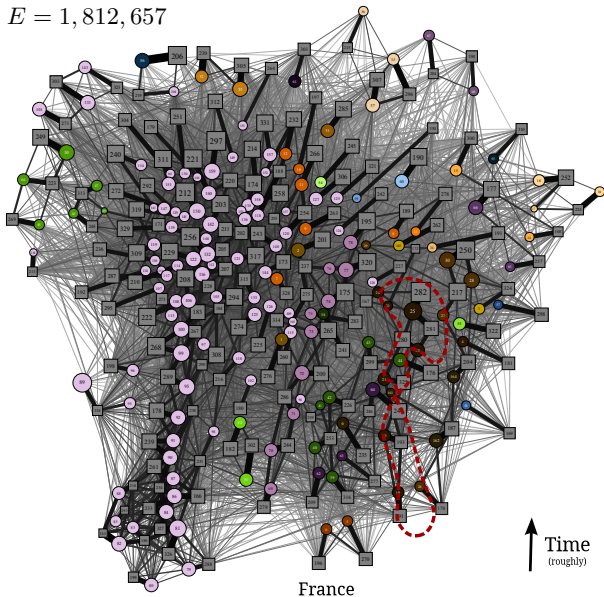
MDL selects: $B = 332$



Bipartiteness is fully uncovered!

Meaningful features:

- ▶ Temporal
- ▶ Spatial (Country)
- ▶ Type/Genre

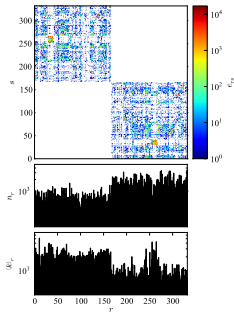


EXAMPLE: THE INTERNET MOVIE DATABASE (IMDB)

Bipartite network of actors and films.

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

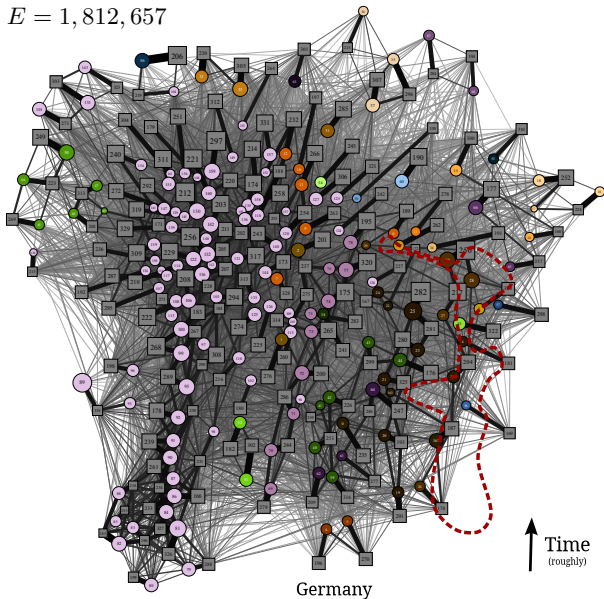
MDL selects: $B = 332$



Bipartiteness is fully uncovered!

Meaningful features:

- ▶ Temporal
- ▶ Spatial (Country)
- ▶ Type/Genre



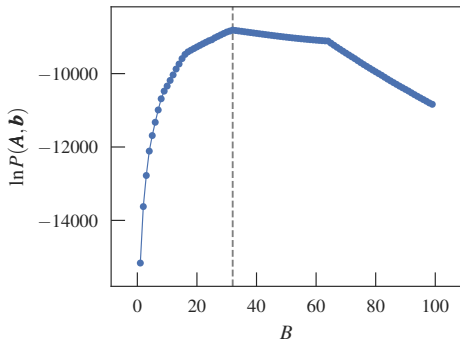
UNDERFITTING: THE “RESOLUTION LIMIT” PROBLEM

64 cliques of size 10

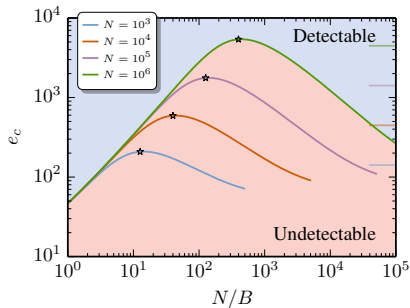
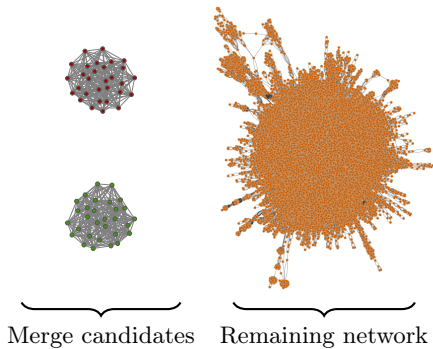


UNDERFITTING: THE “RESOLUTION LIMIT” PROBLEM

64 cliques of size 10



UNDERFITTING: THE “RESOLUTION LIMIT” PROBLEM



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

Why?

MICROCANONICAL EQUIVALENCE

Marginal likelihood:

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

MICROCANONICAL EQUIVALENCE

Marginal likelihood:

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

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

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

MICROCANONICAL EQUIVALENCE

Marginal likelihood:

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

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

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

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

MICROCANONICAL EQUIVALENCE

Marginal likelihood:

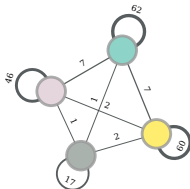
$$\begin{aligned} P(\mathbf{A}|\mathbf{b}) &= \int P(\mathbf{A}|\boldsymbol{\lambda}, \mathbf{b}) P(\boldsymbol{\lambda}|\mathbf{b}) d\boldsymbol{\lambda} \\ &= \frac{\bar{\lambda}^E}{(\bar{\lambda} + 1)^{E+B(B+1)/2}} \times \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!!}{\prod_r n_r^{e_r} \prod_{i<j} A_{ij}!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

