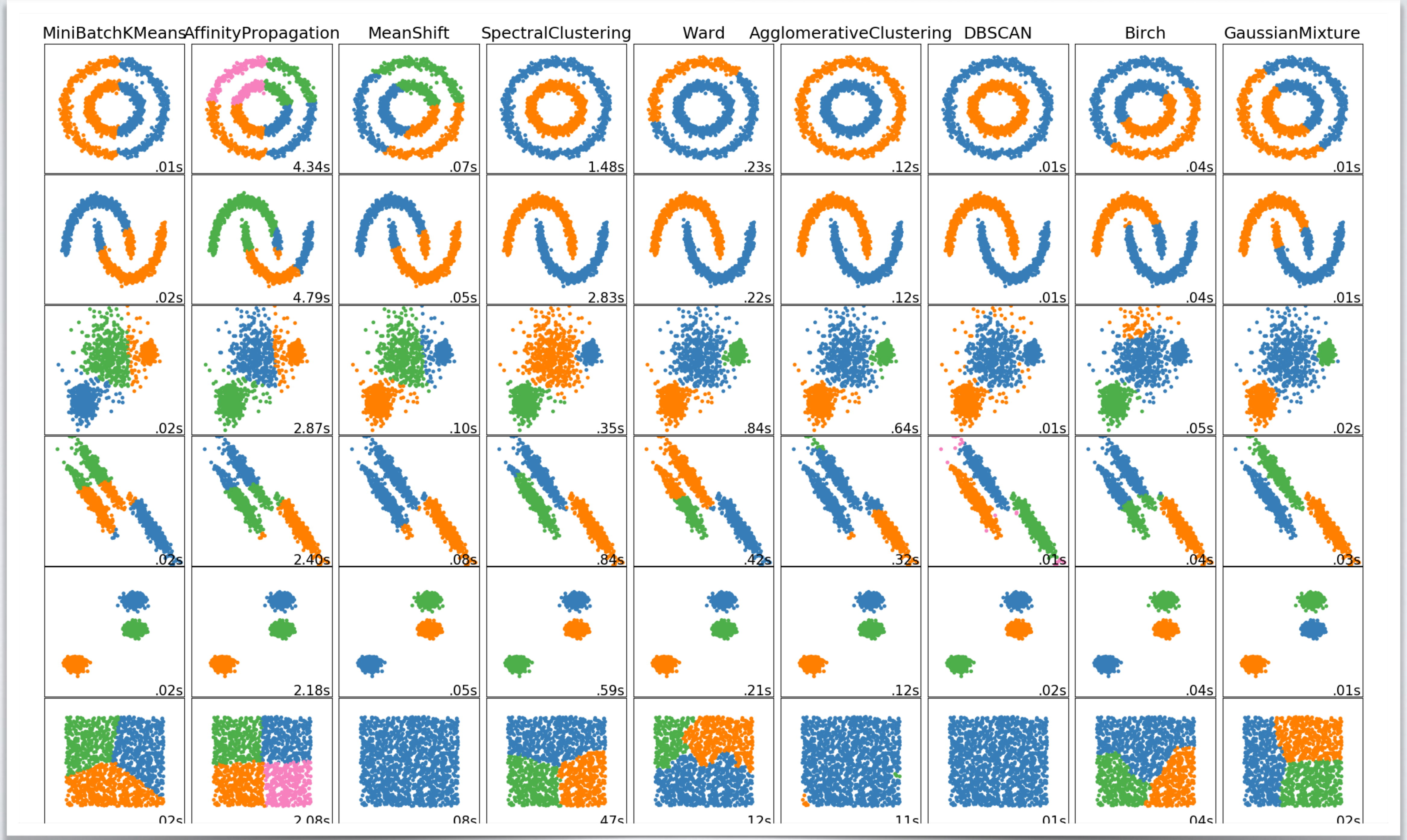
COMMUNITY DETECTION (GRAPH CLUSTERING)

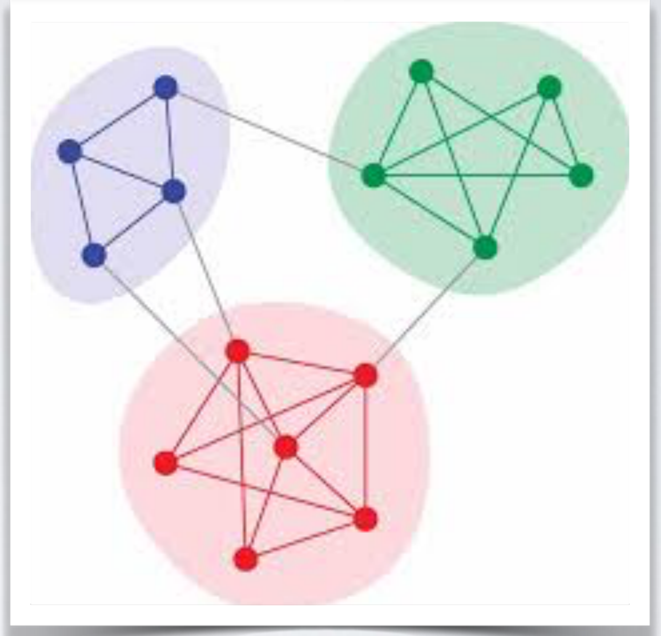
COMMUNITY DETECTION

- Community detection is equivalent to “clustering” in unstructured data
- Clustering: unsupervised machine learning
  - Find groups of elements that are similar to each other
    - People based on DNA, apartments based on characteristics, etc.
  - Hundreds of methods published since 1950 (k-means)
  - Problem: what does “similar to each other” means ?

