

39. Methodenseminar: Big Data Module II



Introduction to Social Network Science with Python

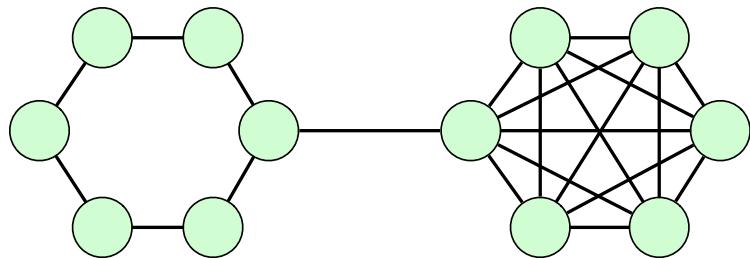
Meso-Scale Analysis

Marcos Oliveira

July 16, 2019

Meso-Scale Structures in Networks

- Real-world networks are heterogeneous.

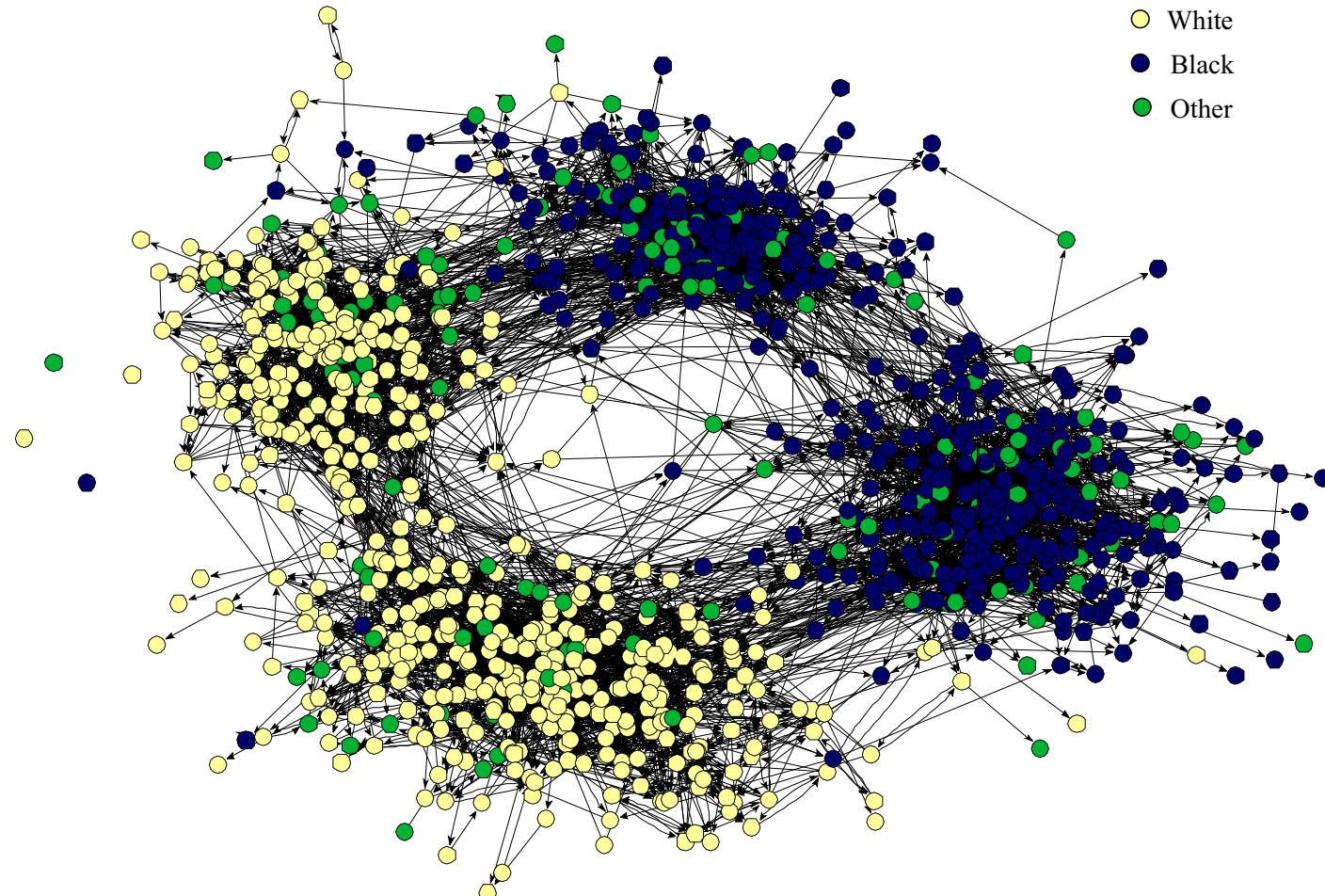


$$C = 0.82$$

- We want a simplified model of something very complicated.
 - Networks are often too large and complex to be analyzed with some kind of simplification.
- Some uses
 - Extrapolation
 - Interpolation
 - Generalization
 - Mechanisms
 - Explanations
 - Useful division
 - Simplifications