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

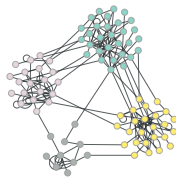
$$P(\mathbf{A}|\mathbf{b}) = P(\mathbf{A}|\mathbf{e}, \mathbf{b}) P(\mathbf{e}|\mathbf{b})$$

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

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



Edge counts, \mathbf{e}



Network, \mathbf{A}

MINIMUM DESCRIPTION LENGTH (MDL)

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

Description length:

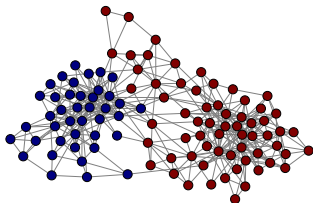
$$\Sigma = -\log_2 P(\mathbf{A}, \mathbf{b}) = \underbrace{-\log_2 P(\mathbf{A}|\mathbf{e}, \mathbf{b})}_{\text{data|model, } \mathcal{S}} - \underbrace{\log_2 P(\mathbf{e}, \mathbf{b})}_{\text{model, } \mathcal{L}}$$

MINIMUM DESCRIPTION LENGTH (MDL)

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

Description length:

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



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



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

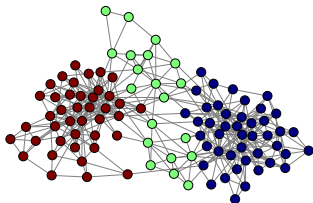
$\Sigma \approx 1927.9$ bits

MINIMUM DESCRIPTION LENGTH (MDL)

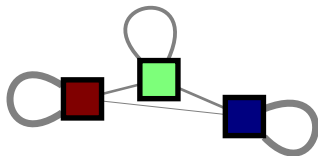
$$P(\mathbf{b}|\mathbf{A}) = \frac{P(\mathbf{A}, \mathbf{b})}{P(\mathbf{A})} = \frac{P(\mathbf{A}, \mathbf{e}, \mathbf{b})}{P(\mathbf{A})} = \frac{2^{-\Sigma}}{P(\mathbf{A})}$$

Description length:

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



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



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

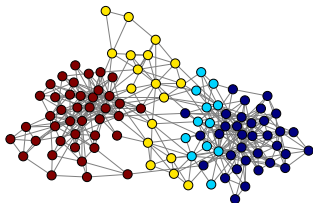
$\Sigma \approx 1891.5$ bits

MINIMUM DESCRIPTION LENGTH (MDL)

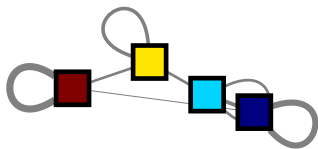
$$P(\mathbf{b}|\mathbf{A}) = \frac{P(\mathbf{A}, \mathbf{b})}{P(\mathbf{A})} = \frac{P(\mathbf{A}, \mathbf{e}, \mathbf{b})}{P(\mathbf{A})} = \frac{2^{-\Sigma}}{P(\mathbf{A})}$$

Description length:

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



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



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

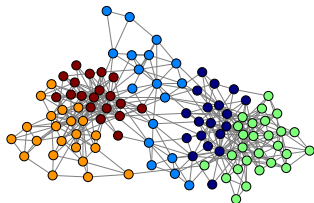
$\Sigma \approx 1911.5$ bits

MINIMUM DESCRIPTION LENGTH (MDL)

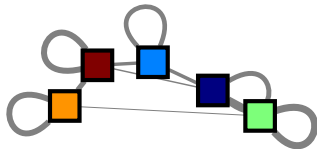
$$P(\mathbf{b}|\mathbf{A}) = \frac{P(\mathbf{A}, \mathbf{b})}{P(\mathbf{A})} = \frac{P(\mathbf{A}, \mathbf{e}, \mathbf{b})}{P(\mathbf{A})} = \frac{2^{-\Sigma}}{P(\mathbf{A})}$$

Description length:

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



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



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

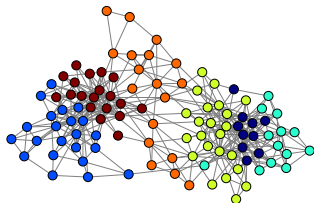
$\Sigma \approx 1921.3$ bits

MINIMUM DESCRIPTION LENGTH (MDL)

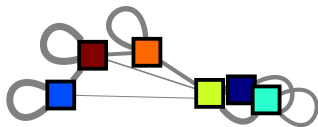
$$P(\mathbf{b}|\mathbf{A}) = \frac{P(\mathbf{A}, \mathbf{b})}{P(\mathbf{A})} = \frac{P(\mathbf{A}, \mathbf{e}, \mathbf{b})}{P(\mathbf{A})} = \frac{2^{-\Sigma}}{P(\mathbf{A})}$$

Description length:

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



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



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

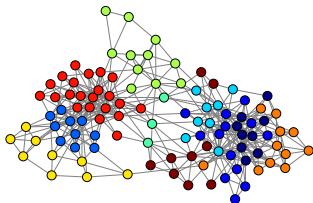
$\Sigma \approx 1940.9$ bits

MINIMUM DESCRIPTION LENGTH (MDL)

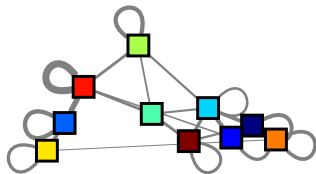
$$P(\mathbf{b}|\mathbf{A}) = \frac{P(\mathbf{A}, \mathbf{b})}{P(\mathbf{A})} = \frac{P(\mathbf{A}, \mathbf{e}, \mathbf{b})}{P(\mathbf{A})} = \frac{2^{-\Sigma}}{P(\mathbf{A})}$$

Description length:

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



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



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

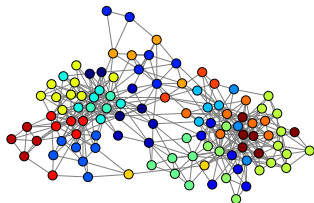
$\Sigma \approx 2041.8$ bits

MINIMUM DESCRIPTION LENGTH (MDL)