COMMUNITY DETECTION



COMMUNITY DETECTION



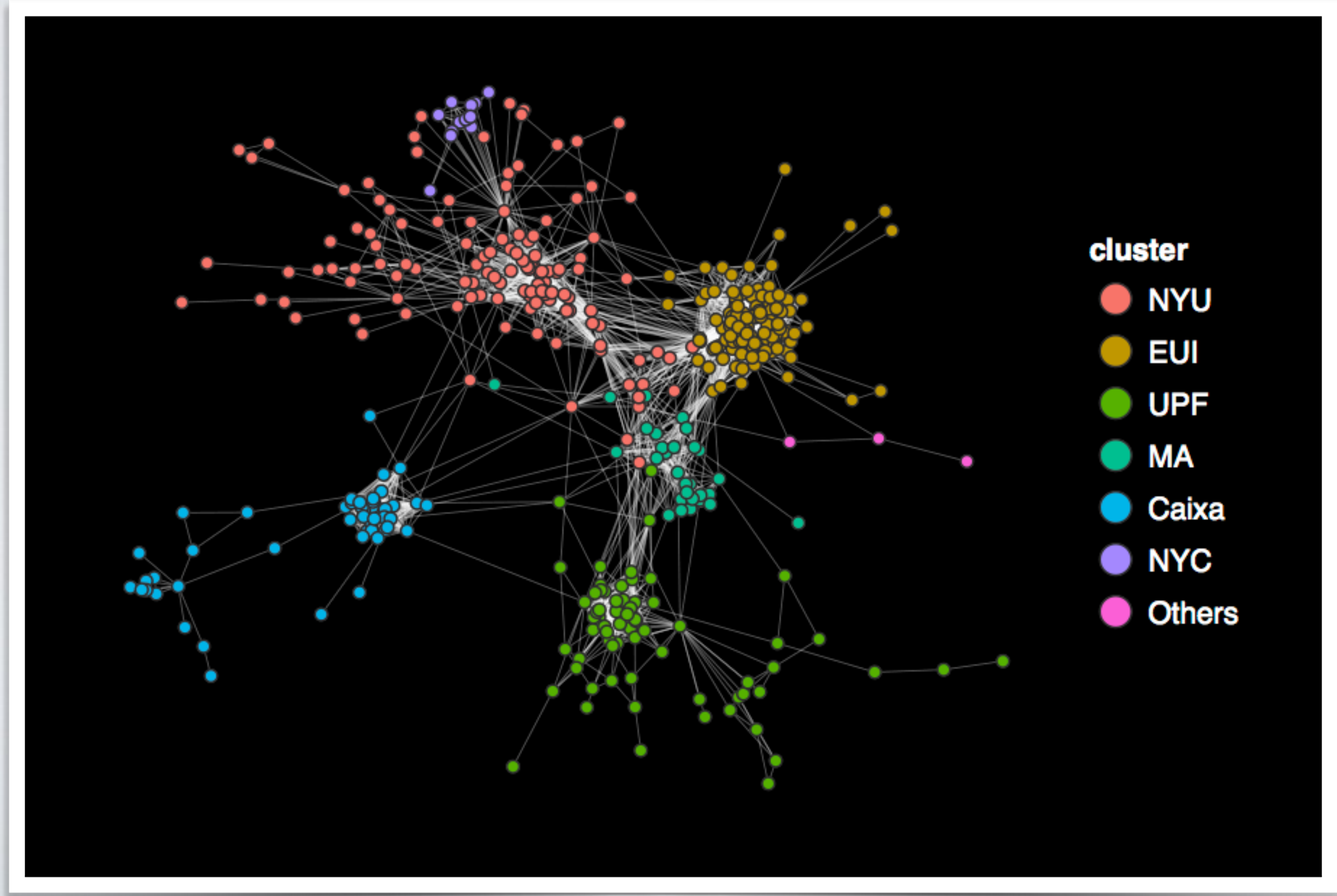
- Community detection:
  - Find groups of nodes that are:
    - Strongly connected to each other
    - Weakly connected to the rest of the network
    - Ideal form: each community is 1)A clique, 2) A separate connected component
  - No formal definition
  - Hundreds of methods published since 2003

WHY COMMUNITY DETECTION ?