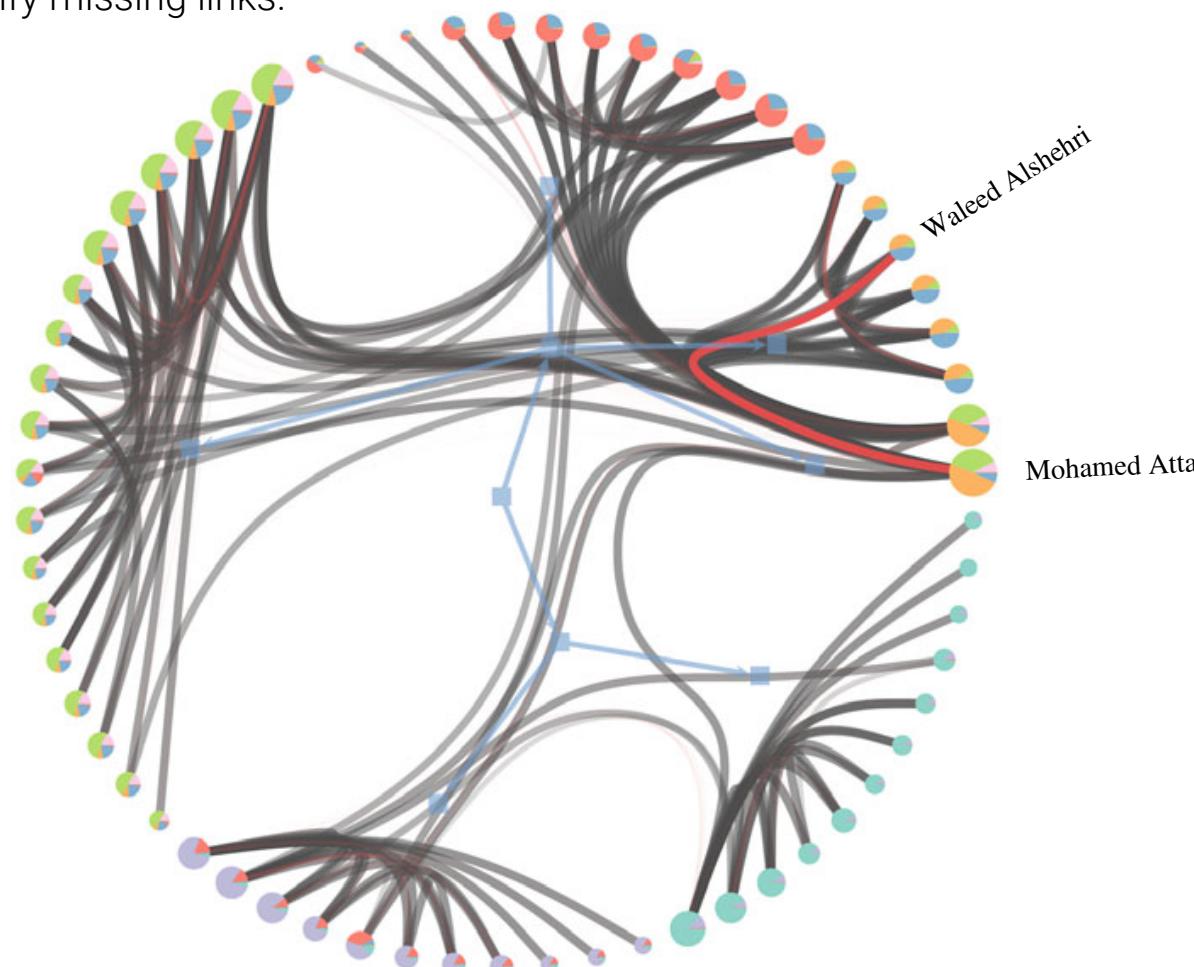
Meso-Scale Structures in Networks

- Why?
 - Extrapolation
 - Make prediction for unseen nodes in the network.



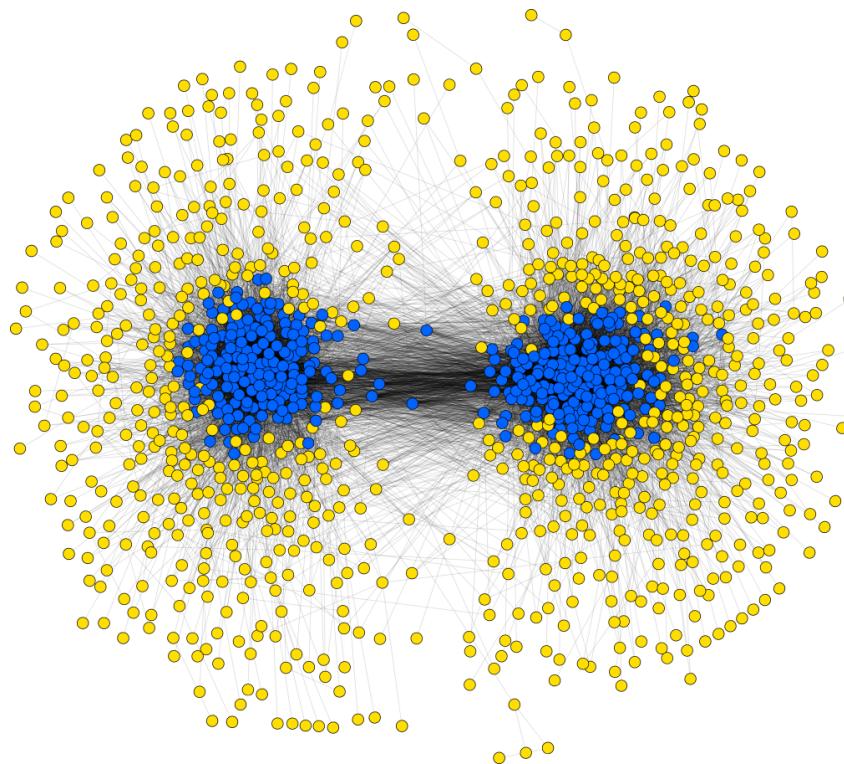
Meso-Scale Structures in Networks

- Why?
 - Interpolation
 - Identify missing links.



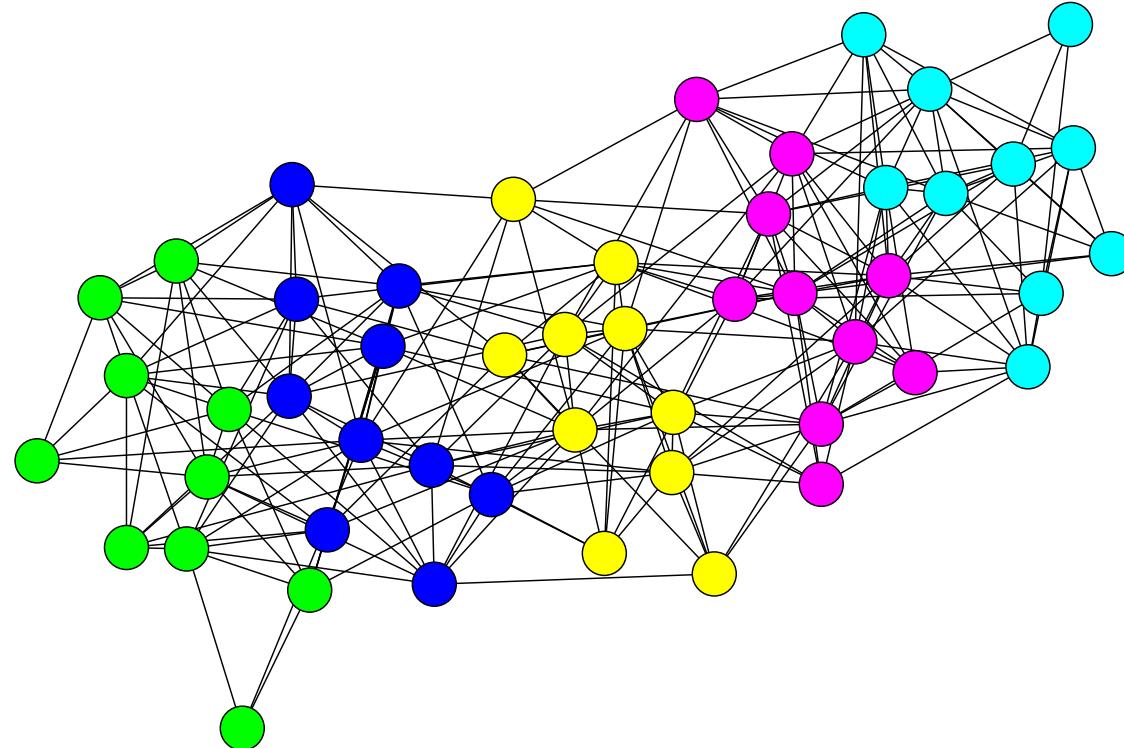
Meso-Scale Structures in Networks

- Why?
 - Generalization
 - Nodes of this type are like the others of the same type.



