Statistical Inference of Modules in Large-Scale Networks

Tiago P. Peixoto

Universität Bremen

Oldenburg, November 2013

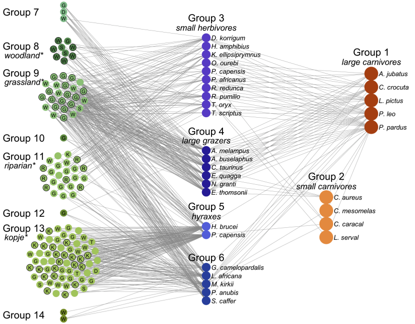
NETWORKS

Networks form the substrate of a wide variety of complex systems.

Networks

Networks form the substrate of a wide variety of complex systems.

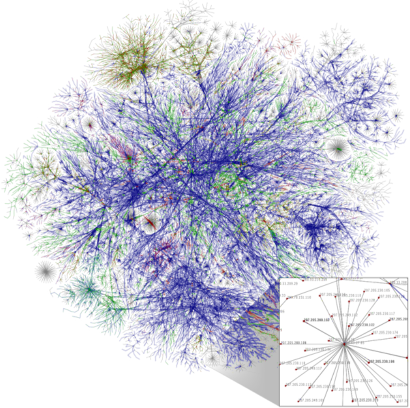
- Food webs



Networks

Networks form the substrate of a wide variety of complex systems.

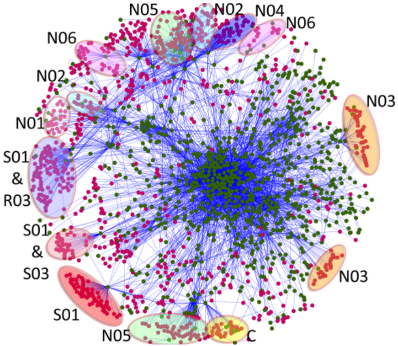
- Food webs
- Communication networks (e.g. Internet)



Networks

Networks form the substrate of a wide variety of complex systems.

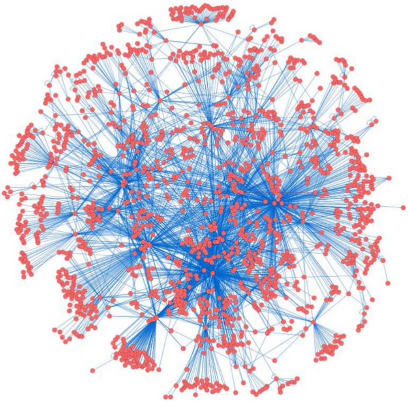
- Food webs
- Communication networks (e.g. Internet)
- Protein-Interaction



Networks

Networks form the substrate of a wide variety of complex systems.

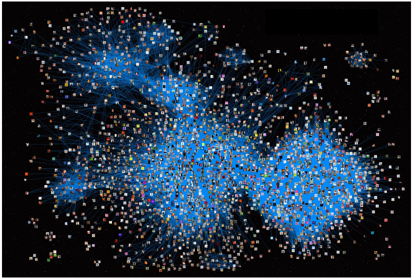
- Food webs
- Communication networks (e.g. Internet)
- Protein-Interaction
- Gene regulation



Networks

Networks form the substrate of a wide variety of complex systems.

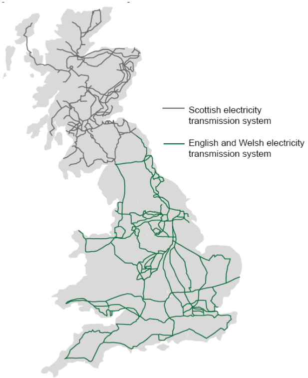
- Food webs
- Communication networks (e.g. Internet)
- Protein-Interaction
- Gene regulation
- Social Networks (Acquaintances, Influence, etc)



Networks

Networks form the substrate of a wide variety of complex systems.

- Food webs
- Communication networks (e.g. Internet)
- Protein-Interaction
- Gene regulation
- Social Networks (Acquaintances, Influence, etc)
- Infrastructure (Power Grids, Gas pipelines)



Networks

Networks form the substrate of a wide variety of complex systems.

- Food webs
- Communication networks (e.g. Internet)
- Protein-Interaction
- Gene regulation
- Social Networks (Acquaintances, Influence, etc)
- Infrastructure (Power Grids, Gas pipelines)
- Transport (Roads, Airports, ...)

NETWORKS

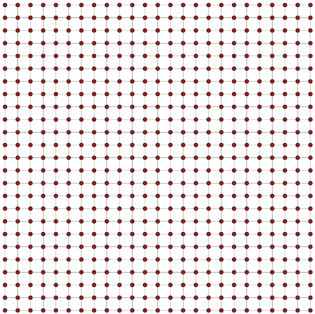
Networks form the substrate of a wide variety of complex systems.

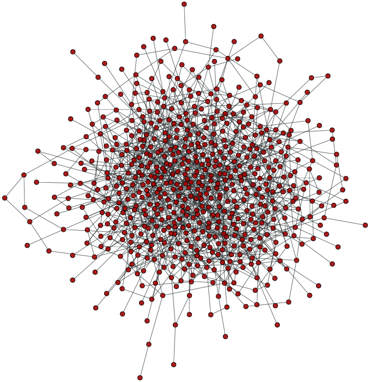
- Food webs
- Communication networks (e.g. Internet)
- Protein-Interaction
- Gene regulation
- Social Networks (Acquaintances, Influence, etc)
- Infrastructure (Power Grids, Gas pipelines)
- Transport (Roads, Airports, ...)

Structure $\leftrightarrow$ Dynamics $\leftrightarrow$ Evolution

Structure of Networks

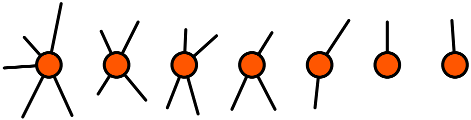
Somewhere between regular and random



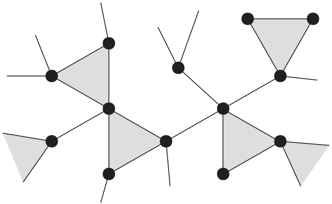


Local (or small-scale) structure

Degree distribution

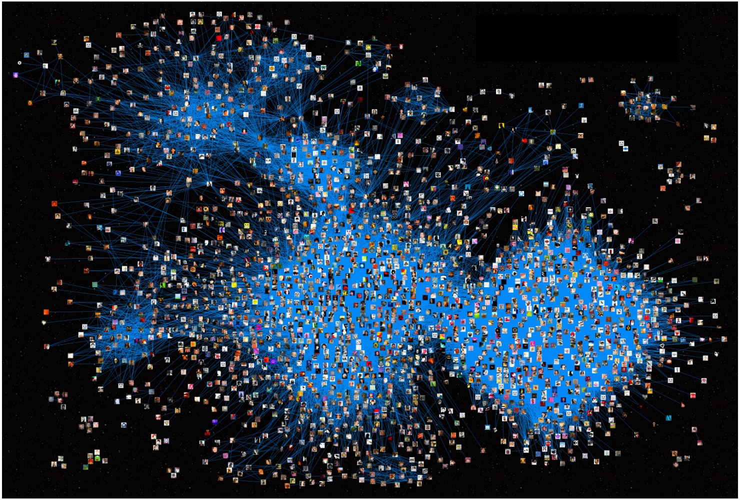


Clustering (Triangles)

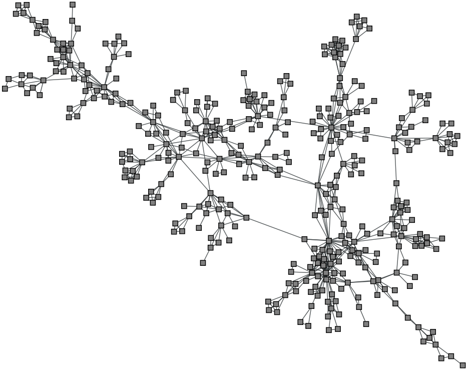


Global (or large-scale) structure

Modular structure

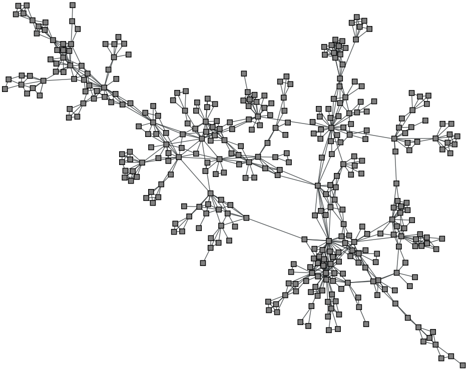


Levels of description

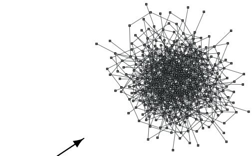


Real Network
(Network scientists)

Levels of description

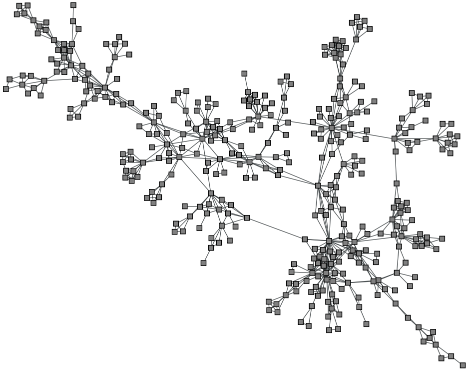


Real Network
(Network scientists)

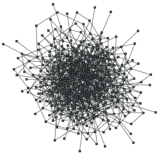


Random graph
(no structure)

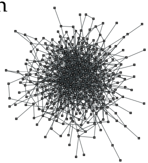
Levels of description



Real Network
(Network scientists)

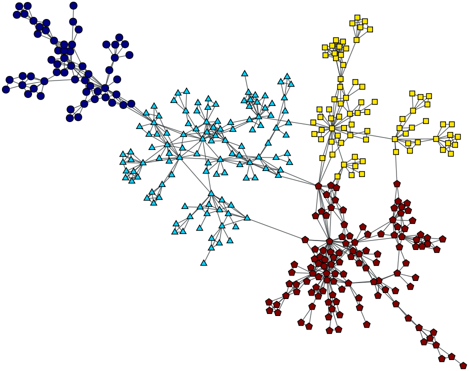


Random graph
(no structure)

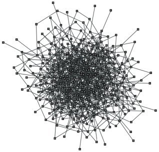


Degree
distribution

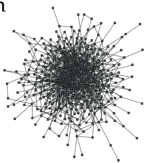
Levels of description



Real Network
(Network scientists)