- One of the key properties of complex networks was
  - High clustering coefficient
  - (friends of my friends are my friends)
- Different from random networks. How to explain it ?
  - Watts strogatz (spatial structure?)
- => In real networks, presence of dense groups: communities
  - Small, dense (random) networks have high density.
  - Large networks could be interpreted as aggregation of smaller, denser networks, with much fewer edges between them

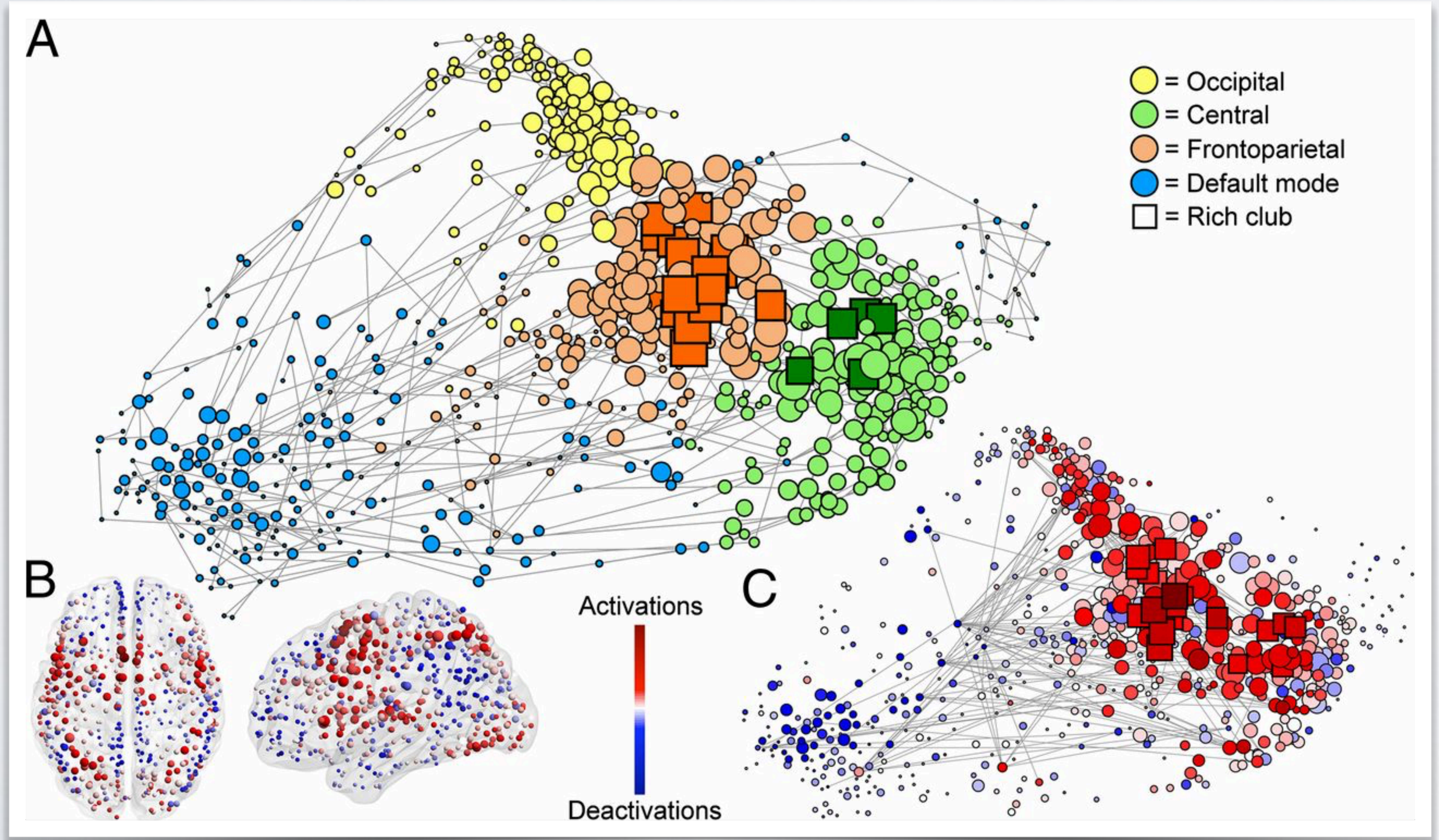
COMMUNITY STRUCTURE IN REAL GRAPHS

- If you plot the graph of your Facebook/linked-in contacts, it looks like this



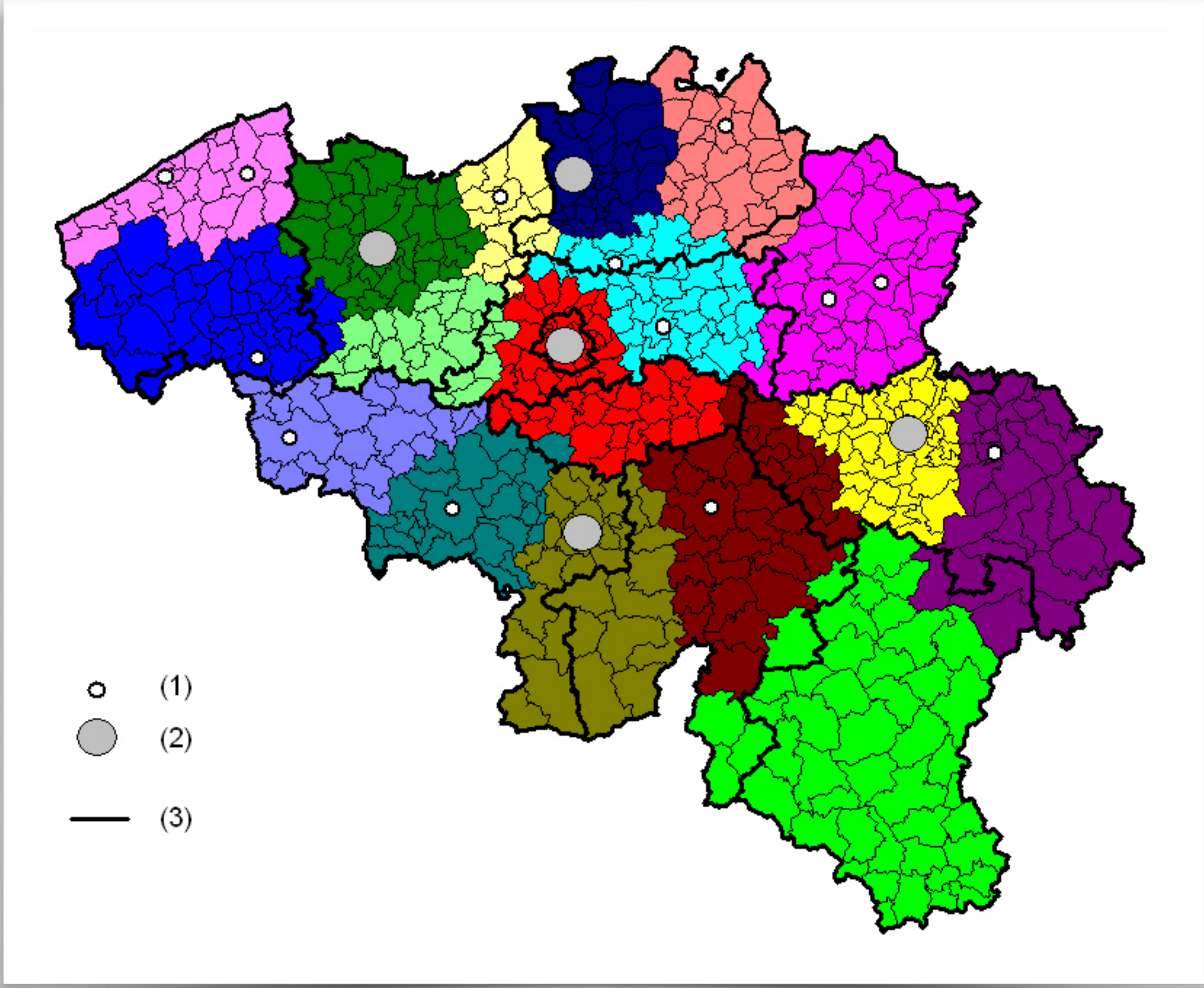
COMMUNITY STRUCTURE IN REAL GRAPHS

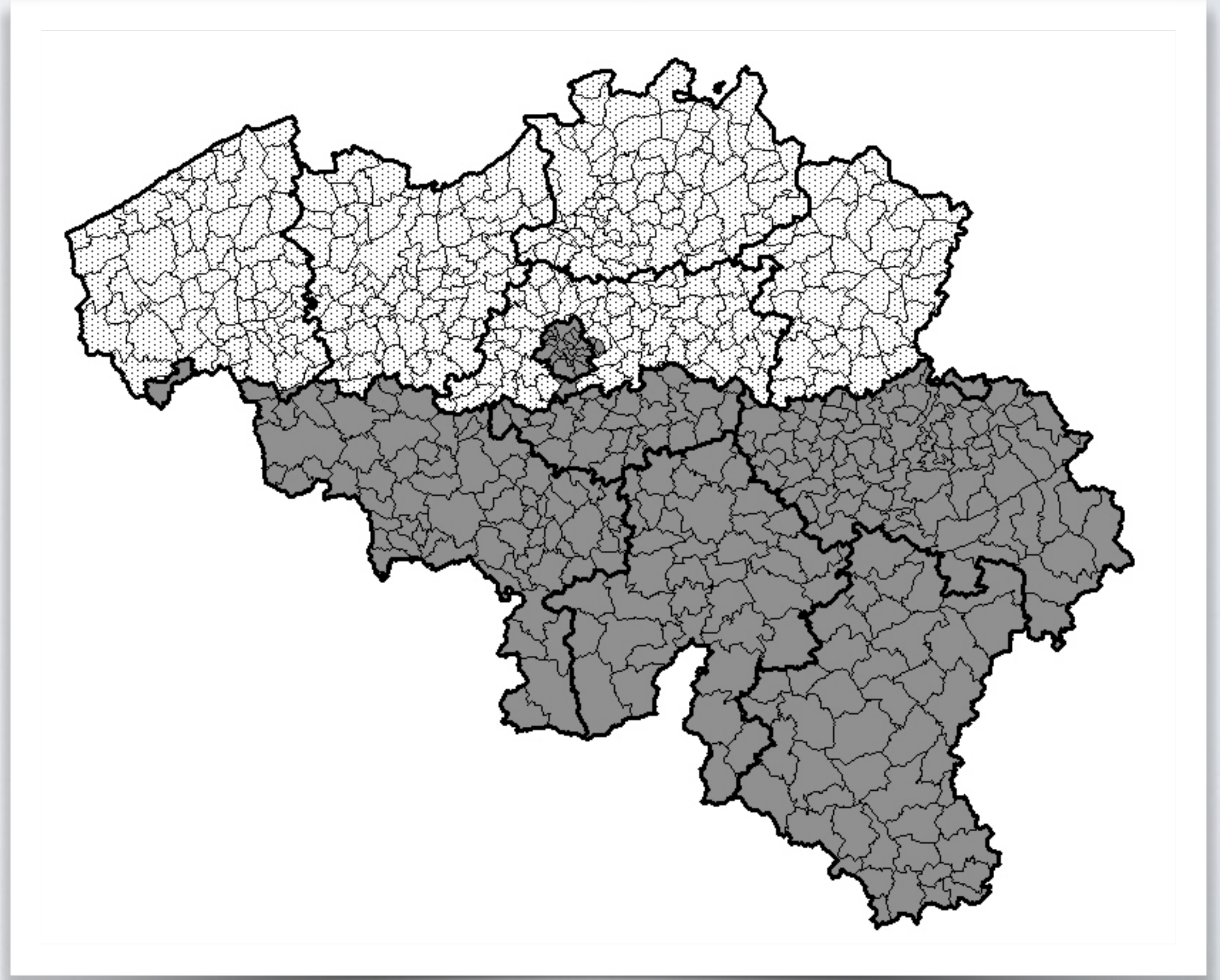
- Connections in the brain ?