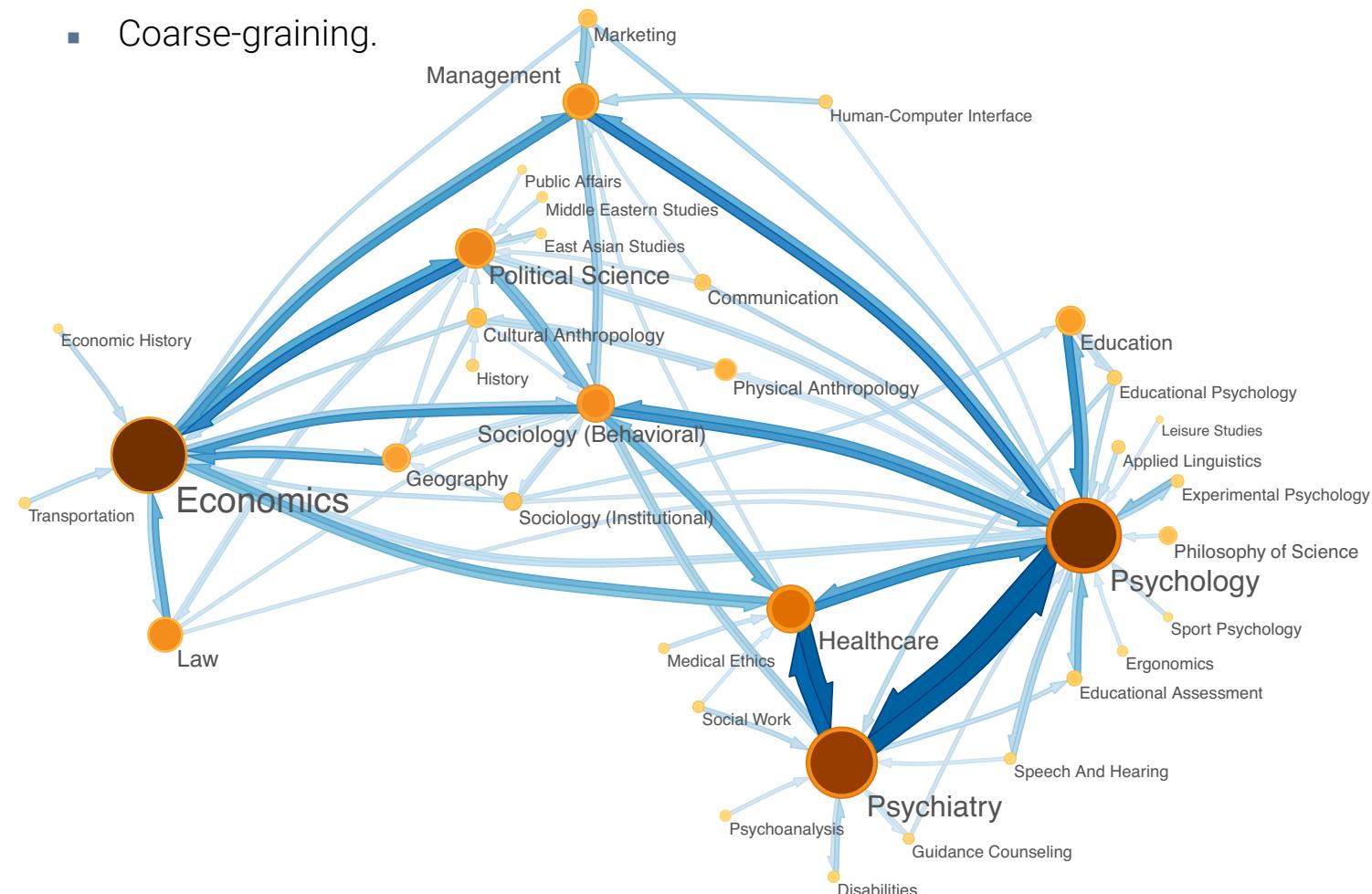
Meso-Scale Structures in Networks

- Why?
 - Mechanisms
 - How did this network arise? What are the underlying mechanisms?



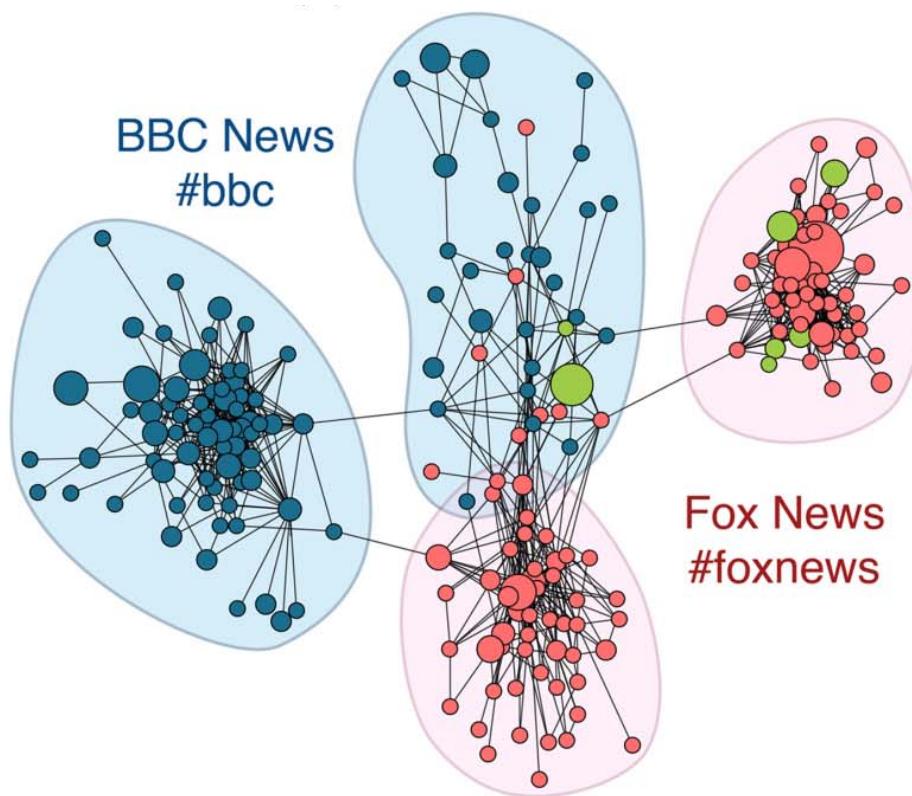
Meso-Scale Structures in Networks

- Why?
 - Explanations
 - Coarse-graining.