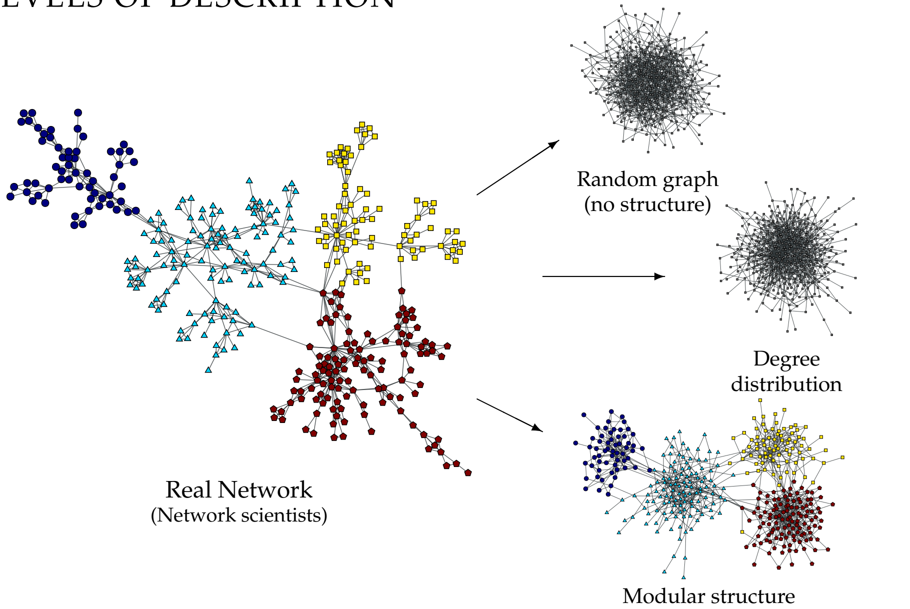


Random graph
(no structure)

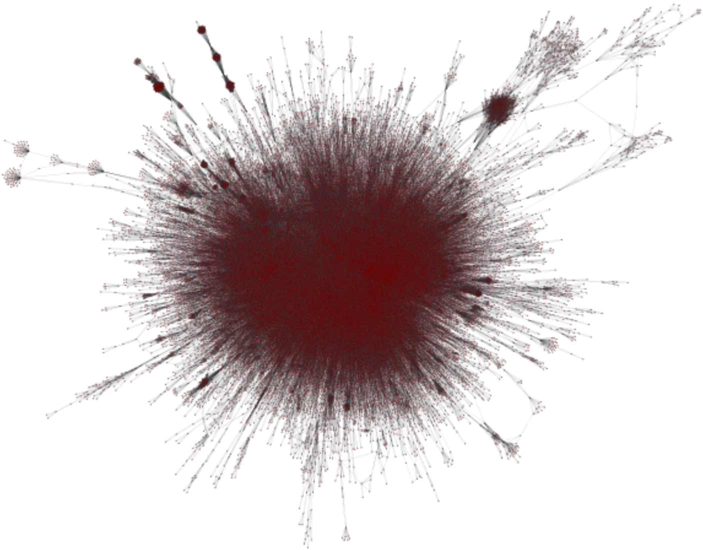


Degree
distribution

Levels of description

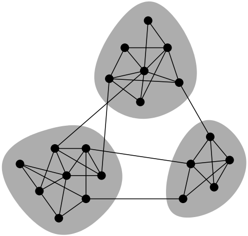


Problem: How to detect and characterize modular structure?

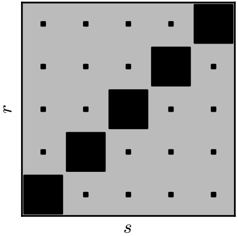


Simpler approach: Community Structure

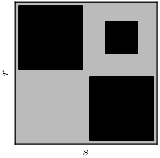
Focus on one of many possible patterns: *Communities*



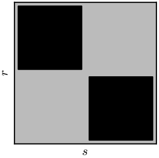
Communities



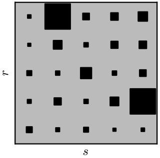
Excludes: Core-periphery, bipartite, multipartite, ...



Core-periphery



Bipartite



Arbitrary

Community Structure: Modularity maximization.

Idea: Find the partition of nodes, such that the fraction of internal edges is higher than expected given a *null model*.

$$Q = \frac{1}{2E} \sum_{ij} (A_{ij} - p_{ij}) \delta_{b_i, b_j}$$

$$A_{ij} = \begin{cases} 1, & \text{if } i, j \text{ are connected} \\ 0, & \text{otherwise.} \end{cases}$$

$$p_{ij} = \frac{k_i k_j}{2E} \rightarrow \text{Random graph}$$

Task: Find the partition $\{b_i\}$ which maximizes Q.

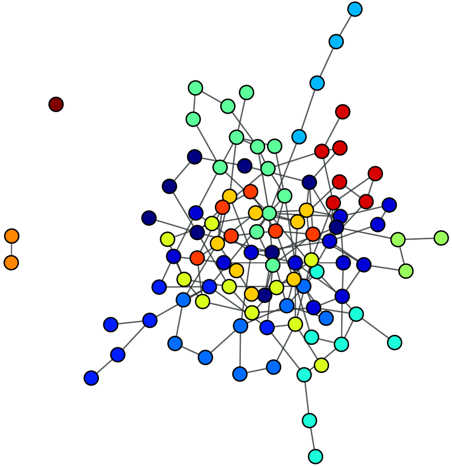
Newman and Girvan, Phys. Rev. E 69 026113 (2004).

Problems with modularity

- Restricted to community structure.
- Maximization of Q is difficult (NP-Complete).
- No built-in validation: Cannot distinguish structure from noise.
- Lack of consistency: Many different partitions with a similar Q.
- Resolution limit: Small modules cannot be detected.

Modularity: No built-in validation

A fully random graph:

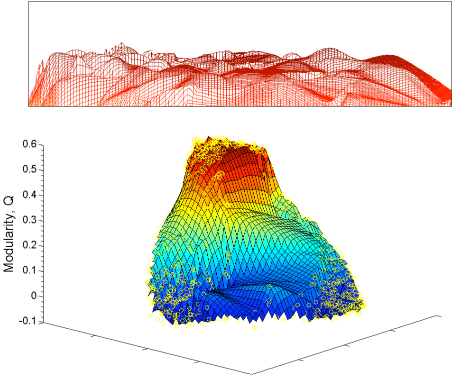


$$Q \simeq 0.55$$

Guimerà et al, Phys. Rev. E 70, 025101 (2004)

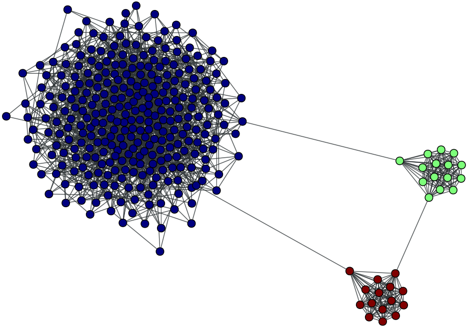
Modularity: Lack of consistency

Partitions are often degenerate.



Good et al, Phys. Rev. E 81, 046106 (2010)

Modularity: Resolution limit



The two smaller communities are merged together if $e_c < \sqrt{E/2}$

(Fortunato et al, PNAS 2007)

Characteristic block size: $\sim \sqrt{E}$

Distorted picture: Heavily skewed towards homogeneous blocks.

Why does modularity fail?

Global model? Wrong null model?

The main problem: It is a rather *ad hoc* approach.

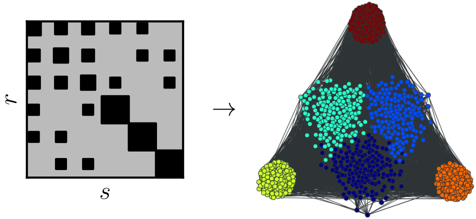
> Comparison with the null model is not connected with statistical significance.

A better approach: Stochastic blockmodels

Statistical inference of *generative models*

Traditional: N nodes divided into B blocks.

Parameters: $b_i \rightarrow$ block membership of node i
$e_{rs} \rightarrow$ number of edges from block r to s.



Degree-corrected: Arbitrary degree sequence: $\{k_i\}$

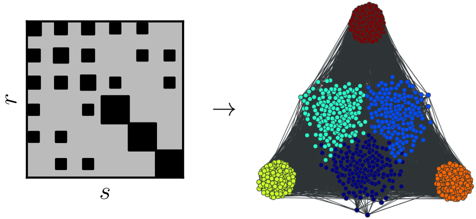
P. W. Holland et al., Soc Networks 5, 109 (1983)
B. Karrer and M. Newman, Phys. Rev. E 83, 016107 (2011)

A better approach: Stochastic blockmodels

Statistical inference of *generative models*

Traditional: N nodes divided into B blocks.

Parameters: $b_i \to$ block membership of node i
$e_{rs} \to$ number of edges from block r to s.



Degree-corrected: Arbitrary degree sequence: $\{k_i\}$

- Not restricted to assortative structures ("communities"). Could be bipartite, multipartite, core-periphery, etc.
- Easily modifiable for directed graphs.
- Inference $\to$ Maximize posterior probability $\mathcal{P}(G|\{e_{rs}\}, \{b_i\})$

P. W. Holland et al., Soc Networks 5, 109 (1983)
B. Karrer and M. Newman, Phys. Rev. E 83, 016107 (2011)

MAXIMUM LIKELIHOOD

Microcanonical formulation: $\mathcal{P}(G|\{e_{rs}\},\{b_i\}) = \dfrac{1}{\Omega(\{e_{rs}\},\{b_i\})}$

Maximum likelihood

Microcanonical formulation: $\mathcal{P}(G|\{e_{rs}\},\{b_i\}) = \dfrac{1}{\Omega(\{e_{rs}\},\{b_i\})}$

Ensemble entropy: $\mathcal{S}(\{e_{rs}\},\{b_i\}) = \ln\Omega(\{e_{rs}\},\{b_i\}) = -\ln\mathcal{P}(G|\{e_{rs}\},\{b_i\})$

T.P.P., PRE 85, 056122 (2012)

Maximum likelihood

Microcanonical formulation: $\mathcal{P}(G|\{e_{rs}\},\{b_i\}) = \dfrac{1}{\Omega(\{e_{rs}\},\{b_i\})}$

Ensemble entropy: $\mathcal{S}(\{e_{rs}\},\{b_i\}) = \ln\Omega(\{e_{rs}\},\{b_i\}) = -\ln\mathcal{P}(G|\{e_{rs}\},\{b_i\})$

$$\mathcal{S} \cong -E - \sum_k N_k \ln k! - \frac{1}{2}\sum_{rs} e_{rs}\ln\left(\frac{e_{rs}}{e_r e_s}\right)$$

Inference ↔ Compression

$$\max_{\{e_{rs}\},\{b_i\}} \ln\mathcal{P} \equiv \min_{\{e_{rs}\},\{b_i\}} \mathcal{S}$$

→ Minimization of information required to describe the network, *when the model is known*.