COMMUNITY STRUCTURE IN REAL GRAPHS

- Phone call communications in Belgium ?

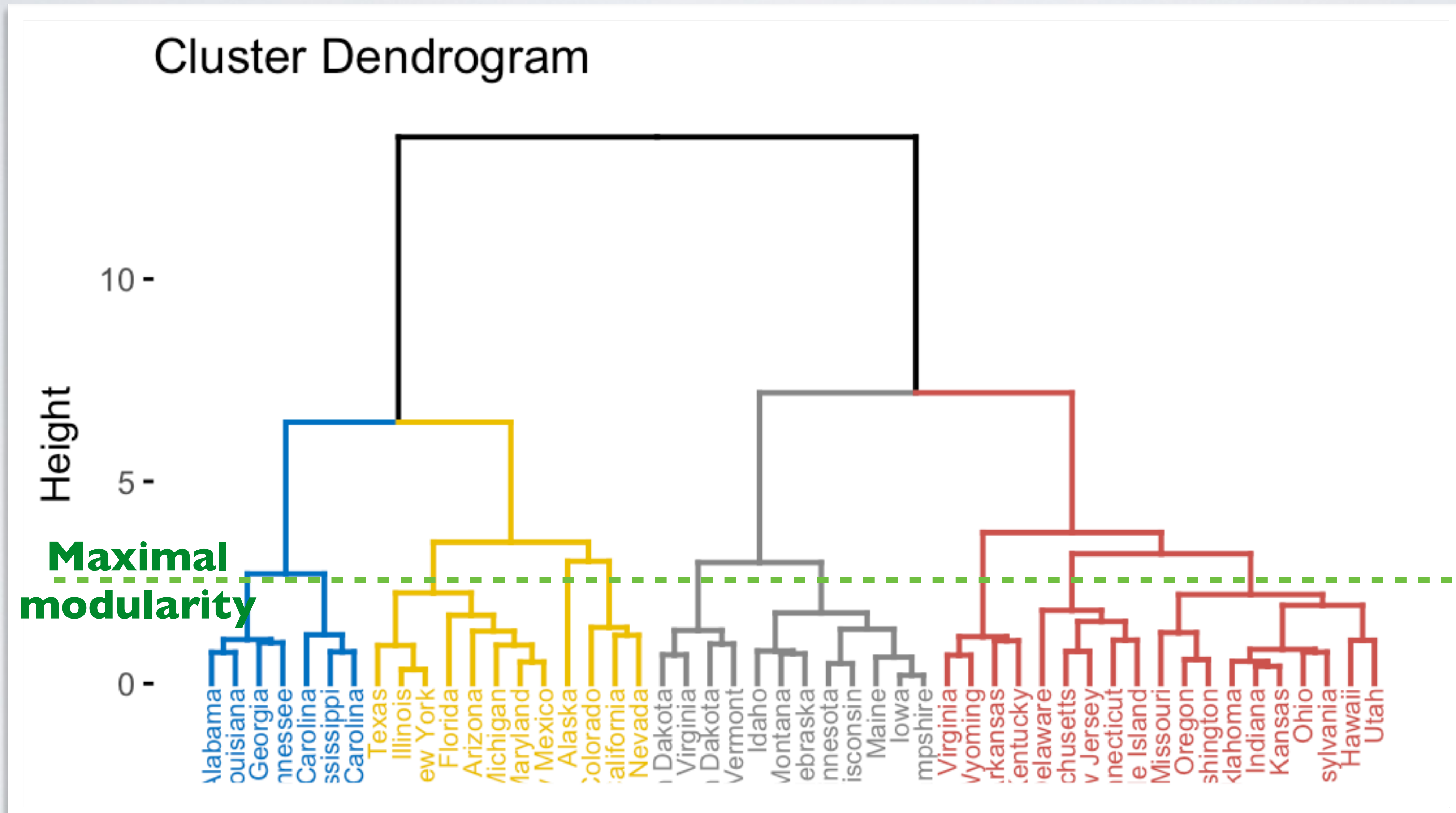




FIRST METHOD BY GIRVAN & NEWMAN

- 1)Compute the betweenness of all edges
- 2)Remove the edge of highest betweenness
- 3)Repeat until all edges have been removed
  - Connected components are communities
- => It is called a *divisive* method
- =>What you obtain is a dendrogram
- How to cut this dendrogram at the *best* level ?

FIRST METHOD BY GIRVAN & NEWMAN



FIRST METHOD BY GIRVAN & NEWMAN

- Introduction of the **Modularity**
- The modularity is computed for a partition of a graph
  - (each node belongs to one and only one community)
- It compares :
  - The **observed** *fraction of edges inside communities*
  - To the **expected** *fraction of edges inside communities* in a random network

MODULARITY

$$Q = \frac{1}{(2m)} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{(2m)} \right] \delta(c_v, c_w)$$

Original formulation

MODULARITY

$$Q = \frac{1}{(2m)} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{(2m)} \right] \delta(c_v, c_w)$$

Sum over all pairs of nodes

MODULARITY

$$Q = \frac{1}{(2m)} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{(2m)} \right] \delta(c_v, c_w)$$

1 if in same community

MODULARITY

$$Q = \frac{1}{(2m)} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{(2m)} \right] \delta(c_v, c_w)$$

1 if there is an edge between them

MODULARITY

$$Q = \frac{1}{(2m)} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{(2m)} \right] \delta(c_v, c_w)$$

Probability of an edge in a configuration model

MODULARITY

- Modularity compares the observed network to a **null model**
  - Usually the configuration model
    - Multi-edges and loops are allowed
  - Other models could be used, such as ER random graphs.