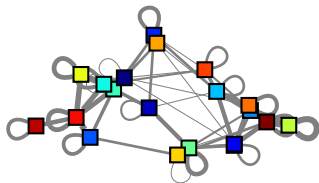
$$P(\mathbf{b}|\mathbf{A}) = \frac{P(\mathbf{A}, \mathbf{b})}{P(\mathbf{A})} = \frac{P(\mathbf{A}, \mathbf{e}, \mathbf{b})}{P(\mathbf{A})} = \frac{2^{-\Sigma}}{P(\mathbf{A})}$$

Description length:

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



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



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

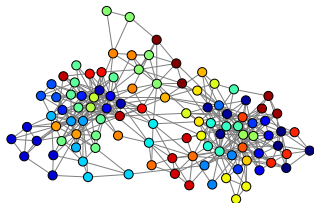
$\Sigma \approx 2338.6$ bits

MINIMUM DESCRIPTION LENGTH (MDL)

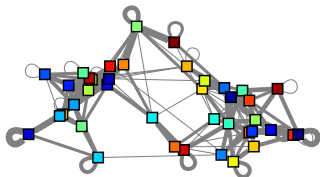
$$P(\mathbf{b}|\mathbf{A}) = \frac{P(\mathbf{A}, \mathbf{b})}{P(\mathbf{A})} = \frac{P(\mathbf{A}, \mathbf{e}, \mathbf{b})}{P(\mathbf{A})} = \frac{2^{-\Sigma}}{P(\mathbf{A})}$$

Description length:

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



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



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

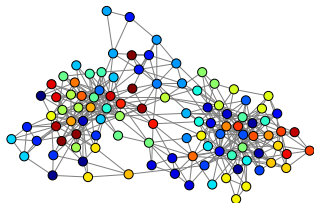
$\Sigma \approx 2823.1$ bits

MINIMUM DESCRIPTION LENGTH (MDL)

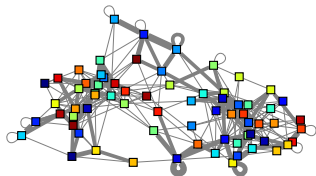
$$P(\mathbf{b}|\mathbf{A}) = \frac{P(\mathbf{A}, \mathbf{b})}{P(\mathbf{A})} = \frac{P(\mathbf{A}, \mathbf{e}, \mathbf{b})}{P(\mathbf{A})} = \frac{2^{-\Sigma}}{P(\mathbf{A})}$$

Description length:

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



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



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

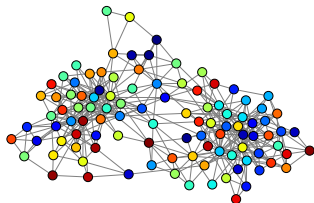
$\Sigma \approx 3308.6$ bits

MINIMUM DESCRIPTION LENGTH (MDL)

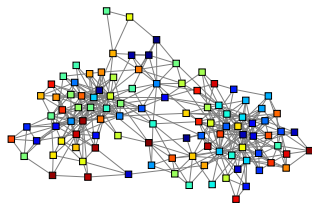
$$P(\mathbf{b}|\mathbf{A}) = \frac{P(\mathbf{A}, \mathbf{b})}{P(\mathbf{A})} = \frac{P(\mathbf{A}, \mathbf{e}, \mathbf{b})}{P(\mathbf{A})} = \frac{2^{-\Sigma}}{P(\mathbf{A})}$$

Description length:

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



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



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

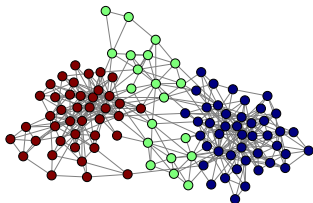
$\Sigma \approx 3714.9$ bits

MINIMUM DESCRIPTION LENGTH (MDL)

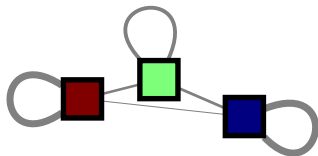
$$P(\mathbf{b}|\mathbf{A}) = \frac{P(\mathbf{A}, \mathbf{b})}{P(\mathbf{A})} = \frac{P(\mathbf{A}, \mathbf{e}, \mathbf{b})}{P(\mathbf{A})} = \frac{2^{-\Sigma}}{P(\mathbf{A})}$$

Description length:

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



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

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

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?

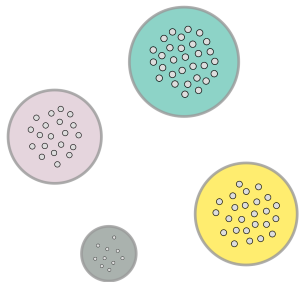
A: Construct a model for the model!

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

Microcanonical:

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

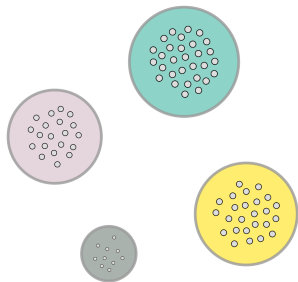
PRIOR FOR THE PARTITION



PRIOR FOR THE PARTITION

Option 1: Noninformative

$$P(\mathbf{b}) = \frac{1}{\sum_{B=1}^N \left\{ \begin{matrix} N \\ B \end{matrix} \right\} B!}$$



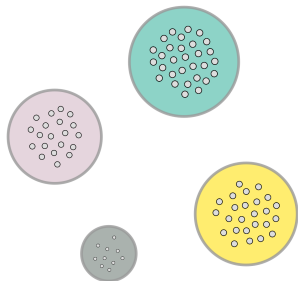
PRIOR FOR THE PARTITION

Option 1: Noninformative

$$P(\mathbf{b}) = \frac{1}{\sum_{B=1}^N \left\{ \begin{matrix} N \\ B \end{matrix} \right\} B!}$$