Maximum likelihood

Microcanonical formulation: $\mathcal{P}(G|\{e_{rs}\},\{b_i\}) = \dfrac{1}{\Omega(\{e_{rs}\},\{b_i\})}$

Ensemble entropy: $\mathcal{S}(\{e_{rs}\},\{b_i\}) = \ln\Omega(\{e_{rs}\},\{b_i\}) = -\ln\mathcal{P}(G|\{e_{rs}\},\{b_i\})$

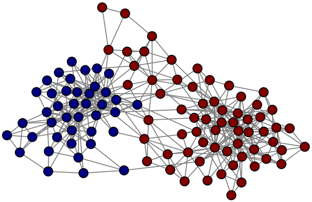
$$\mathcal{S} \cong -E - \sum_k N_k \ln k! - \frac{1}{2}\sum_{rs} e_{rs}\ln\left(\frac{e_{rs}}{e_r e_s}\right)$$

$$\max_{\{e_{rs}\},\{b_i\}} \ln\mathcal{P} \equiv \min_{\{e_{rs}\},\{b_i\}} \mathcal{S} \quad \rightarrow$$

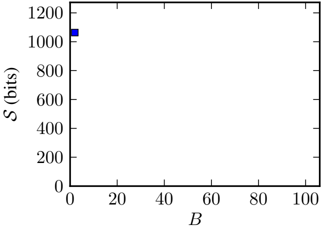
Inference ↔ Compression

Minimization of information required to describe the network, *when the model is known*.

Works like a charm if the number of blocks B is known...



$B = 2$



MAXIMUM LIKELIHOOD

Microcanonical formulation: $\mathcal{P}(G|\{e_{rs}\},\{b_i\}) = \dfrac{1}{\Omega(\{e_{rs}\},\{b_i\})}$

Ensemble entropy: $\mathcal{S}(\{e_{rs}\},\{b_i\}) = \ln \Omega(\{e_{rs}\},\{b_i\}) = -\ln \mathcal{P}(G|\{e_{rs}\},\{b_i\})$

$$\mathcal{S} \cong -E - \sum_k N_k \ln k! - \frac{1}{2}\sum_{rs} e_{rs} \ln\left(\frac{e_{rs}}{e_r e_s}\right)$$

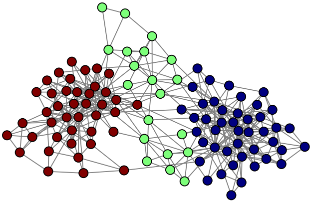
$$\max_{\{e_{rs}\},\{b_i\}} \ln \mathcal{P} \equiv \min_{\{e_{rs}\},\{b_i\}} \mathcal{S}$$

→

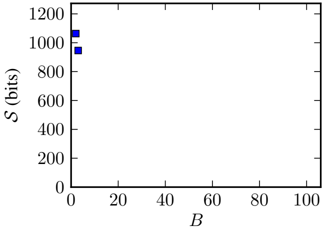
Inference ↔ Compression

Minimization of information required to describe the network, *when the model is known*.

Works like a charm if the number of blocks B is known...



$B = 3$



T.P.P., PRE 85, 056122 (2012)

Maximum likelihood

Microcanonical formulation: $\mathcal{P}(G|\{e_{rs}\},\{b_i\}) = \dfrac{1}{\Omega(\{e_{rs}\},\{b_i\})}$

Ensemble entropy: $\mathcal{S}(\{e_{rs}\},\{b_i\}) = \ln\Omega(\{e_{rs}\},\{b_i\}) = -\ln\mathcal{P}(G|\{e_{rs}\},\{b_i\})$

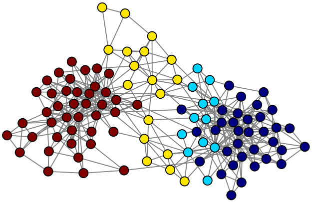
$$\mathcal{S} \cong -E - \sum_k N_k \ln k! - \frac{1}{2}\sum_{rs} e_{rs}\ln\left(\frac{e_{rs}}{e_r e_s}\right)$$

$$\max_{\{e_{rs}\},\{b_i\}} \ln\mathcal{P} \equiv \min_{\{e_{rs}\},\{b_i\}} \mathcal{S} \quad \rightarrow$$

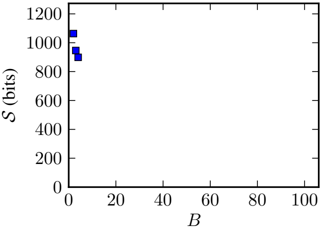
Inference ↔ Compression

Minimization of information required to describe the network, *when the model is known*.

Works like a charm if the number of blocks B is known...



$B = 4$



T.P.P., PRE 85, 056122 (2012)

Maximum likelihood

Microcanonical formulation: $\mathcal{P}(G|\{e_{rs}\},\{b_i\}) = \dfrac{1}{\Omega(\{e_{rs}\},\{b_i\})}$

Ensemble entropy: $\mathcal{S}(\{e_{rs}\},\{b_i\}) = \ln\Omega(\{e_{rs}\},\{b_i\}) = -\ln\mathcal{P}(G|\{e_{rs}\},\{b_i\})$

$$\mathcal{S} \cong -E - \sum_k N_k \ln k! - \frac{1}{2}\sum_{rs} e_{rs}\ln\left(\frac{e_{rs}}{e_r e_s}\right)$$

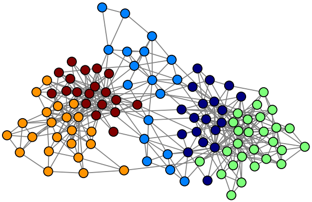
$$\max_{\{e_{rs}\},\{b_i\}} \ln\mathcal{P} \equiv \min_{\{e_{rs}\},\{b_i\}} \mathcal{S}$$

$\rightarrow$

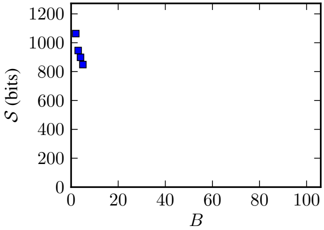
Inference $\leftrightarrow$ Compression

Minimization of information required to describe the network, *when the model is known*.

Works like a charm if the number of blocks B is known...



$B = 5$



T.P.P., PRE 85, 056122 (2012)

Maximum likelihood

Microcanonical formulation: $\mathcal{P}(G|\{e_{rs}\},\{b_i\}) = \dfrac{1}{\Omega(\{e_{rs}\},\{b_i\})}$

Ensemble entropy: $\mathcal{S}(\{e_{rs}\},\{b_i\}) = \ln\Omega(\{e_{rs}\},\{b_i\}) = -\ln\mathcal{P}(G|\{e_{rs}\},\{b_i\})$

$$\mathcal{S} \cong -E - \sum_k N_k \ln k! - \frac{1}{2}\sum_{rs} e_{rs} \ln\left(\frac{e_{rs}}{e_r e_s}\right)$$

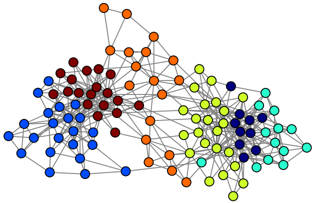
$$\max_{\{e_{rs}\},\{b_i\}} \ln\mathcal{P} \equiv \min_{\{e_{rs}\},\{b_i\}} \mathcal{S}$$

$\rightarrow$

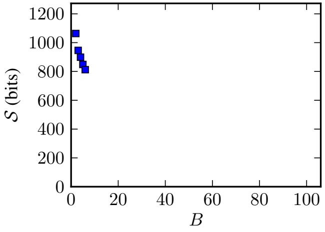
Inference $\leftrightarrow$ Compression

Minimization of information required to describe the network, *when the model is known*.

Works like a charm if the number of blocks B is known...



$B = 6$



T.P.P., PRE 85, 056122 (2012)

Maximum likelihood

Microcanonical formulation: $\mathcal{P}(G|\{e_{rs}\},\{b_i\}) = \dfrac{1}{\Omega(\{e_{rs}\},\{b_i\})}$

Ensemble entropy: $\mathcal{S}(\{e_{rs}\},\{b_i\}) = \ln\Omega(\{e_{rs}\},\{b_i\}) = -\ln\mathcal{P}(G|\{e_{rs}\},\{b_i\})$

$$\mathcal{S} \cong -E - \sum_k N_k \ln k! - \frac{1}{2}\sum_{rs} e_{rs}\ln\left(\frac{e_{rs}}{e_r e_s}\right)$$

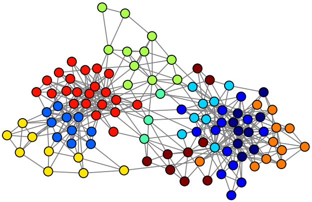
$$\max_{\{e_{rs}\},\{b_i\}} \ln\mathcal{P} \equiv \min_{\{e_{rs}\},\{b_i\}} \mathcal{S}$$

$\rightarrow$

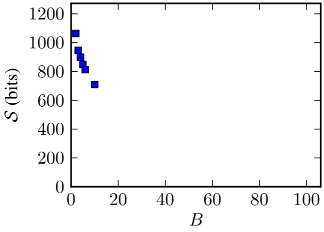
Inference $\leftrightarrow$ Compression

Minimization of information required to describe the network, *when the model is known*.

Works like a charm if the number of blocks B is known...



$B = 10$



Maximum likelihood

Microcanonical formulation: $\mathcal{P}(G|\{e_{rs}\},\{b_i\}) = \frac{1}{\Omega(\{e_{rs}\},\{b_i\})}$

Ensemble entropy: $\mathcal{S}(\{e_{rs}\},\{b_i\}) = \ln\Omega(\{e_{rs}\},\{b_i\}) = -\ln\mathcal{P}(G|\{e_{rs}\},\{b_i\})$

$$\mathcal{S} \cong -E - \sum_k N_k \ln k! - \frac{1}{2}\sum_{rs} e_{rs} \ln\left(\frac{e_{rs}}{e_r e_s}\right)$$

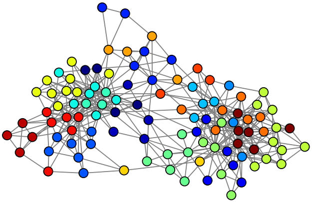
$$\max_{\{e_{rs}\},\{b_i\}} \ln\mathcal{P} \equiv \min_{\{e_{rs}\},\{b_i\}} \mathcal{S}$$

$\rightarrow$

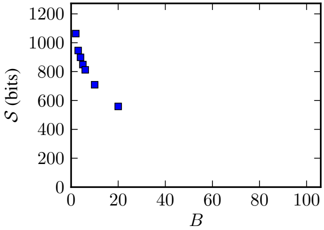
Inference $\leftrightarrow$ Compression

Minimization of information required to describe the network, *when the model is known*.