- Natural extension to weighted/multi-edge networks

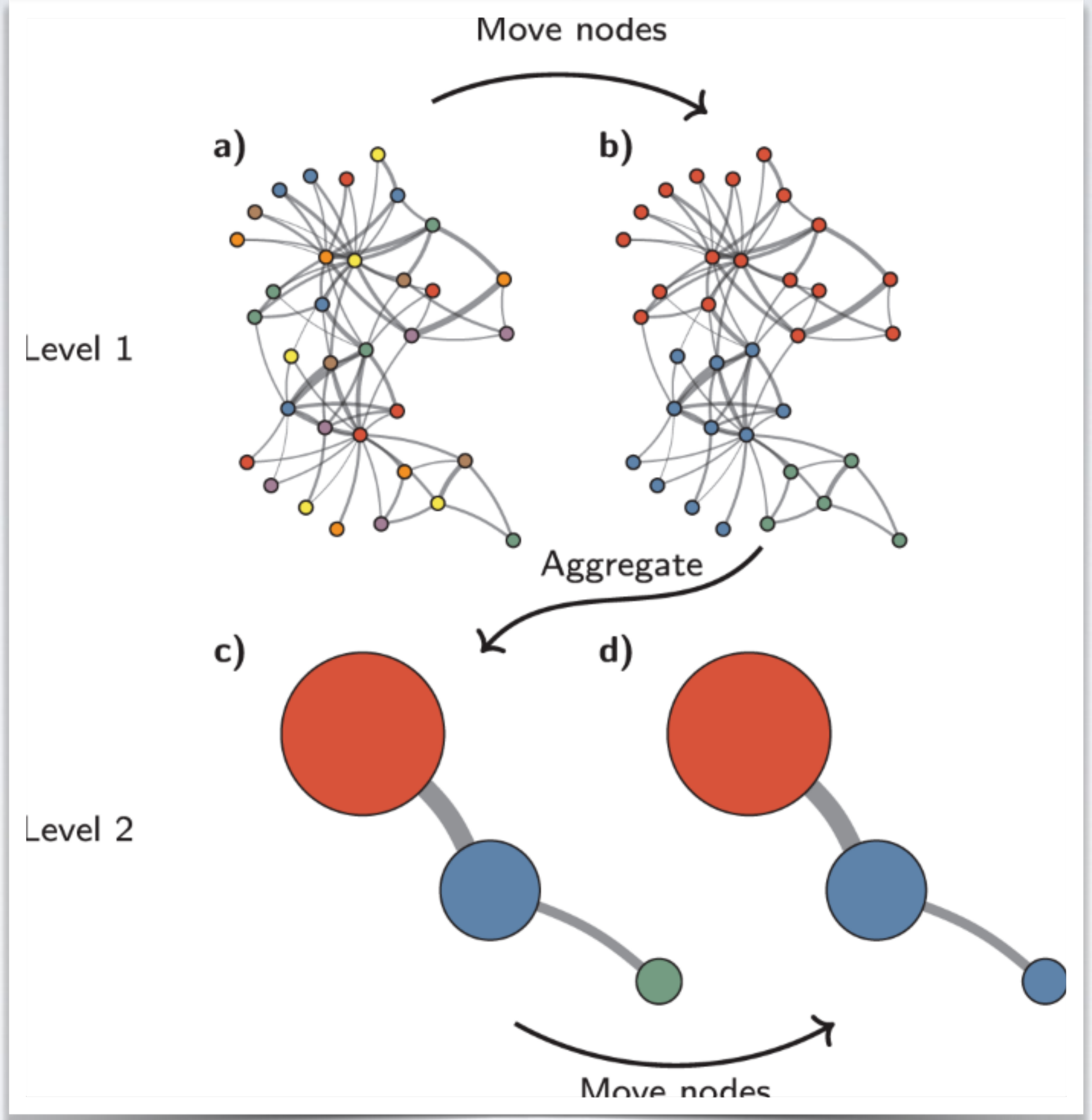
FIRST METHOD BY GIRVAN & NEWMAN

- Back to the method:
  - Create a dendrogram by removing edges
  - Cut the dendrogram at the best level using modularity
- =>In the end, your objective is… to optimize the Modularity, right ?
- Why not optimizing it directly !
  - But NP complete problem

LOUVAIN ALGORITHM

- Simple, greedy approach