Option 2: Slightly less noninformative

$$P(\mathbf{b}|B) = \frac{1}{\left\{ \begin{matrix} N \\ B \end{matrix} \right\} B!} \quad P(B) = \frac{1}{N}$$



PRIOR FOR THE PARTITION

Option 1: Noninformative

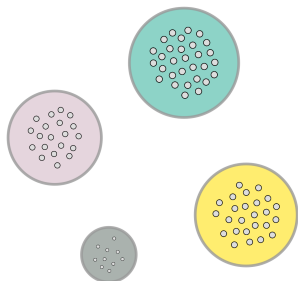
$$P(\mathbf{b}) = \frac{1}{\sum_{B=1}^N \left\{ \begin{matrix} N \\ B \end{matrix} \right\} B!}$$

Option 2: Slightly less noninformative

$$P(\mathbf{b}|B) = \frac{1}{\left\{ \begin{matrix} N \\ B \end{matrix} \right\} B!} \quad P(B) = \frac{1}{N}$$

Option 3: Conditioned on group sizes

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



PRIOR FOR THE PARTITION

Option 1: Noninformative

$$P(\mathbf{b}) = \frac{1}{\sum_{B=1}^N \left\{ \begin{matrix} N \\ B \end{matrix} \right\} B!}$$

Option 2: Slightly less noninformative

$$P(\mathbf{b}|B) = \frac{1}{\left\{ \begin{matrix} N \\ B \end{matrix} \right\} B!} \quad P(B) = \frac{1}{N}$$

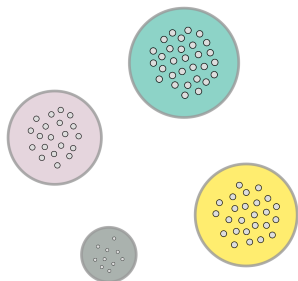
Option 3: Conditioned on group sizes

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

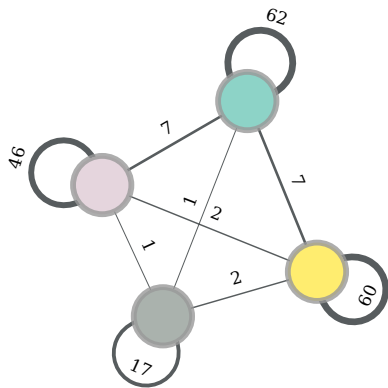
For $N \gg B$:

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

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



PRIOR FOR THE EDGE COUNTS



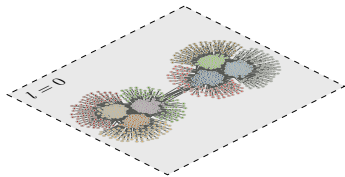
Option 1: Noninformative

$$P(\mathbf{e}) = \left(\left(\begin{pmatrix} B \\ 2 \\ E \end{pmatrix} \right) \right)^{-1}$$

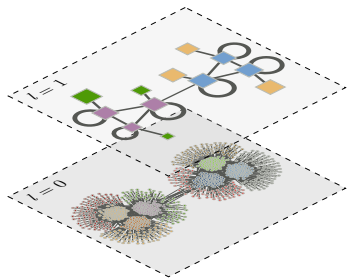
$$\binom{n}{m} = \binom{n+m-1}{m}$$

NESTED SBM: GROUP HIERARCHIES

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

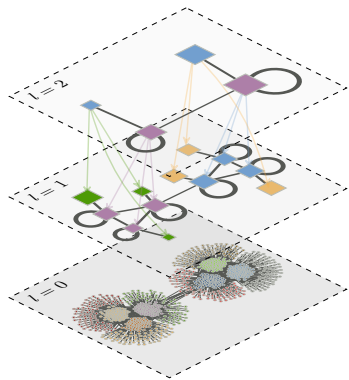


NESTED SBM: GROUP HIERARCHIES



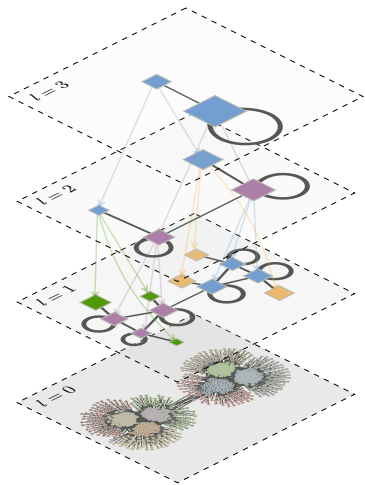
Condition $P(e)$ on... another SBM!

NESTED SBM: GROUP HIERARCHIES



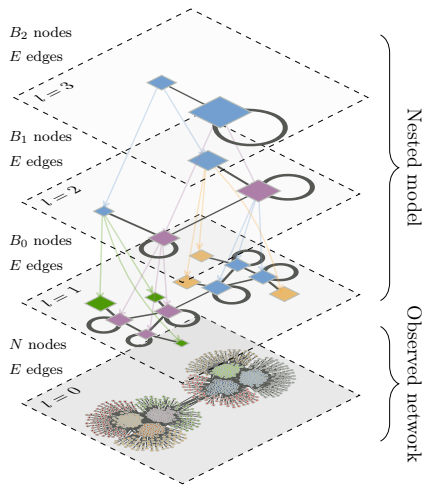
Condition $P(e)$ on... another SBM!

NESTED SBM: GROUP HIERARCHIES



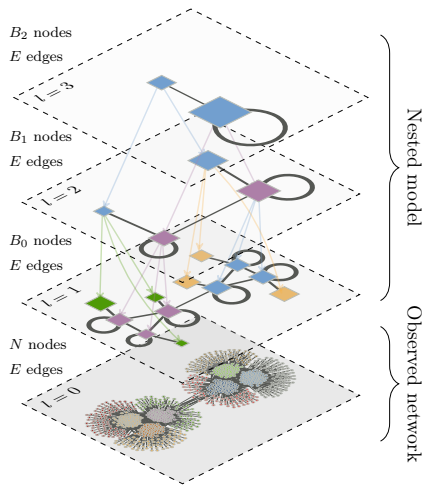
Condition $P(e)$ on... another SBM!