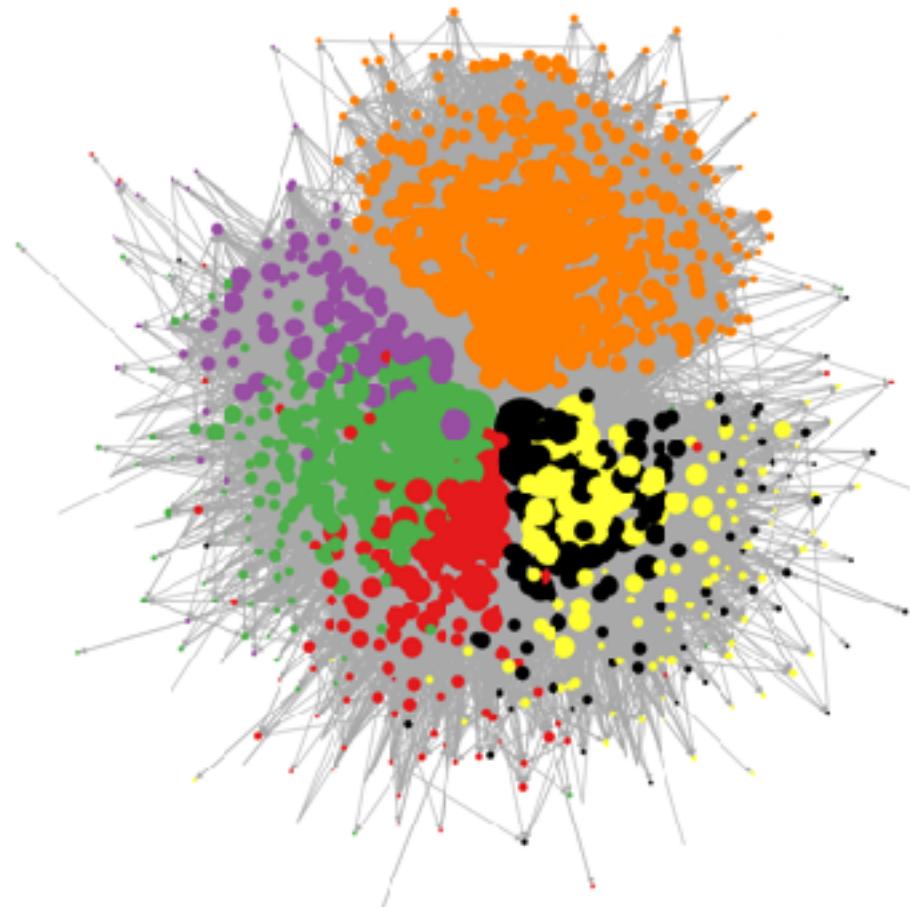
Meso-Scale Structures in Networks

- Why?
 - Useful division
 - Finding groups to assign treatments.



Meso-Scale Structures in Networks

- Why?
 - Simplification
 - Regression models need ranks or groups.



Meso-Scale Structures: Communities

Meso-Scale Structures: Communities

- Communities

- In network science, we call a community a group of nodes that have a higher likelihood of connecting to each other than to nodes from other communities.



- Homophily or assortative mixing: “like links with like.”

- The so-called “hypotheses of community detection”:
 - H1: Fundamental hypothesis
 - A network’s community structure is uniquely encoded in its wiring diagram.
 - H2: Connectedness and density hypothesis
 - A community is a locally dense connected subgraph in a network.
 - H3: Random hypothesis
 - Randomly wired networks lack an inherent community structure.
 - H4: Maximal modularity hypothesis

- Modularity
 - How much more often do attributes match across edges than expected at random?

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - p_{ij}) \delta(x_i, x_j)$$

$\delta(x_i, x_j) = 1$, if $x_i = x_j$
 $\delta(x_i, x_j) = 0$, otherwise

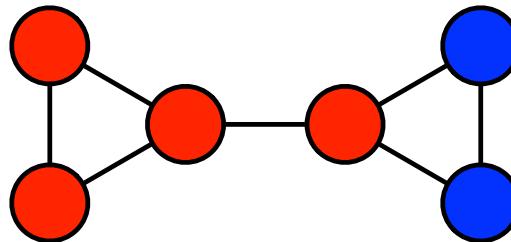
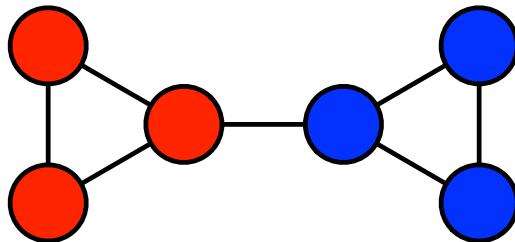
- A good null model should account for the degree of the nodes:

$$p_{ij} = \frac{k_i k_j}{2m}$$

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(x_i, x_j)$$

Meso-Scale Structures: Communities

- Modularity
 - An example:

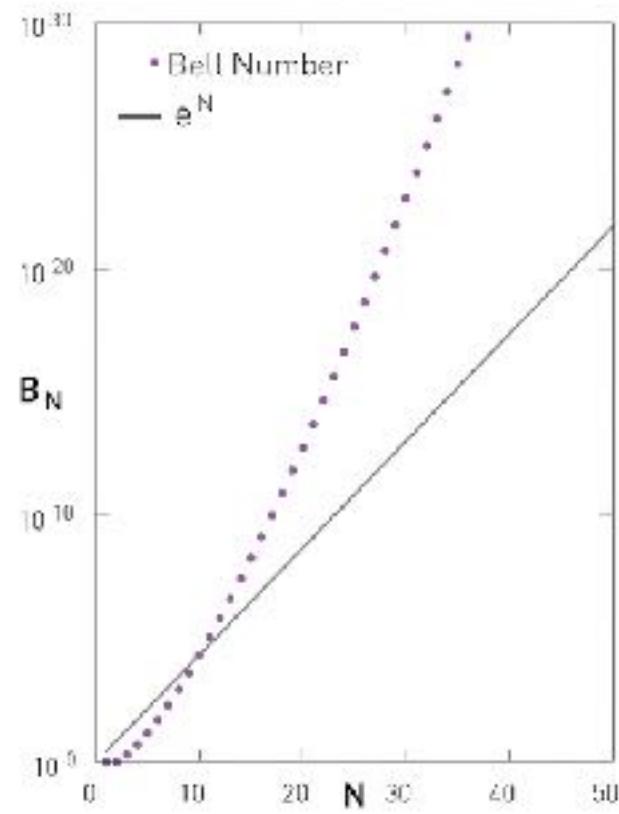


- Resolution limit:
 - Modularity has an artifact in that small communities might be merged into a large one.
 - We will see this during the demo.

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(x_i, x_j)$$

- The hypotheses of community detection:
 - H1: Fundamental hypothesis
 - A network's community structure is uniquely encoded in its wiring diagram.
 - H2: Connectedness and density hypothesis
 - A community is a locally dense connected subgraph in a network.
 - H3: Random hypothesis
 - Randomly wired networks lack an inherent community structure.
 - H4: Maximal modularity hypothesis
 - For a given network the partition with maximum modularity corresponds to the optimal community structure.

- Community detection algorithms
 - Clauset-Newman-Moore: greedy modularity
 - The Louvain method
 - Girvan-Newman: a centrality-based method
 - Label propagation



- Community detection algorithms: greedy modularity.
 - General idea:
 - Start with every vertex being in its own group, then recursively and greedily choose a pair of groups to merge based on modularity.
 - Algorithm:
 1. Start with every node being its own group.
 2. Choose the pair of groups that their merge would produce the greatest increase (or smallest decrease) in the modularity.
 3. Repeat step 2. until there is only one group remaining.
 - Advantage: simple to implement and understand.
 - Disadvantage: not the fastest.

Meso-Scale Structures: Communities

- Community detection algorithms: the Louvain method.
 - Algorithm:

1. Start with every node being its own group.

Phase 1:

2. For each node **i**

2.1. Evaluate the gain in modularity if node **i** joins the community of its neighbors **j**.

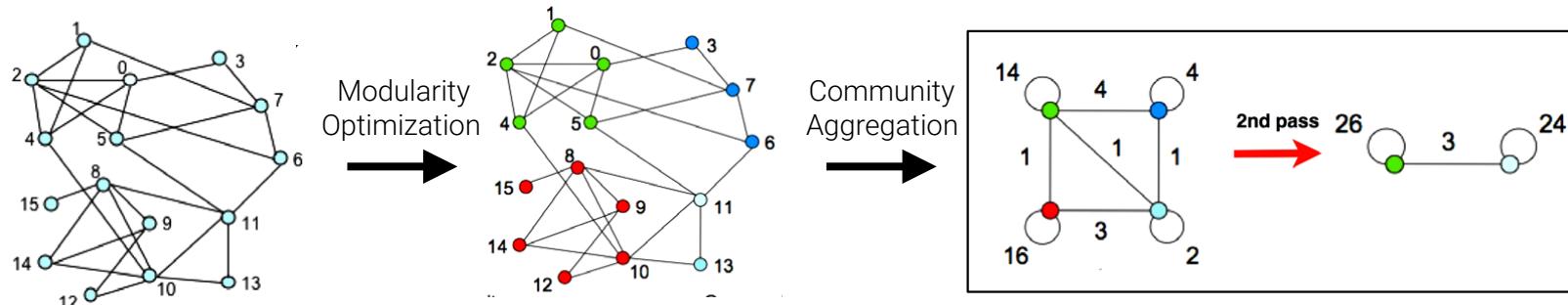
2.2. The node **i** is then placed in the community of maximum gain.