Works like a charm if the number of blocks B is known...



$B = 20$



Maximum likelihood

Microcanonical formulation: $\mathcal{P}(G|\{e_{rs}\},\{b_i\}) = \dfrac{1}{\Omega(\{e_{rs}\},\{b_i\})}$

Ensemble entropy: $\mathcal{S}(\{e_{rs}\},\{b_i\}) = \ln\Omega(\{e_{rs}\},\{b_i\}) = -\ln\mathcal{P}(G|\{e_{rs}\},\{b_i\})$

$$\mathcal{S} \cong -E - \sum_k N_k \ln k! - \frac{1}{2}\sum_{rs} e_{rs}\ln\left(\frac{e_{rs}}{e_r e_s}\right)$$

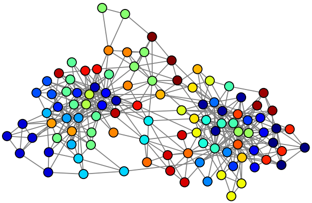
$$\max_{\{e_{rs}\},\{b_i\}} \ln\mathcal{P} \equiv \min_{\{e_{rs}\},\{b_i\}} \mathcal{S}$$

$\rightarrow$

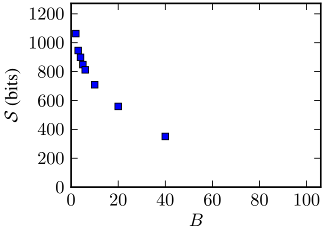
Inference $\leftrightarrow$ Compression

Minimization of information required to describe the network, *when the model is known*.

Works like a charm if the number of blocks B is known...



$B = 40$



T.P.P., PRE 85, 056122 (2012)

Maximum likelihood

Microcanonical formulation: $\mathcal{P}(G|\{e_{rs}\},\{b_i\}) = \dfrac{1}{\Omega(\{e_{rs}\},\{b_i\})}$

Ensemble entropy: $\mathcal{S}(\{e_{rs}\},\{b_i\}) = \ln\Omega(\{e_{rs}\},\{b_i\}) = -\ln\mathcal{P}(G|\{e_{rs}\},\{b_i\})$

$$\mathcal{S} \cong -E - \sum_k N_k \ln k! - \frac{1}{2}\sum_{rs} e_{rs}\ln\left(\frac{e_{rs}}{e_r e_s}\right)$$

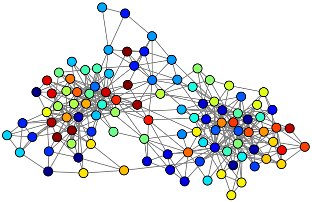
$$\max_{\{e_{rs}\},\{b_i\}} \ln\mathcal{P} \equiv \min_{\{e_{rs}\},\{b_i\}} \mathcal{S}$$

→

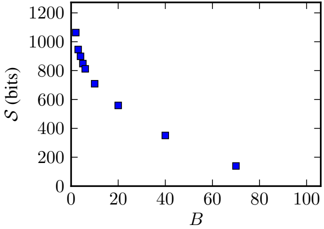
Inference ↔ Compression

Minimization of information required to describe the network, *when the model is known*.

Works like a charm if the number of blocks B is known...



$B = 70$



T.P.P., PRE 85, 056122 (2012)

Maximum likelihood

Microcanonical formulation: $\mathcal{P}(G|\{e_{rs}\},\{b_i\}) = \dfrac{1}{\Omega(\{e_{rs}\},\{b_i\})}$

Ensemble entropy: $\mathcal{S}(\{e_{rs}\},\{b_i\}) = \ln\Omega(\{e_{rs}\},\{b_i\}) = -\ln\mathcal{P}(G|\{e_{rs}\},\{b_i\})$

$$\mathcal{S} \cong -E - \sum_k N_k \ln k! - \frac{1}{2}\sum_{rs} e_{rs}\ln\left(\frac{e_{rs}}{e_r e_s}\right)$$

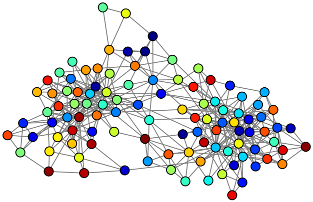
$$\max_{\{e_{rs}\},\{b_i\}} \ln\mathcal{P} \equiv \min_{\{e_{rs}\},\{b_i\}} \mathcal{S}$$

→

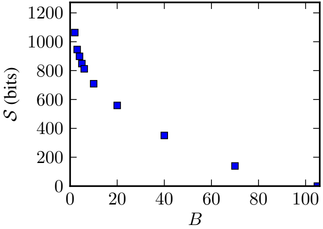
Inference ↔ Compression

Minimization of information required to describe the network, *when the model is known*.

Works like a charm if the number of blocks B is known...



$B = N$



... otherwise: **Overfitting!**

T.P.P., PRE 85, 056122 (2012)

Solution: Minimum Description Length

$\mathcal{S} \rightarrow$ Information required to describe the network, *when the model is known*.

Solution: Minimum Description Length

$\mathcal{S} \rightarrow$ Information required to describe the network, *when the model is known*.

$\mathcal{L} \rightarrow$ Information required to describe *the model*.

SOLUTION: MINIMUM DESCRIPTION LENGTH

$\mathcal{S} \rightarrow$ Information required to describe the network, *when the model is known*.

$\mathcal{L} \rightarrow$ Information required to describe *the model*.

$$\mathcal{L} = \ln \left(\!\! \binom{\left(\!\binom{B}{2}\!\right)}{E} \!\! \right) + \ln N! - \sum_r \ln n_r! - N \sum_k p_k \ln p_k$$

Solution: Minimum Description Length

$\mathcal{S} \rightarrow$ Information required to describe the network, *when the model is known*.

$\mathcal{L} \rightarrow$ Information required to describe *the model*.

$$\mathcal{L} = \ln \left(\!\! \binom{\left(\!\binom{B}{2}\!\right)}{E} \!\! \right) + \ln N! - \sum_r \ln n_r! - N \sum_k p_k \ln p_k$$

$$\Sigma = \mathcal{S} + \mathcal{L}$$

Total information necessary, without a priori knowledge of the model!

Solution: Minimum Description Length

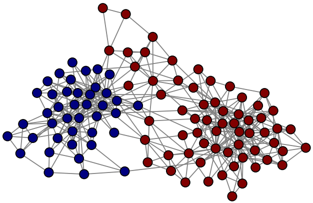
$\mathcal{S} \rightarrow$ Information required to describe the network, *when the model is known*.

$\mathcal{L} \rightarrow$ Information required to describe *the model*.

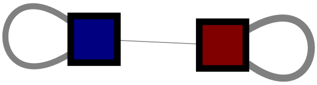
$$\mathcal{L} = \ln \left(\!\!\binom{\left(\!\binom{B}{2}\!\right)}{E}\!\!\right) + \ln N! - \sum_r \ln n_r! - N \sum_k p_k \ln p_k$$

$$\Sigma = \mathcal{S} + \mathcal{L}$$

Total information necessary, without a priori knowledge of the model!



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



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

$\Sigma \simeq 1926.9$ bits

Solution: Minimum Description Length

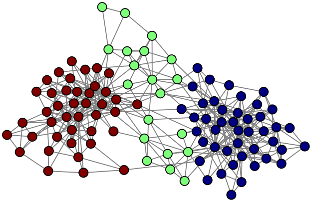
$\mathcal{S} \rightarrow$ Information required to describe the network, *when the model is known*.

$\mathcal{L} \rightarrow$ Information required to describe *the model*.

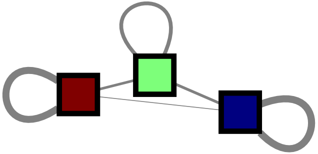
$$\mathcal{L} = \ln \left(\!\!\binom{\left(\!\binom{B}{2}\!\right)}{E}\!\!\right) + \ln N! - \textstyle\sum_r \ln n_r! - N \sum_k p_k \ln p_k$$

$$\Sigma = \mathcal{S} + \mathcal{L}$$

Total information necessary, without a priori knowledge of the model!



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



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

$\Sigma \simeq 1891.6$ bits

T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

SOLUTION: MINIMUM DESCRIPTION LENGTH

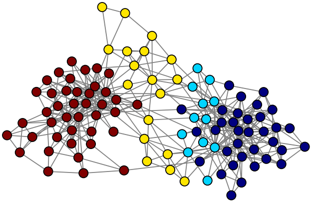
$\mathcal{S} \rightarrow$ Information required to describe the network, *when the model is known*.

$\mathcal{L} \rightarrow$ Information required to describe *the model*.

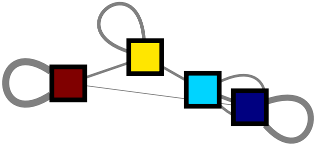
$$\mathcal{L} = \ln \left(\!\!\binom{\binom{B}{2}}{E}\!\!\right) + \ln N! - \sum_r \ln n_r! - N \sum_k p_k \ln p_k$$

$$\Sigma = \mathcal{S} + \mathcal{L}$$

Total information necessary, without a priori knowledge of the model!



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



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

$\Sigma \simeq 1911.5$ bits

T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

Solution: Minimum Description Length

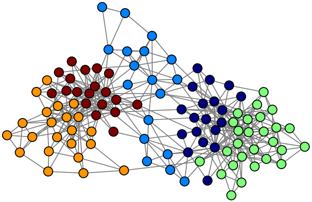
$\mathcal{S} \rightarrow$ Information required to describe the network, *when the model is known*.

$\mathcal{L} \rightarrow$ Information required to describe *the model*.

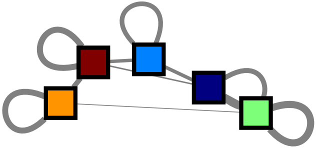
$$\mathcal{L} = \ln \left(\!\!\left(\begin{array}{c} \left(\!\binom{B}{2}\!\right) \\ E \end{array} \right)\!\!\right) + \ln N! - \textstyle\sum_r \ln n_r! - N \sum_k p_k \ln p_k$$

$$\Sigma = \mathcal{S} + \mathcal{L}$$

Total information necessary, without a priori knowledge of the model!