NESTED SBM: GROUP HIERARCHIES



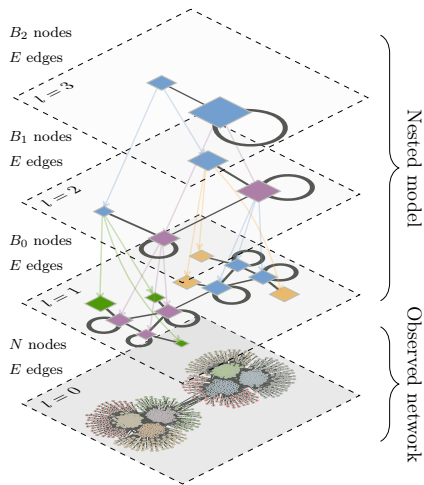
Condition $P(e)$ on... another SBM!

NESTED SBM: GROUP HIERARCHIES

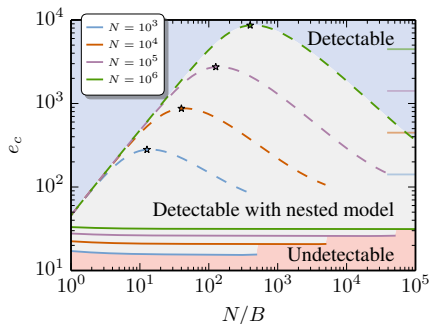


Condition $P(e)$ on... another SBM!

NESTED SBM: GROUP HIERARCHIES



Condition $P(e)$ on... another SBM!

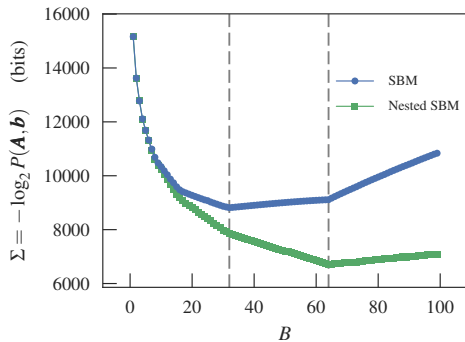


For planted partition:

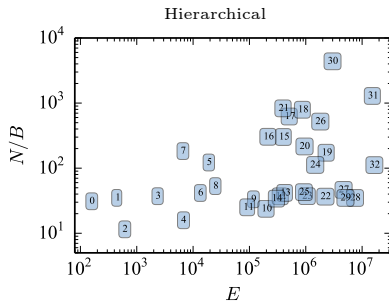
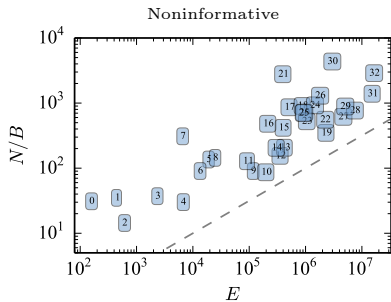
$$B_{\max} = O\left(\frac{N}{\log N}\right) \gg \sqrt{N}$$

NESTED SBM PREVENTS UNDERFITTING

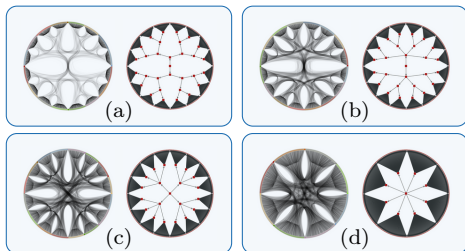
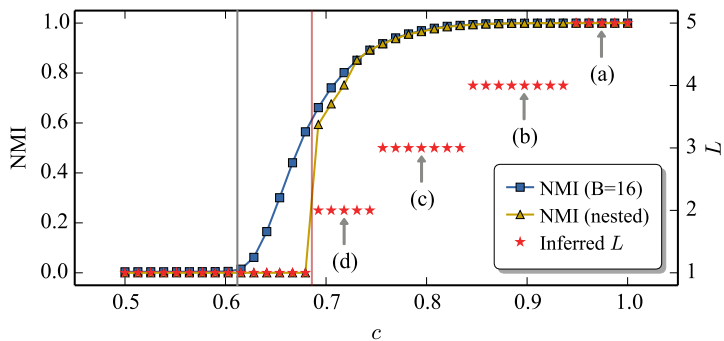
64 cliques of size 10