3. Repeat step 2. until no individual move can improve the modularity.

Phase 2:

4. Create a network from the communities found in Phase 1.

5. Use this new network as input for the Phase 1 until there are no more changes.



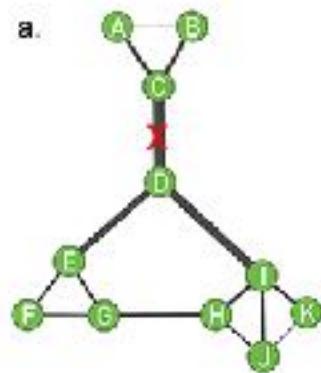
- Community detection algorithms: the Louvain method.
 - Advantage: quite fast.

	Karate	Arxiv	Internet	Web nd.edu	Phone	Web uk-2005	Web WebBase 2001
Nodes/links	34/77	9k/24k	70k/351k	325k/1M	2.04M/5.4M	39M/783M	118M/1B
	0.38/0 s	0.772/3.6 s	0.692/799 s	0.927/5034 s	—/—	—/—	—/—
	0.42/0 s	0.813/0 s	0.781/1 s	0.935/3 s	0.76/44 s	0.979/738 s	0.984/152 mn

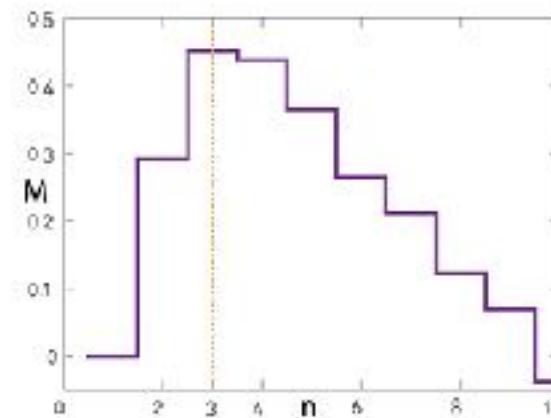
Greedy
Louvain

Meso-Scale Structures: Communities

- Community detection algorithms: a centrality-based method.
 - General idea: remove links concerning nodes that belong to different communities, eventually breaking a network into communities.



...



- Community detection algorithms: label propagation.
 - General idea: labels are propagated across the network; the algorithm converges when each node has the majority label of its neighborhood.
 - Algorithm:
 1. Start with every node with a unique label.
 2. For each node **i**
 - 2.1.** Set its label to the label occurring with the highest frequency among its neighbors.
 3. Repeat step **2.** until every node has a label that the maximum number of neighbors have.

- Other measures of quality

- Coverage

- The ratio of *intra-community* edges by the total number of edges.
 - When all the communities are disconnected from each other, coverage is equal to 1.

$$\frac{\text{\# intra-community edges}}{\text{total number of edges}}$$

- Performance

- The ratio between the number of connections inside the communities plus the non-edges between communities and the total number of potential edges
 - It counts the number of correctly "interpreted" pair of vertices.

$$\frac{\text{\# intra-community edges} + \text{\# inter-community non-edges}}{\text{total possible number of edges}}$$

- Being critical about community detection
 - Do we really have communities?
 - Community detection algorithms will find communities, regardless they exist or not.
 - Hypotheses of community detection?
 - We cannot prove the correctness; how to turn them into theorems?
 - What does modularity really mean?
 - High modularity means the presence of significant deviations from the null model.
 - It is all about assortative mixing.
- Other meso-scale structures might explain the data better.
 - We will see this in the SBM part of the course.

Next: Community Detection (Demo)

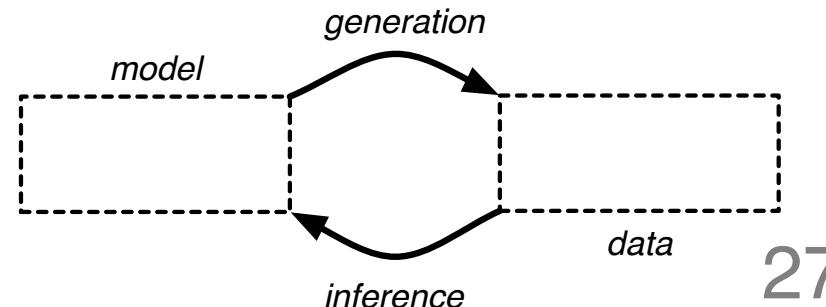


Leibniz Institute
for the Social Sciences



The Stochastic Block Model

- Block models
 - Divide the network into a set of nodes where all nodes in the same block had the same patterns of connection to nodes in other blocks.
- The Stochastic Block Model (SBM)
 - They are random graph ensembles in which:
 - **vertices** are separated into groups (or ‘blocks’),
 - and the probability of an **edge** existing between two vertices is determined according to their group membership.
 - The *standard* stochastic block model assumes that all vertices belonging to the same block are statistically indistinguishable, they all have the same .
- A generative model approach:



- The Stochastic Block Model definition:

N Number of vertices

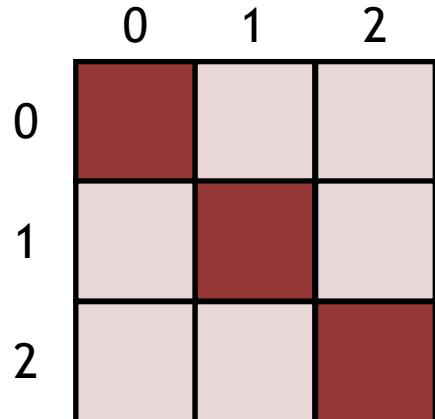
B Number of blocks

n_r Number of vertices in block $r \in [0, B - 1]$

e_{rs} The number of edges between blocks r and s (micro-canonical)

- SBM generation

$$N = 99 \quad B = 3 \quad n_r = 33$$



Meso-Scale Structures in Networks: the SBM

- Block matrix

	0	1	2
0	dark red	light pink	light pink
1	light pink	dark red	light pink
2	light pink	light pink	dark red

	0	1	2
0	light pink	dark red	dark red
1	dark red	light pink	dark red
2	dark red	dark red	light pink

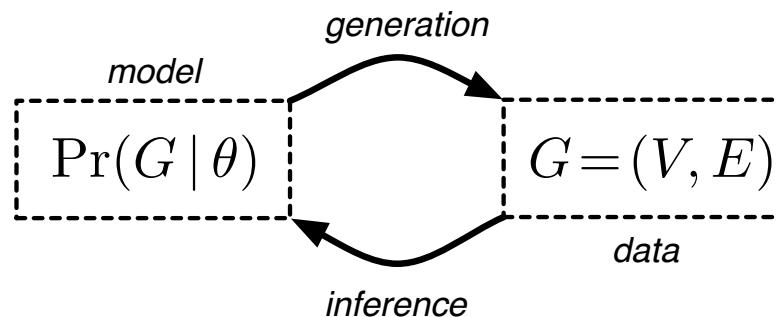
	0	1	2
0	dark red	dark red	light pink
1	dark red	dark red	light pink
2	light pink	light pink	dark red

	0	1	2
0	dark red	dark red	light pink
1	dark red	dark red	dark red
2	light pink	dark red	dark red

We will play with these during the demo.