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



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

$\Sigma \simeq 1921.3$ bits

T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

Solution: Minimum Description Length

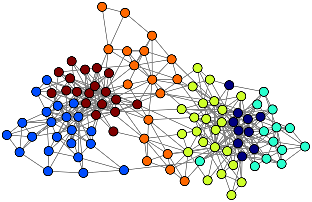
$\mathcal{S} \rightarrow$ Information required to describe the network, *when the model is known*.

$\mathcal{L} \rightarrow$ Information required to describe *the model*.

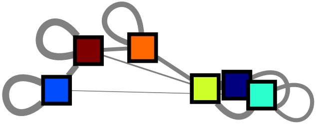
$$\mathcal{L} = \ln \left(\!\!\binom{\binom{B}{2}}{E}\!\!\right) + \ln N! - \sum_r \ln n_r! - N \sum_k p_k \ln p_k$$

$$\Sigma = \mathcal{S} + \mathcal{L}$$

Total information necessary, without a priori knowledge of the model!



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



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

$\Sigma \simeq 1940.9$ bits

T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

Solution: Minimum Description Length

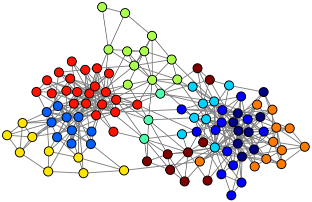
$\mathcal{S} \to$ Information required to describe the network, *when the model is known*.

$\mathcal{L} \to$ Information required to describe *the model*.

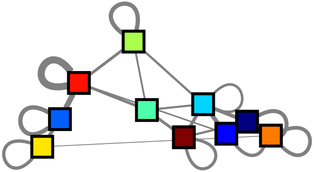
$$\mathcal{L} = \ln\left(\!\!\left(\begin{matrix}\left(\!\binom{B}{2}\!\right)\\ E\end{matrix}\right)\!\!\right) + \ln N! - \textstyle\sum_r \ln n_r! - N\sum_k p_k \ln p_k$$

$$\Sigma = \mathcal{S} + \mathcal{L}$$

Total information necessary, without a priori knowledge of the model!



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



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

$\Sigma \simeq 2041.8$ bits

T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

Solution: Minimum Description Length

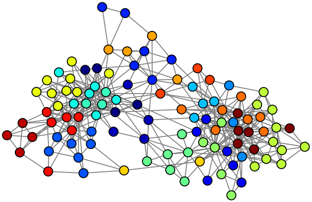
$\mathcal{S} \rightarrow$ Information required to describe the network, *when the model is known*.

$\mathcal{L} \rightarrow$ Information required to describe *the model*.

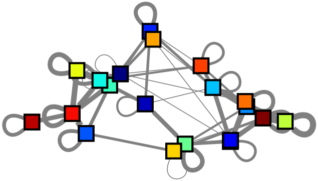
$$\mathcal{L} = \ln \left(\!\!\binom{\binom{B}{2}}{E}\!\!\right) + \ln N! - \textstyle\sum_r \ln n_r! - N \sum_k p_k \ln p_k$$

$$\Sigma = \mathcal{S} + \mathcal{L}$$

Total information necessary, without a priori knowledge of the model!



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



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

$\Sigma \simeq 2338.6$ bits

Solution: Minimum Description Length

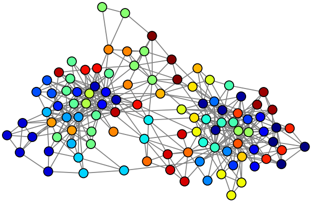
$\mathcal{S} \rightarrow$ Information required to describe the network, *when the model is known*.

$\mathcal{L} \rightarrow$ Information required to describe *the model*.

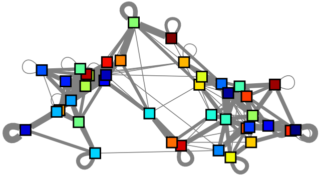
$$\mathcal{L} = \ln \left(\!\!\binom{\left(\!\binom{B}{2}\!\right)}{E}\!\!\right) + \ln N! - \sum_r \ln n_r! - N \sum_k p_k \ln p_k$$

$$\Sigma = \mathcal{S} + \mathcal{L}$$

Total information necessary, without a priori knowledge of the model!



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



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

$\Sigma \simeq 2823.1$ bits

T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

Solution: Minimum Description Length

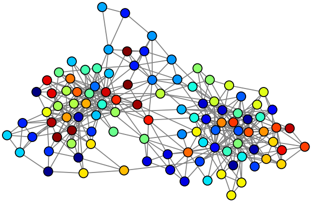
$\mathcal{S} \rightarrow$ Information required to describe the network, *when the model is known*.

$\mathcal{L} \rightarrow$ Information required to describe *the model*.

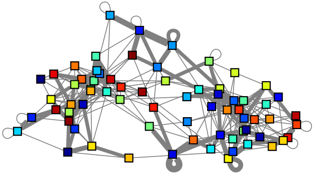
$$\mathcal{L} = \ln \left(\!\!\binom{\left(\!\binom{B}{2}\!\right)}{E}\!\!\right) + \ln N! - \textstyle\sum_r \ln n_r! - N \sum_k p_k \ln p_k$$

$$\Sigma = \mathcal{S} + \mathcal{L}$$

Total information necessary, without a priori knowledge of the model!



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



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

$\Sigma \simeq 3308.7$ bits

T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

Solution: Minimum Description Length

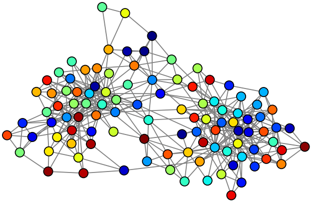
$\mathcal{S} \rightarrow$ Information required to describe the network, *when the model is known*.

$\mathcal{L} \rightarrow$ Information required to describe *the model*.

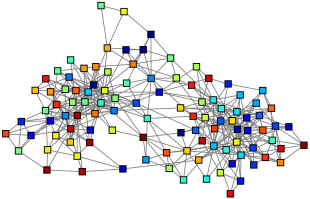
$$\mathcal{L} = \ln \left(\!\!\binom{\left(\!\binom{B}{2}\!\right)}{E}\!\!\right) + \ln N! - \textstyle\sum_r \ln n_r! - N \sum_k p_k \ln p_k$$

$\Sigma = \mathcal{S} + \mathcal{L}$

Total information necessary, without a priori knowledge of the model!



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



Model, $\mathcal{L} \simeq 2973.0$ bits

$\Sigma \simeq 3714.9$ bits

T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

Solution: Minimum Description Length

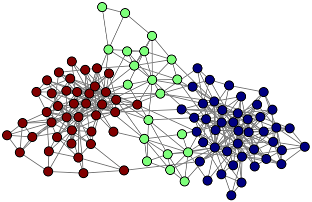
$\mathcal{S} \rightarrow$ Information required to describe the network, *when the model is known*.

$\mathcal{L} \rightarrow$ Information required to describe *the model*.

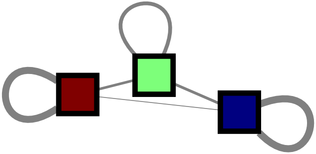
$$\mathcal{L} = \ln \left(\!\!\left({\left(\!\binom{B}{2}\!\right) \atop E} \right)\!\!\right) + \ln N! - \sum_r \ln n_r! - N \sum_k p_k \ln p_k$$

$$\Sigma = \mathcal{S} + \mathcal{L}$$

Total information necessary, without a priori knowledge of the model!



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



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

$\Sigma \simeq 1891.6$ bits

T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

Solution: Minimum Description Length

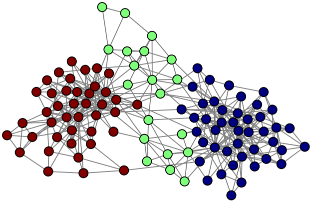
$\mathcal{S} \rightarrow$ Information required to describe the network, *when the model is known*.

$\mathcal{L} \rightarrow$ Information required to describe *the model*.

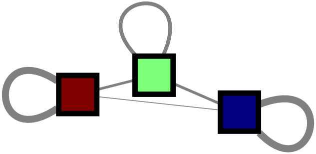
$$\mathcal{L} = \ln \left(\!\!\binom{\binom{B}{2}}{E}\!\!\right) + \ln N! - \textstyle\sum_r \ln n_r! - N \sum_k p_k \ln p_k$$

$\Sigma = \mathcal{S} + \mathcal{L}$

Total information necessary, without a priori knowledge of the model!

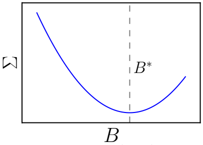


$B = 3$, $\mathcal{S} \simeq 1688.1$ bits



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

$\Sigma \simeq 1891.6$ bits



$\rightarrow$

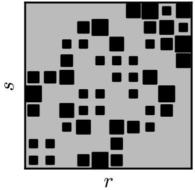
Occam's razor

The best model is the one which most compresses the data.

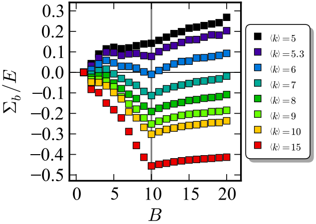
T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

WORKS VERY WELL... IF THE BLOCK STRUCTURE IS DETECTABLE!

Generated ($B = 10$)



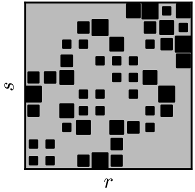
Planted



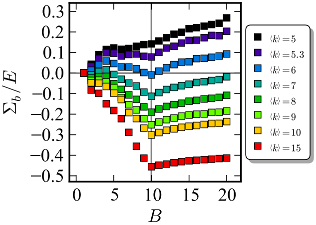
T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

WORKS VERY WELL... IF THE BLOCK STRUCTURE IS DETECTABLE!