RESOLUTION PROBLEM IN THE WILD

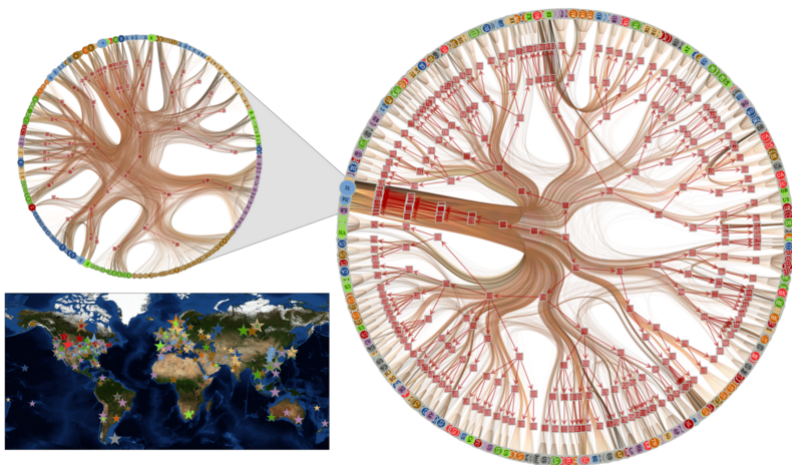


HIERARCHICAL MODEL: BUILT-IN VALIDATION



NESTED SBM: MODELING AT MULTIPLE SCALES

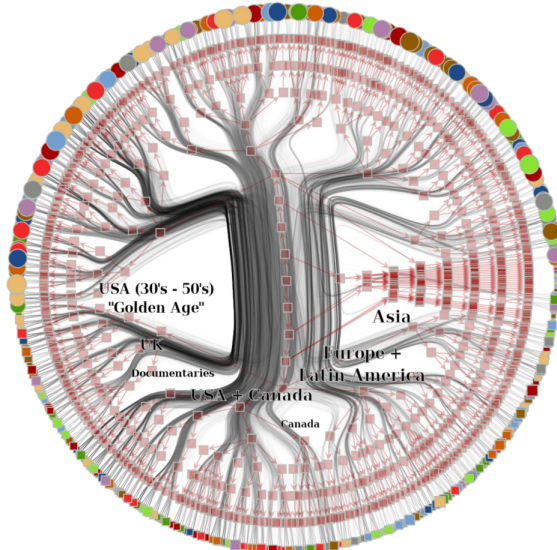
Internet (AS)
($N = 52,104$, $E = 399,625$)



NESTED SBM: MODELING AT MULTIPLE SCALES

IMDB FILM-ACTOR NETWORK ($N = 372,447$, $E = 1,812,312$, $B = 717$)

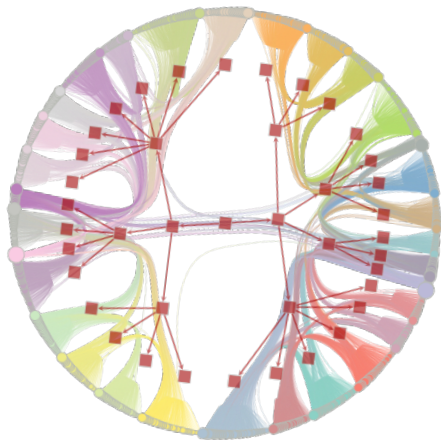
Actors



Films

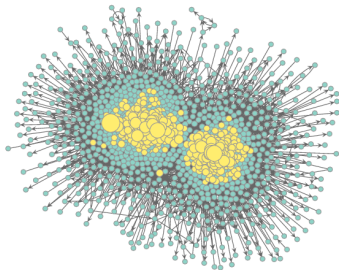
NESTED SBM: MODELING AT MULTIPLE SCALES

HUMAN CONNECTOME



DEGREE-CORRECTION

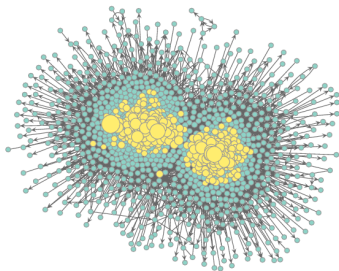
Traditional SBM



DEGREE-CORRECTION

Traditional SBM

Degree-corrected SBM



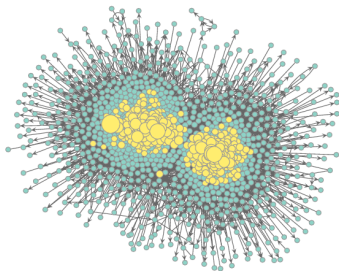
$$P(\mathbf{A}|\boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{b}) = \prod_{i < j} \frac{(\theta_i \theta_j \lambda_{b_i b_j})^{A_{ij}} e^{-\theta_i \theta_j \lambda_{b_i b_j}}}{A_{ij}!} \times \prod_i \frac{(\theta_i^2 \lambda_{b_i b_i} / 2)^{A_{ii}/2} e^{-\theta_i^2 \lambda_{b_i b_i} / 2}}{(A_{ii}/2)!}$$

$\theta_i \rightarrow$ average degree of node i

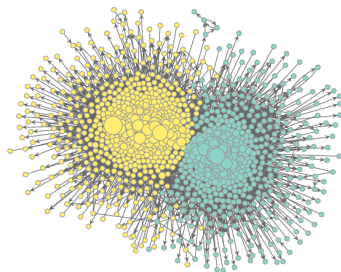
(Karrer and Newman 2011)

DEGREE-CORRECTION

Traditional SBM



Degree-corrected SBM



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

$\theta_i \rightarrow$ average degree of node i

(Karrer and Newman 2011)

DEGREE-CORRECTED MICROCANONICAL SBM

Noninformative priors:

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

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

DEGREE-CORRECTED MICROCANONICAL SBM

Noninformative priors:

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

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

Marginal likelihood:

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

DEGREE-CORRECTED MICROCANONICAL SBM

Noninformative priors:

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

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

Marginal likelihood:

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

Microcanonical equivalence:

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

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

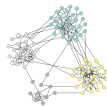
$$P(\mathbf{k}|\mathbf{e}, \mathbf{b}) = \prod_r \binom{n_r}{e_r}^{-1}$$



Edge counts \mathbf{e} .



Degrees, \mathbf{k} .

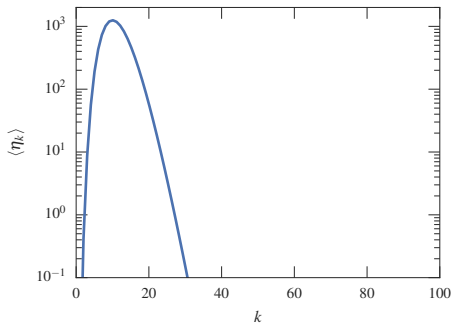
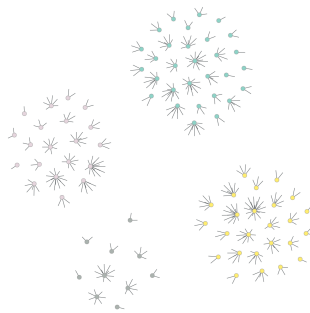
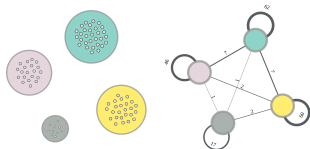


Network, \mathbf{A} .