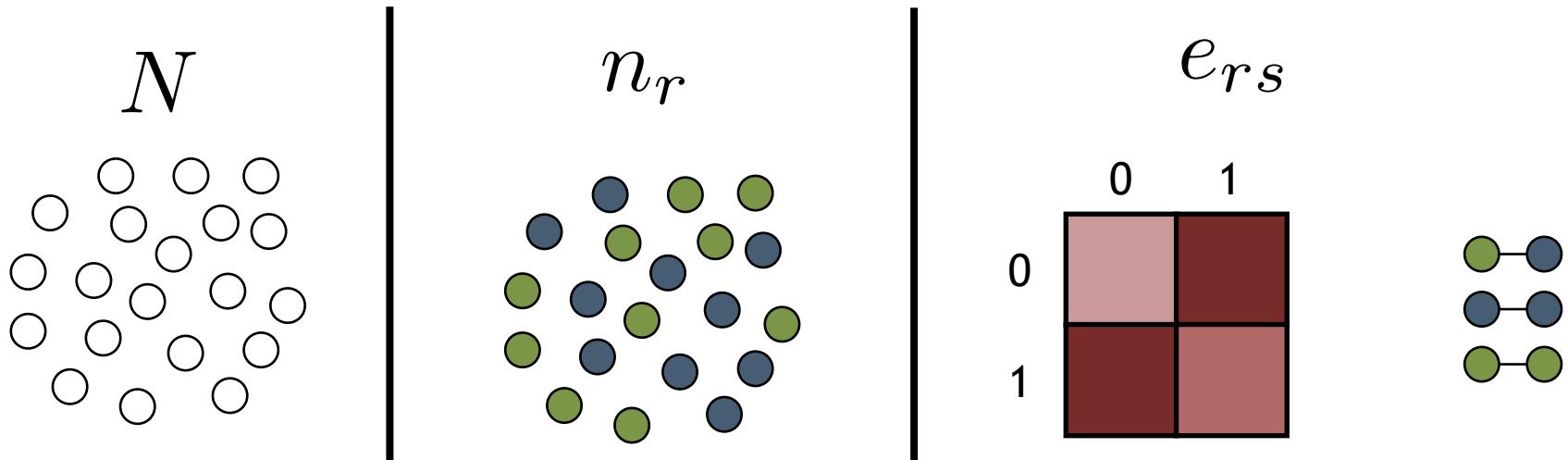
- SBM Inference

- Given that we know B



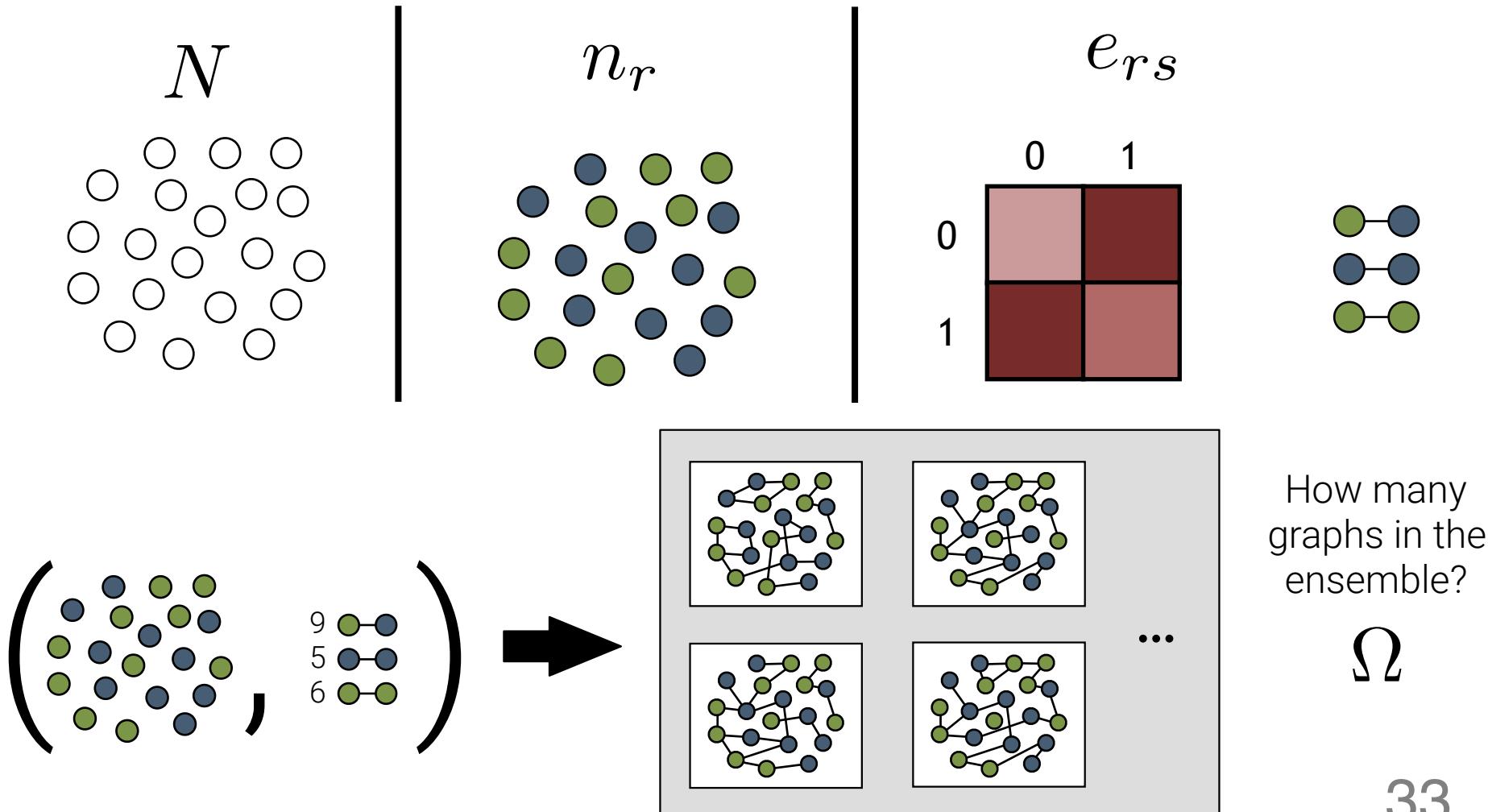
Meso-Scale Structures in Networks: the SBM