Generated ($B = 10$)

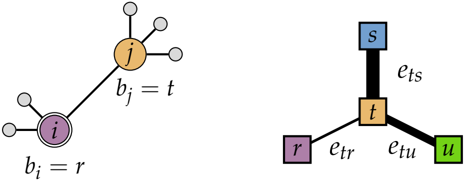


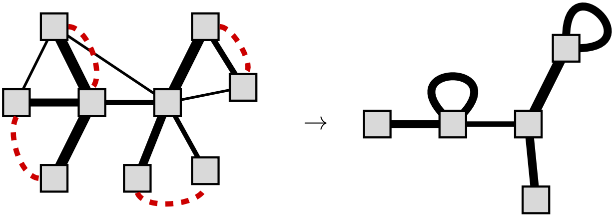
Planted



Advantage of MDL: Very efficient

Scalable algorithm: (fast) MCMC / Greedy Aggomeration





arXiv:1310.4378

Total running time: $O(N \ln^2 N)$

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)

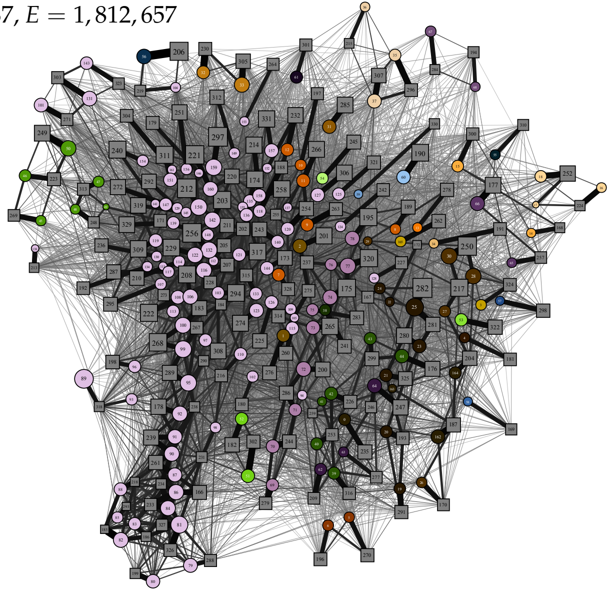
Bipartite network of actors and films.
Fairly large: $N = 372,787$, $E = 1,812,657$
MDL selects: $B = 332$

Example: The Internet Movie Database (IMDB)

Bipartite network of actors and films.

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

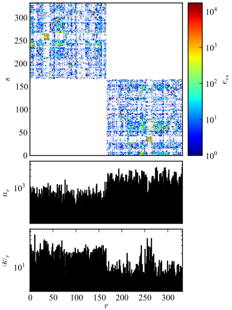
MDL selects: $B = 332$



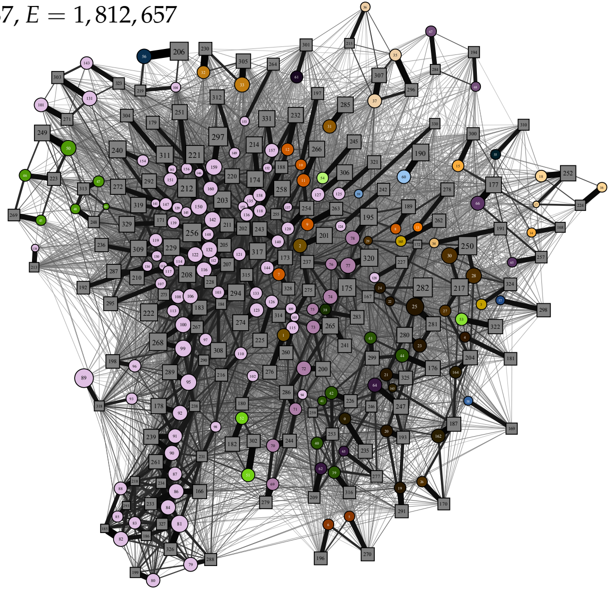
T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

Example: The Internet Movie Database (IMDB)

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



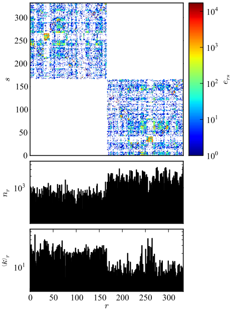
Bipartiteness is fully uncovered!

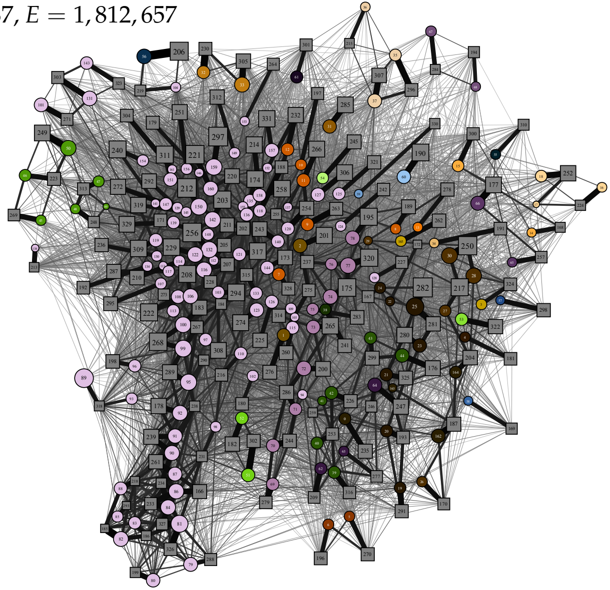


T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

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!

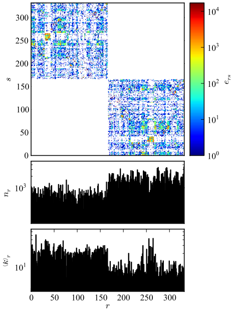
Detects meaningful features:

- Temporal
- Spatial (Country)
- Type/Genre

T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

Example: The Internet Movie Database (IMDB)

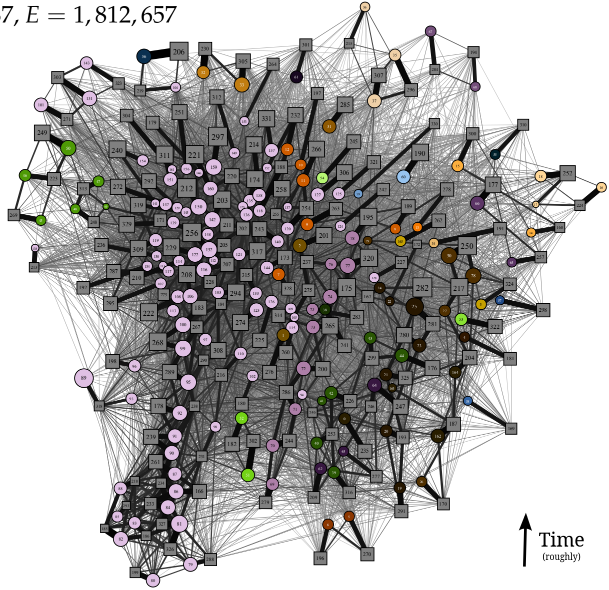
Bipartite network of actors and films.
Fairly large: $N = 372,787$, $E = 1,812,657$
MDL selects: $B = 332$



Bipartiteness is fully uncovered!

Detects meaningful features:

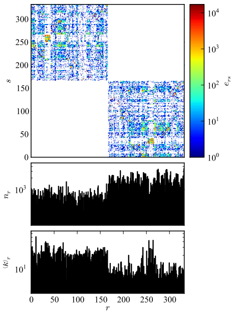
- Temporal
- Spatial (Country)
- Type/Genre



T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

Example: The Internet Movie Database (IMDB)

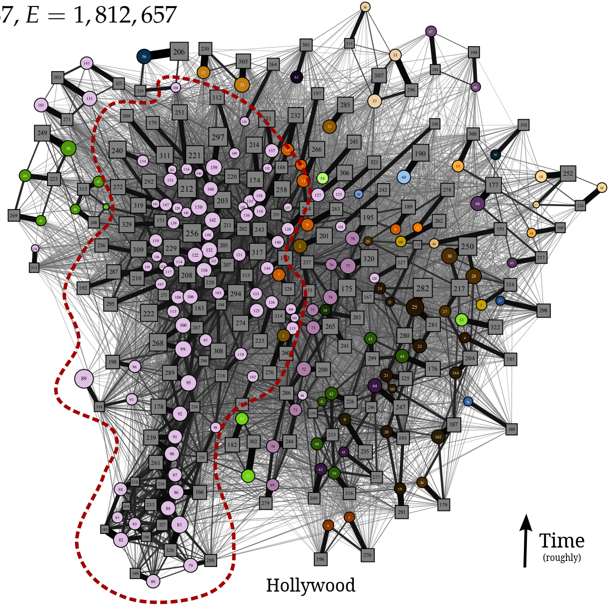
Bipartite network of actors and films.
Fairly large: $N = 372{,}787$, $E = 1{,}812{,}657$
MDL selects: $B = 332$



Bipartiteness is fully uncovered!

Detects meaningful features:

- Temporal
- Spatial (Country)
- Type/Genre



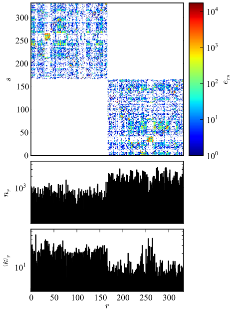
T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

Example: The Internet Movie Database (IMDB)

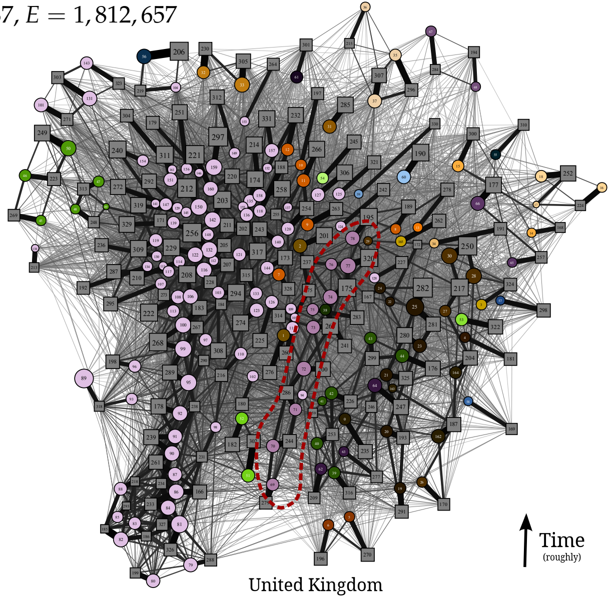
Bipartite network of actors and films.

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

MDL selects: $B = 332$



Bipartiteness is fully uncovered!