  - Easy to implement
  - Fast
- Yields a hierarchical community structure
- Beat state of the art on all aspects (when introduced)
  - Speed
  - Max modularity obtained
  - Do not fall in some traps (see later)

LOUVAIN ALGORITHM



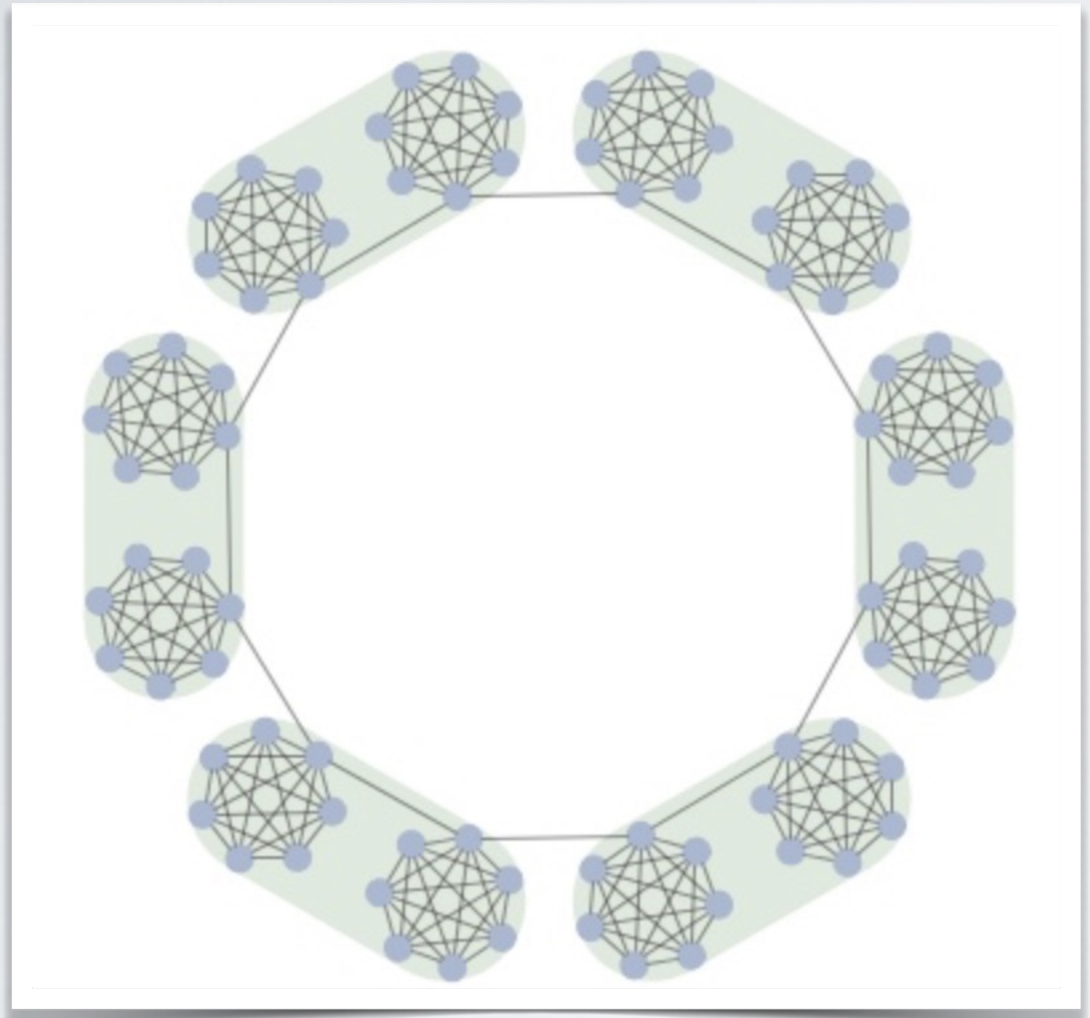
Blondel, Vincent D., et al. "Fast unfolding of communities in large networks." *Journal of statistical mechanics: theory and experiment* 2008.10 (2008): P10008.