PRIOR FOR THE DEGREES

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

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



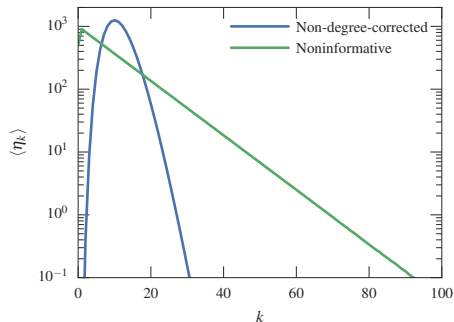
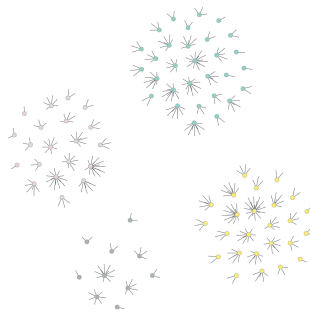
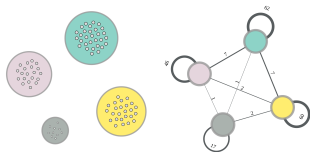
PRIOR FOR THE DEGREES

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

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

Option 1: Noninformative

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



PRIOR FOR THE DEGREES

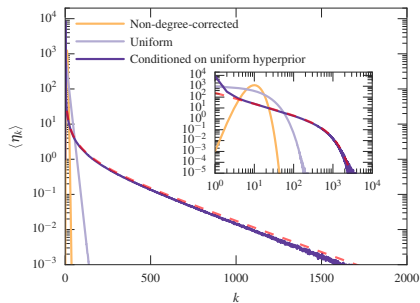
Option 2: Conditioned on degree distribution

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

PRIOR FOR THE DEGREES

Option 2: Conditioned on degree distribution

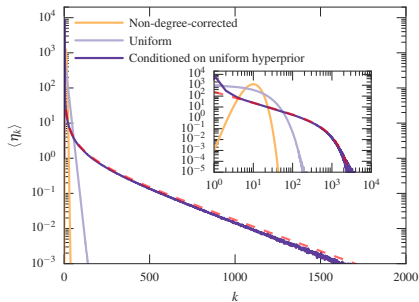
$$\begin{aligned} P(\mathbf{k}|\mathbf{b}, \mathbf{e}) &= P(\mathbf{k}|\boldsymbol{\eta})P(\boldsymbol{\eta}) \\ &= \prod_r \frac{n_r!}{\prod_r \eta_k^r!} \times \prod_r q(e_r, n_r)^{-1} \end{aligned}$$



PRIOR FOR THE DEGREES

Option 2: Conditioned on degree distribution

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



Bose-Einstein statistics

N “bosons” (nodes)

$k_i \rightarrow$ “energy level” (degree)

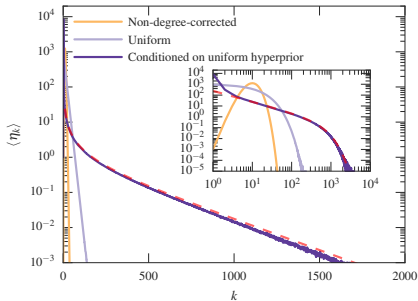
$$\langle \eta_k \rangle \approx \frac{1}{\exp\left(k\sqrt{\zeta(2)/E}\right) - 1}$$

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

PRIOR FOR THE DEGREES

Option 2: Conditioned on degree distribution

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



Bose-Einstein statistics

N “bosons” (nodes)

$k_i \rightarrow$ “energy level” (degree)

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

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

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

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

PRIOR FOR THE DEGREES: INTEGER PARTITIONS

$q(m, n) \rightarrow$ 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
 $O(M^2)$.

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

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

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

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

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

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

(Szekeres, 1951)

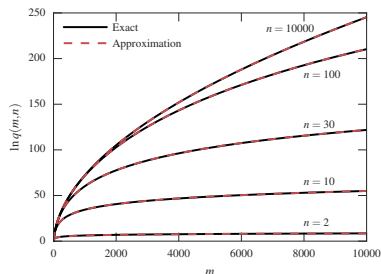
PRIOR FOR THE DEGREES: INTEGER PARTITIONS

$q(m, n) \rightarrow$ 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

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

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

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

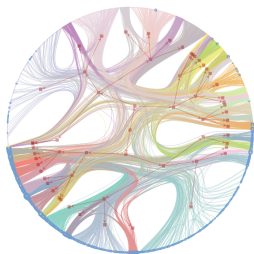
$$g(u) = \frac{2v(u)}{u} - u \ln(1 - e^{-v(u)})$$

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

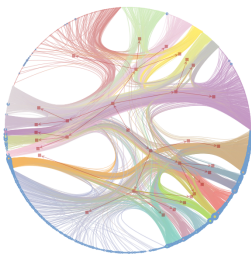
(Szekeres, 1951)

EXAMPLE: POLITICAL BLOGS REDUX

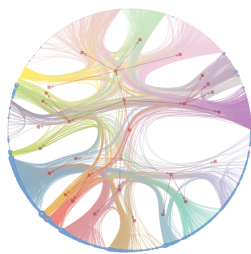
$N = 1,222, E = 19,027$



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



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



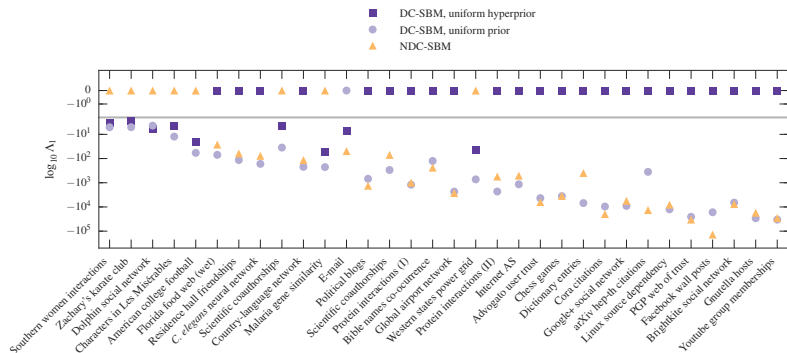
(c) DC-SBM, uniform
hyperprior, $B_1 = 20$,
 $\Sigma \approx 84890$ bits