Detects meaningful features:

- Temporal
- Spatial (Country)
- Type/Genre

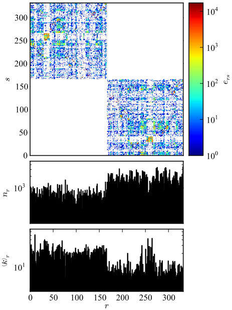
T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

Example: The Internet Movie Database (IMDB)

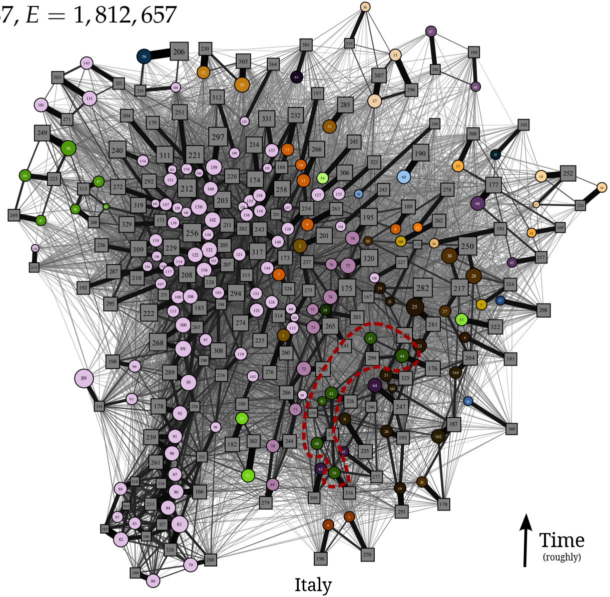
Bipartite network of actors and films.

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

MDL selects: $B = 332$



Bipartiteness is fully uncovered!



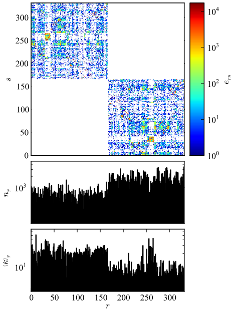
Detects meaningful features:

- Temporal
- Spatial (Country)
- Type/Genre

T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

Example: The Internet Movie Database (IMDB)

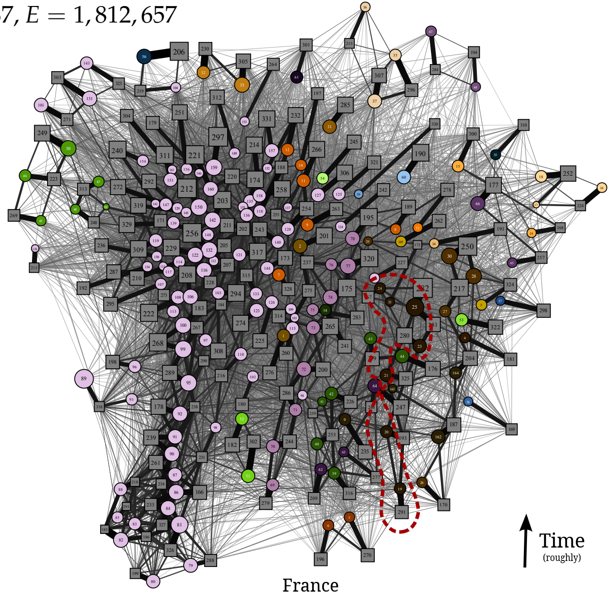
Bipartite network of actors and films.
Fairly large: $N = 372,787$, $E = 1,812,657$
MDL selects: $B = 332$



Bipartiteness is fully uncovered!

Detects meaningful features:

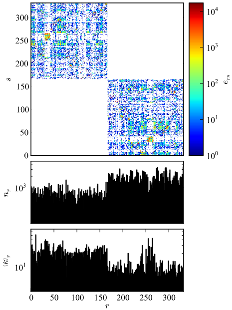
- Temporal
- Spatial (Country)
- Type/Genre



T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

Example: The Internet Movie Database (IMDB)

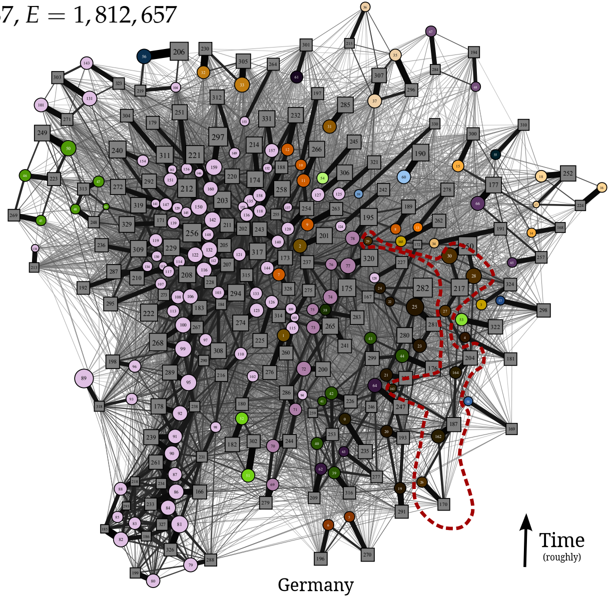
Bipartite network of actors and films.
Fairly large: $N = 372,787$, $E = 1,812,657$
MDL selects: $B = 332$



Bipartiteness is fully uncovered!

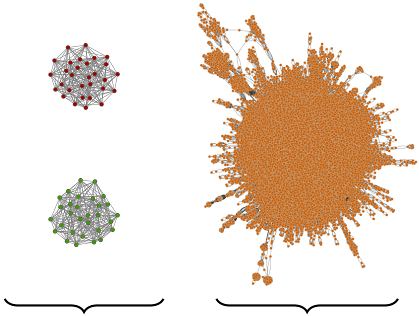
Detects meaningful features:

- Temporal
- Spatial (Country)
- Type/Genre



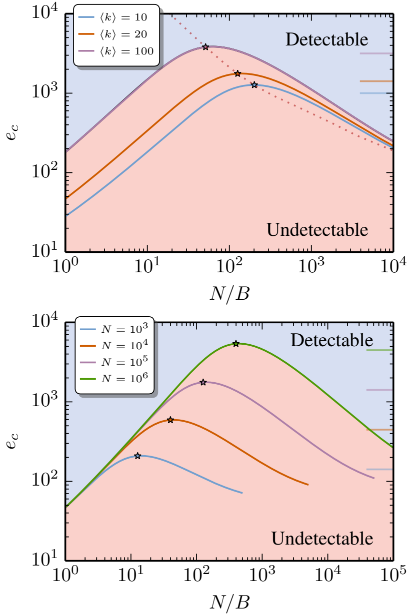
T.P.P., Phys. Rev. Lett. 110, 148701 (2013)

Problem: Resolution limit !?



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

Similar to modularity!
What is going on?

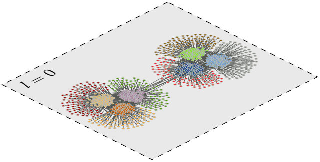


Resolution limit: Lack of prior information

Assumption that all block structures (block graphs) occur with the same probability.

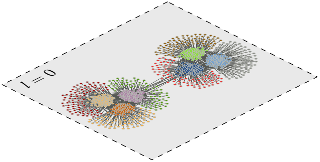
$$\mathcal{L} \sim B^2 \ln E + N \ln B$$

Lack of prior information: Solution → model the model

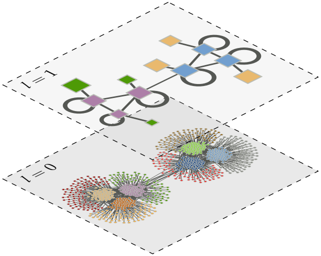


T.P.P., arXiv:1310.4377

Lack of prior information: Solution → model the model

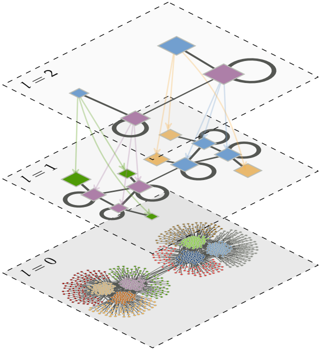


Lack of prior information: Solution → model the model



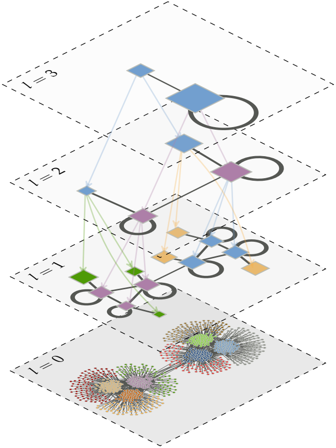
T.P.P., arXiv:1310.4377

Lack of prior information: Solution → model the model



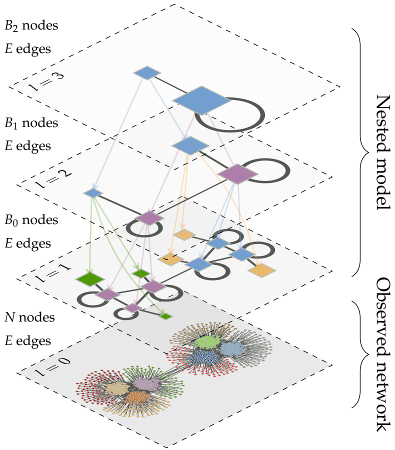
T.P.P., arXiv:1310.4377

Lack of prior information: Solution → model the model



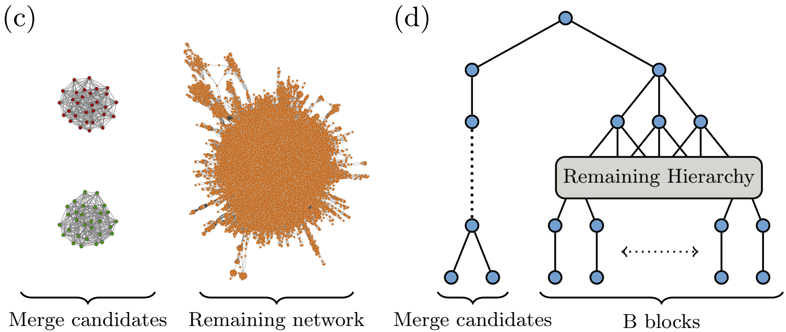
T.P.P., arXiv:1310.4377

LACK OF PRIOR INFORMATION: SOLUTION → MODEL THE MODEL



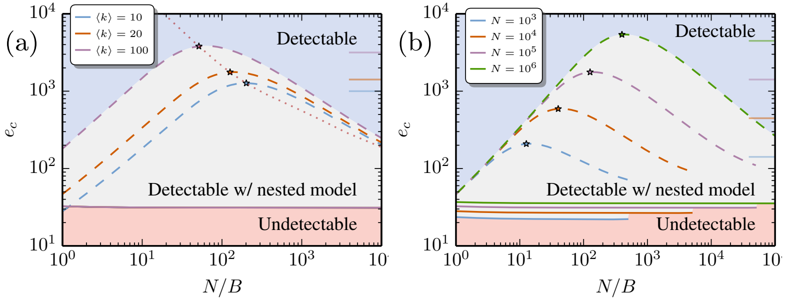
T.P.P., arXiv:1310.4377

Hierarchical model: Increased resolution



$$e_c^* \approx [\ln(B+N) - \ln n_c] / \ln 2$$

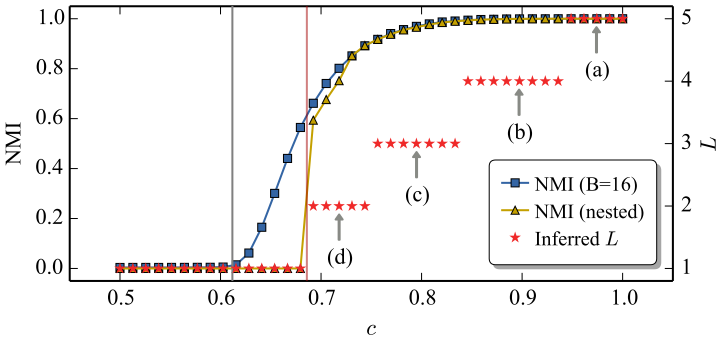
Hierarchical model: Increased resolution