- SBM Inference
 - Generating data for a specific value of B



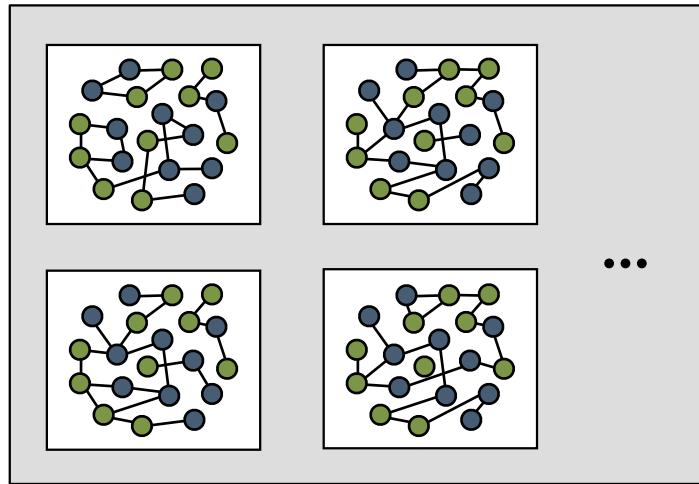
Meso-Scale Structures in Networks: the SBM

- SBM Inference
 - Generating data for a specific value of B



- SBM Inference
 - The entropy of an SBM

$$N, n_r, e_{rs}$$



Ω Number of graphs in the ensemble

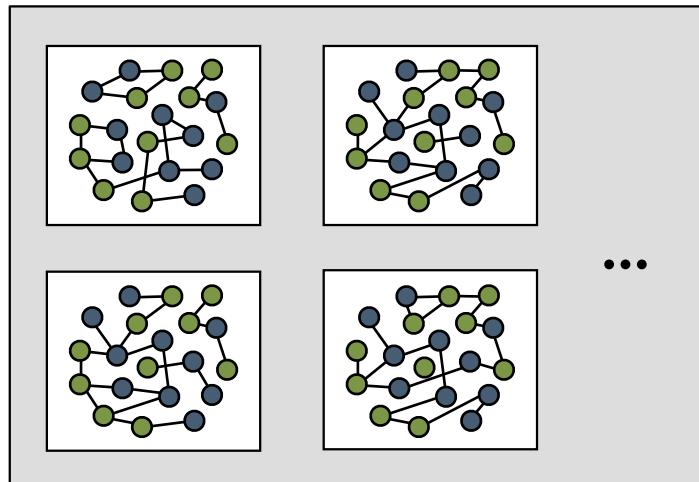
The entropy is given as: $S = \ln \Omega$

Entropy measures the degree of “order” of an ensemble. Higher entropy = more disorder.

Meso-Scale Structures in Networks: the SBM

- SBM Inference
 - The entropy of an SBM

$$N, n_r, e_{rs}$$



- Deriving a log-likelihood function:

If we assume that each graph has the same probability:

$$\mathcal{P} = 1/\Omega$$

Ω Number of graphs in the ensemble

The entropy is given as: $\mathcal{S} = \ln \Omega$

Entropy measures the degree of “order” of an ensemble. Higher entropy = more disorder.

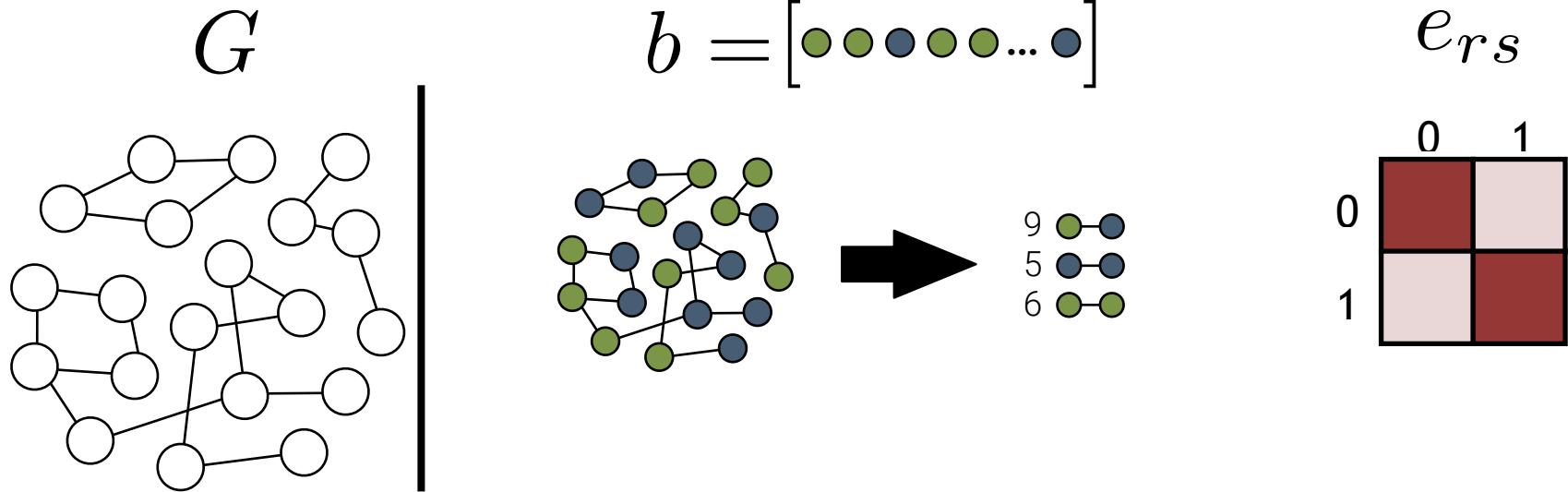
We can define the log-likelihood function as:

$$L = \ln \mathcal{P}$$

$$L = -\mathcal{S}$$

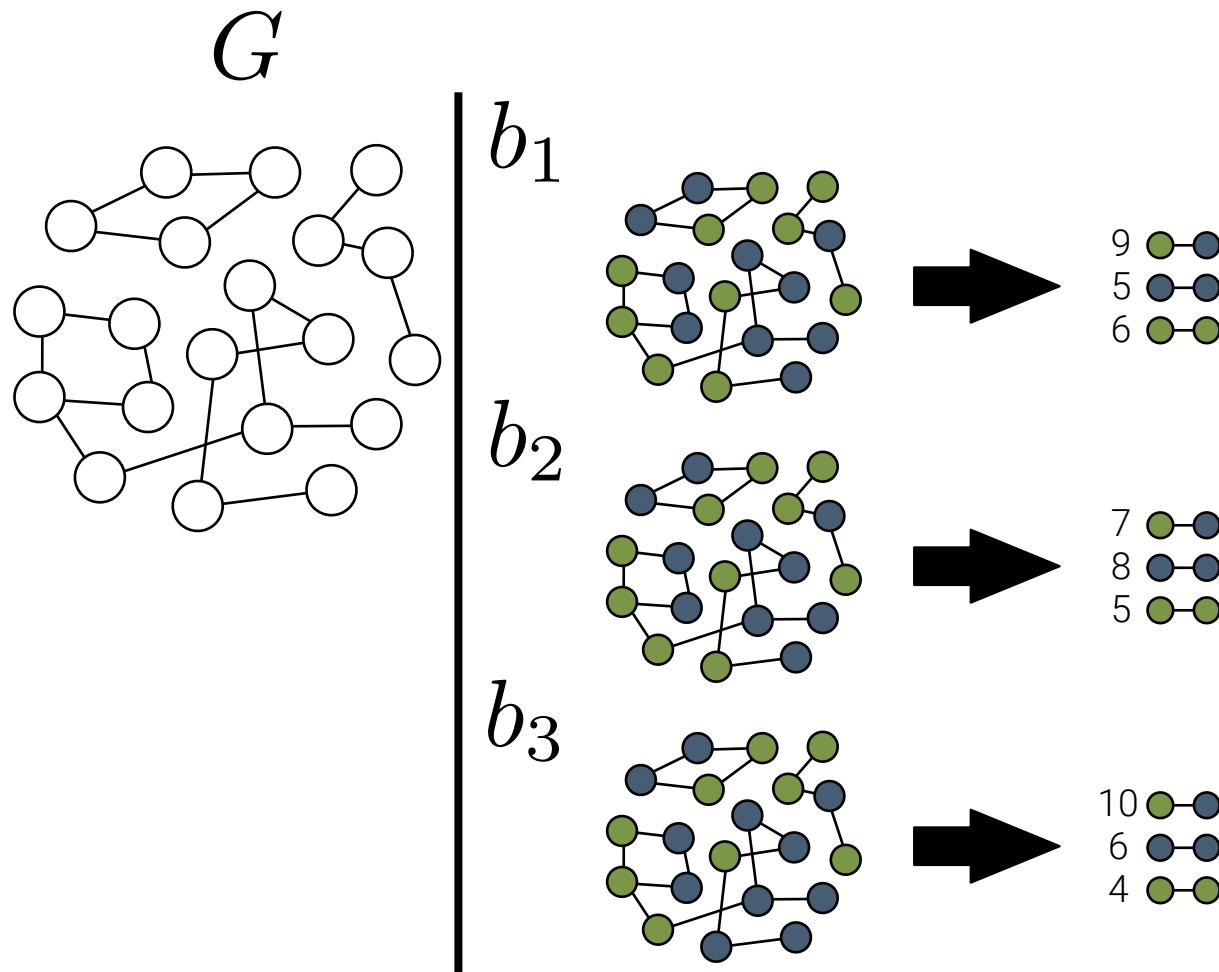
Meso-Scale Structures in Networks: the SBM

- SBM Inference



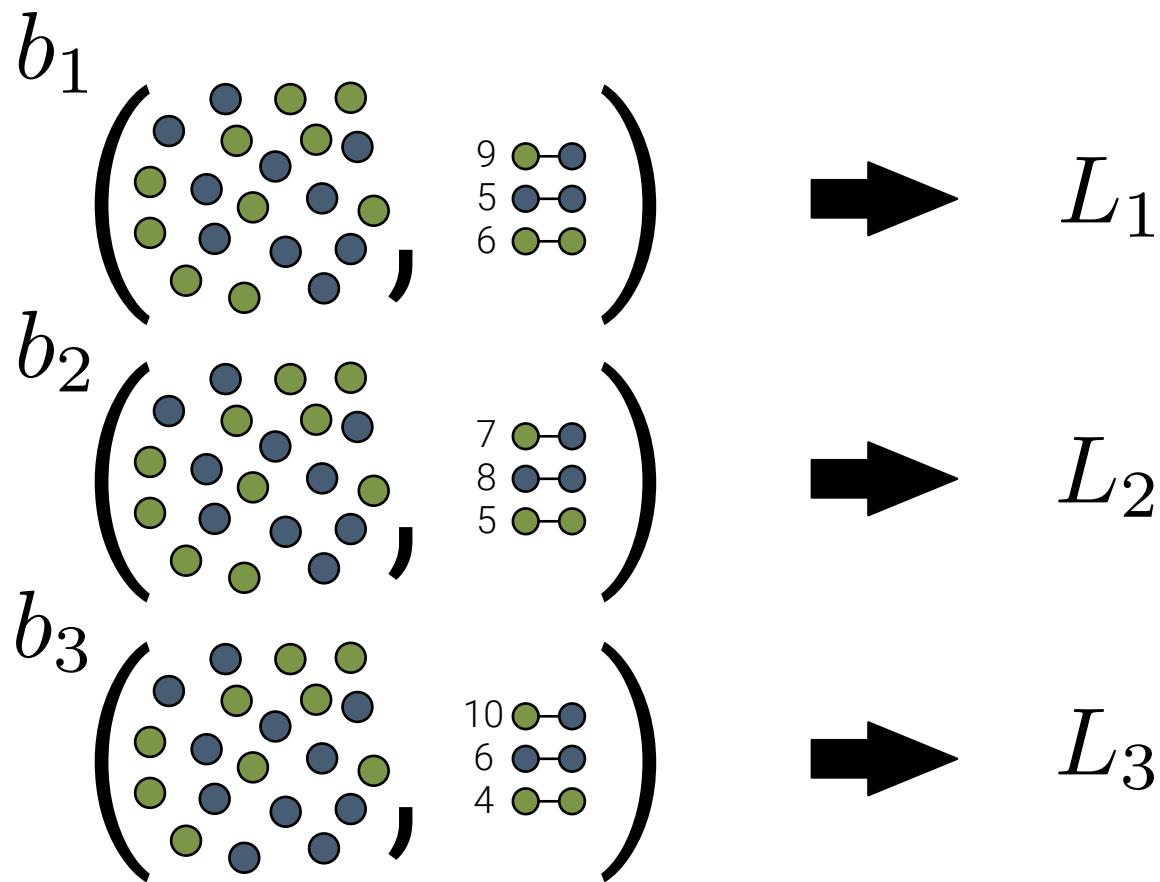
Meso-Scale Structures in Networks: the SBM

- SBM Inference



Meso-Scale Structures in Networks: the SBM

■ SBM Inference



- SBM Inference
 - Maximize the likelihood L (or equivalently, minimize S)
 - A simple “greedy” algorithm:
 1. Start with a random partition of b
 2. For each vertex in the graph:
 - 2.1. Change its block membership to the value that maximizes the likelihood.
 3. Repeat step 2. until no further improvement is possible.
 - We need many runs to find the optimal solution.

- SBM Inference
 - Most of the time we are unaware of B
 - As B increases, we start to memorize data.
 - The “best model” is when $B = N$ (a.k.a. the worst model).

- SBM Inference
 - The Description Length:
 - The information-theoretic definition: “the amount of bits required to send the compressed message plus the number of bits in the encoding scheme”
 - The Description Length of the Stochastic Block Model:
 - The entropy of data given the model plus the entropy of the model.
$$\Sigma = \mathcal{S} + \mathcal{L}$$
 - It is basically a way to penalize the complexity of the model.
 - To find blocks, we follow the Minimum Description Length principle:
 - The best choice of model which fits given data is the one which most compresses it; the one that minimizes the total amount of information required to describe it.

- SBM Inference when we are unaware of B
 - Minimizing the description length \sum
 - A straightforward algorithm:
 1. Start with $B = 0$.
 2. Increase the number of blocks B .
 3. Fit the SBM with B blocks and check the description length of it.
 4. Repeat step 2. until the description length stop decreasing.
 - The fitting part (i.e., step 3.) is based on Metropolis-Hastings:
 - Move across the “partition space” by proposing to move a node from a group into another group.
 - For this, we use local-level information about the neighbors of the node.

- Degree-corrected Stochastic Block Model (DC-SBM)
 - The standard stochastic block model assumes that all vertices belonging to the same block are statistically indistinguishable, they all have the same.
 - Many observed networks, however, show a broad range of degrees.

Next: Stochastic Block Model (Demo)



Leibniz Institute
for the Social Sciences