MODEL SELECTION

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

MODEL SELECTION

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



INFERRING GROUP ASSIGNMENTS: SAMPLING VS. MAXIMIZATION

$\hat{\mathbf{b}}$ \rightarrow estimator, \mathbf{b}^* \rightarrow true partition

Maximum *a posteriori* (MAP)

Indicator loss function:

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

Maximization over the posterior

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

yields the MAP estimator

$$\hat{\mathbf{b}} = \operatorname{argmax}_{\mathbf{b}} P(\mathbf{b} | \mathbf{A}).$$

(Identical to MDL.)

Node marginals

Overlap loss function:

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

Maximization over the posterior

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

yields the marginal estimator

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

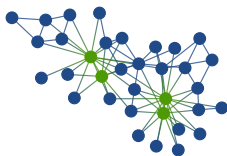
$$\pi_i(r) = \sum_{\mathbf{b} \setminus b_i} P(b_i = r, \mathbf{b} \setminus b_i | \mathbf{A}).$$

SAMPLING VS. MAXIMIZATION

Zachary's karate club



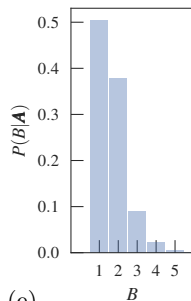
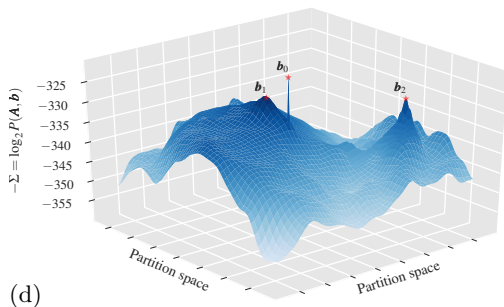
(a) \mathbf{b}_0 , $\Sigma = 321.3$ bits



(b) \mathbf{b}_1 , $\Sigma = 327.5$ bits



(c) \mathbf{b}_2 , $\Sigma = 329.3$ bits



SAMPLING VS. MAXIMIZATION

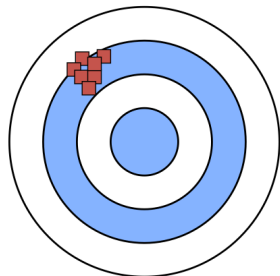
BIAS-VARIANCE TRADE-OFF

Maximization (MDL)

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

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

Lower variance, higher bias



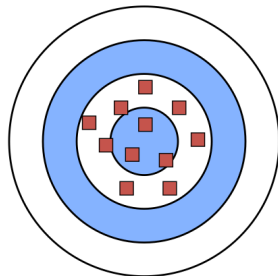
*Tendency to **underfit**.*

Sampling

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

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

Higher variance, lower bias



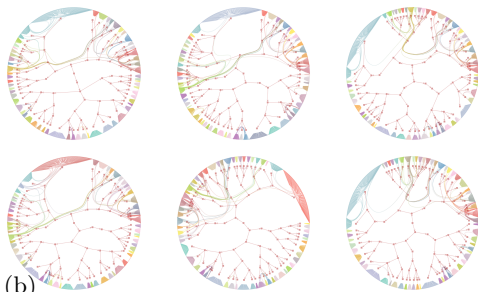
*Tendency to **overfit**.*

SAMPLING VS. MAXIMIZATION

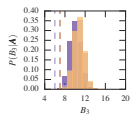
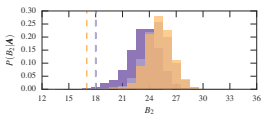
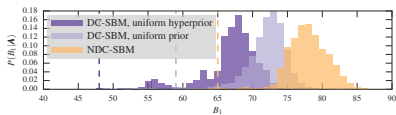
Collaborations between scientists ($N = 1,589$, $E = 2,742$)



(a)



(b)



INFERENCE ALGORITHM: METROPOLIS-HASTINGS

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

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

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 .

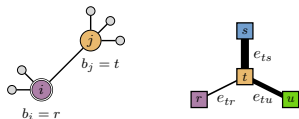
In contrast to:

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

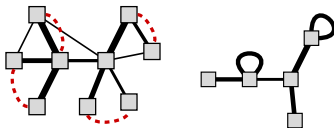
- ▶ Choose a random vertex i (happens to belong to group r).
- ▶ Move it to a random group $s \in [1, B]$, chosen with a probability $p(r \rightarrow s|t)$ proportional to $e_{ts} + \epsilon$, where t is the group membership of a randomly chosen neighbour of i .



Fast mixing times.

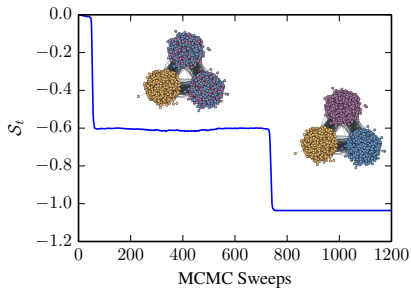
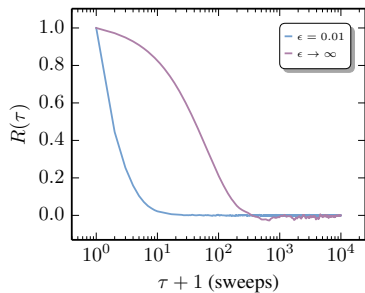
Agglomerative initialization

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



Avoids metastable states.

EFFICIENT INFERENCE: OTHER IMPROVEMENTS



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