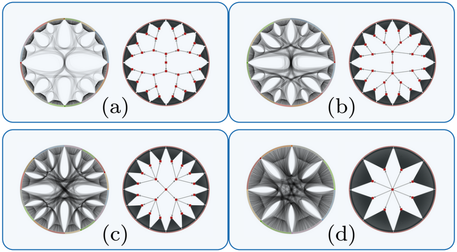


Hierarchical model: $N/B_{\max} \sim \ln N \qquad (\ll \sqrt{N})$

T.P.P., arXiv:1310.4377

Hierarchical model: Built-in validation

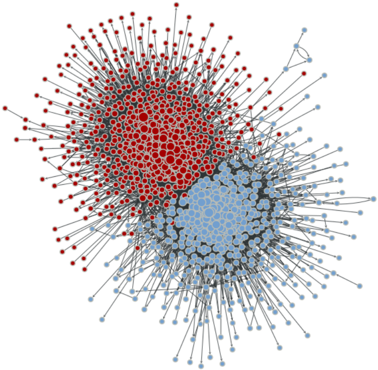




T.P.P., arXiv:1310.4377

EMPIRICAL NETWORKS

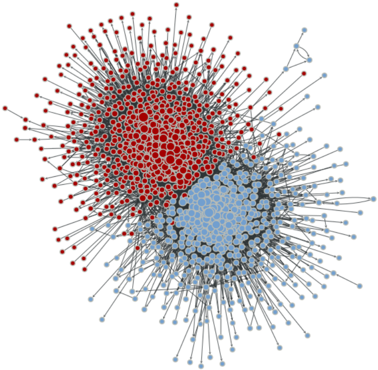
POLITICAL BLOGS ($N = 1,222, E = 19,027$)

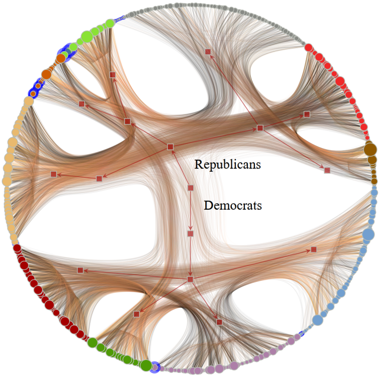


T.P.P., arXiv:1310.4377

EMPIRICAL NETWORKS

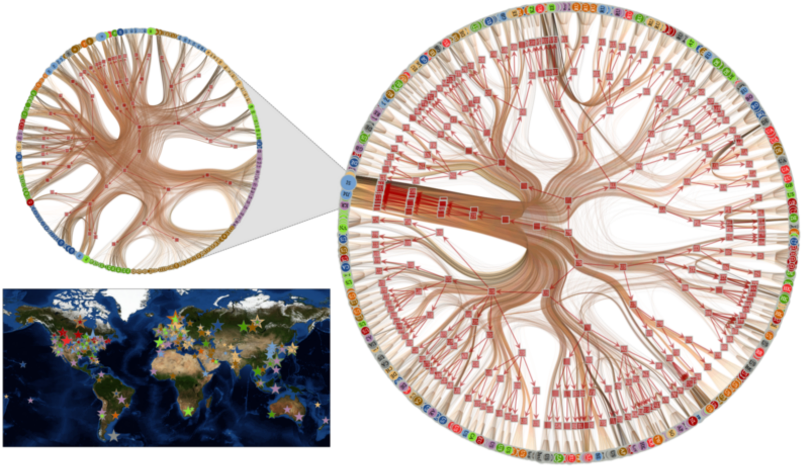
POLITICAL BLOGS ($N = 1,222, E = 19,027$)





EMPIRICAL NETWORKS

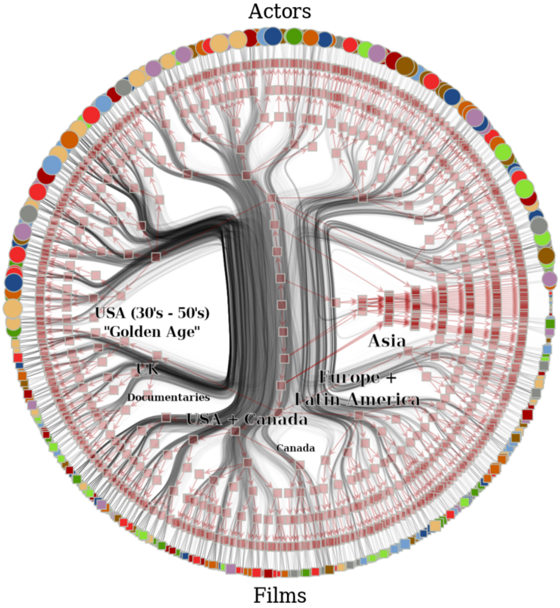
INTERNET (AUTONOMOUS SYSTEMS) ($N = 52,104, E = 399,625$)



T.P.P., arXiv:1310.4377

Empirical networks

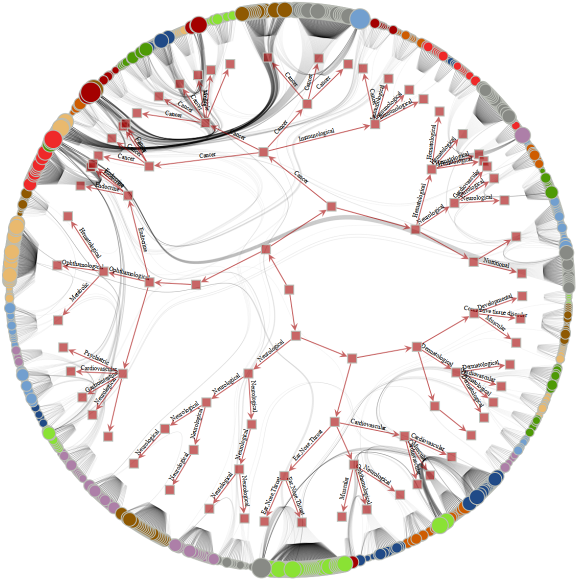
IMDB Film-Actor Network ($N = 372,447, E = 1,812,312, B = 717$)



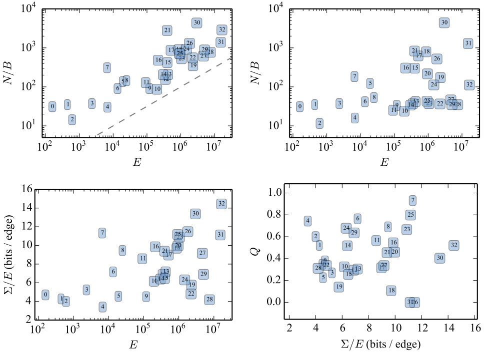
T.P.P., arXiv:1310.4377

Empirical networks

Human Disease Genes ($N = 903, E = 6{,}760$)



EMPIRICAL NETWORKS



No.	N	E	Dir.	No.	N	E	Dir.	No.	N	E	Dir.
0	62	159	No	11	21,363	91,286	No	22	255,265	2,234,572	Yes
1	105	441	No	12	27,400	352,504	Yes	23	317,080	1,049,866	No
2	115	613	No	13	34,401	421,441	Yes	24	325,729	1,469,679	Yes
3	297	2,345	Yes	14	39,796	301,498	Yes	25	334,863	925,872	No
4	903	6,760	No	15	52,104	399,625	Yes	26	372,547	1,812,312	No
5	1,222	19,021	Yes	16	56,739	212,945	No	27	449,087	4,690,321	Yes
6	4,158	13,422	No	17	75,877	508,836	Yes	28	654,782	7,499,425	Yes
7	4,941	6,594	No	18	82,168	870,161	Yes	29	855,802	5,066,842	Yes
8	8,638	24,806	No	19	105,628	2,299,623	No	30	1,134,890	2,987,624	No
9	11,204	117,619	No	20	196,591	950,327	No	31	1,637,868	15,205,016	No
10	17,903	196,972	No	21	224,832	394,400	Yes	32	3,764,117	16,511,740	Yes

No.	Network	No.	Network	No.	Network
0	Dolphins	11	arXiv Co-Authors (cond-mat)	22	Web graph of `stanford.edu`.
1	Political Books	12	arXiv Citations (hep-th)	23	DBLP collaboration
2	American Football	13	arXiv Citations (hep-ph)	24	WWW
3	C. Elegans Neurons	14	PGP	25	Amazon product network
4	Disease Genes	15	Internet AS (Caida)	26	IMDB film-actor (bipartite)
5	Political Blogs	16	Brightkite social network	27	APS citations
6	arXiv Co-Authors (gr-qc)	17	Epinions.com trust network	28	Berkeley/Stanford web graph
7	Power Grid	18	Slashdot	29	Google web graph
8	arXiv Co-Authors (hep-th)	19	Flickr	30	Youtube social network
9	arXiv Co-Authors (hep-ph)	20	Gowalla social network	31	Yahoo groups (bipartite)
10	arXiv Co-Authors (astro-ph)	21	EU email	32	US patent citations

Conclusion

- Statistical inference → more principled
- Stochastic blockmodels → simple, tractable, **general block structure!**
- Minimum description length (MDL) → non-parametric, elegant, principled
- MDL → **Built-in validation!**
- Nested-stochastic blockmodel → block structure at multiple scales
- Nested MDL → **Vanishing resolution limit!**
- Overall very efficient → meaningful results from very large data sets.

graph-tool

Very fast, freely available C++ code as part of the graph-tool Python library!
`http://graph-tool.skewed.de`