RESOLUTION LIMIT

- Modularity == Definition of good communities ?
- 2006-2008: series of articles [Fortunato,Lancicchinetti,Barthelemy]
  - Resolution limit of Modularity
- Let's see an example

Fortunato, Santo, and Marc Barthelemy. "Resolution limit in community detection." *Proceedings of the national academy of sciences* 104.1 (2007): 36-41.

RESOLUTION LIMIT



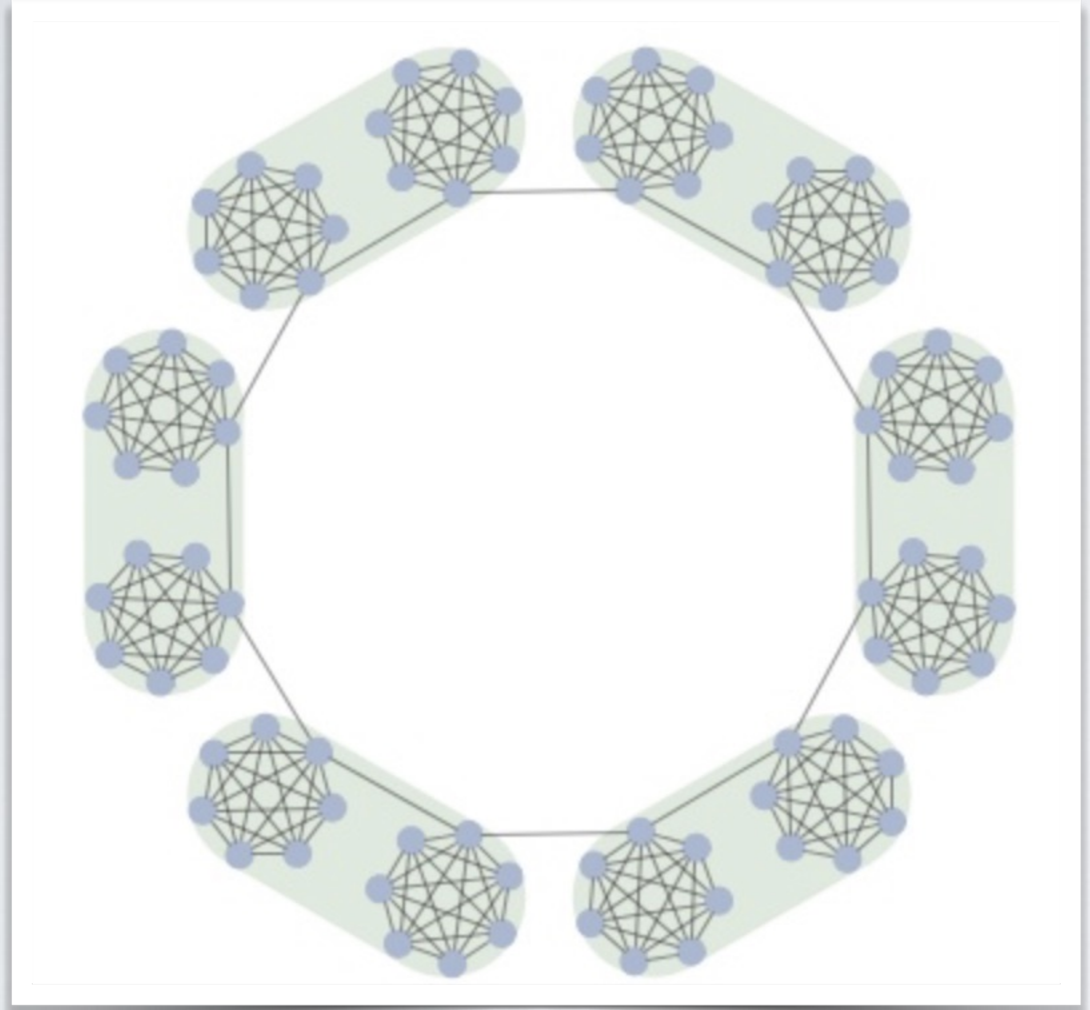
Let's consider a ring of cliques

Cliques are as dense as possible

Single edge between them:
=>As separated as possible

Any acceptable algorithm=>Each clique is a community

RESOLUTION LIMIT



But with modularity:

Small graphs=> OK

Large graphs=>
The max of modularity obtained
by merging cliques

RESOLUTION LIMIT

- Discovery that Modularity has a “favorite scale”:
- For a graph of given **density** and **size:**
  - Communities cannot be smaller than a fraction of nodes
  - Communities cannot be larger than a fraction of nodes
- Modularity optimisation will never discover
  - Small communities in large networks
  - Large communities in small networks

RESOLUTION LIMIT

- Multi-resolution modularity

$$\sum_{i}^{c} e_{ii} - a_i^2 \quad \Rightarrow \quad \sum_{i}^{c} e_{ii} - \lambda a_i^2$$

λ = Resolution parameter

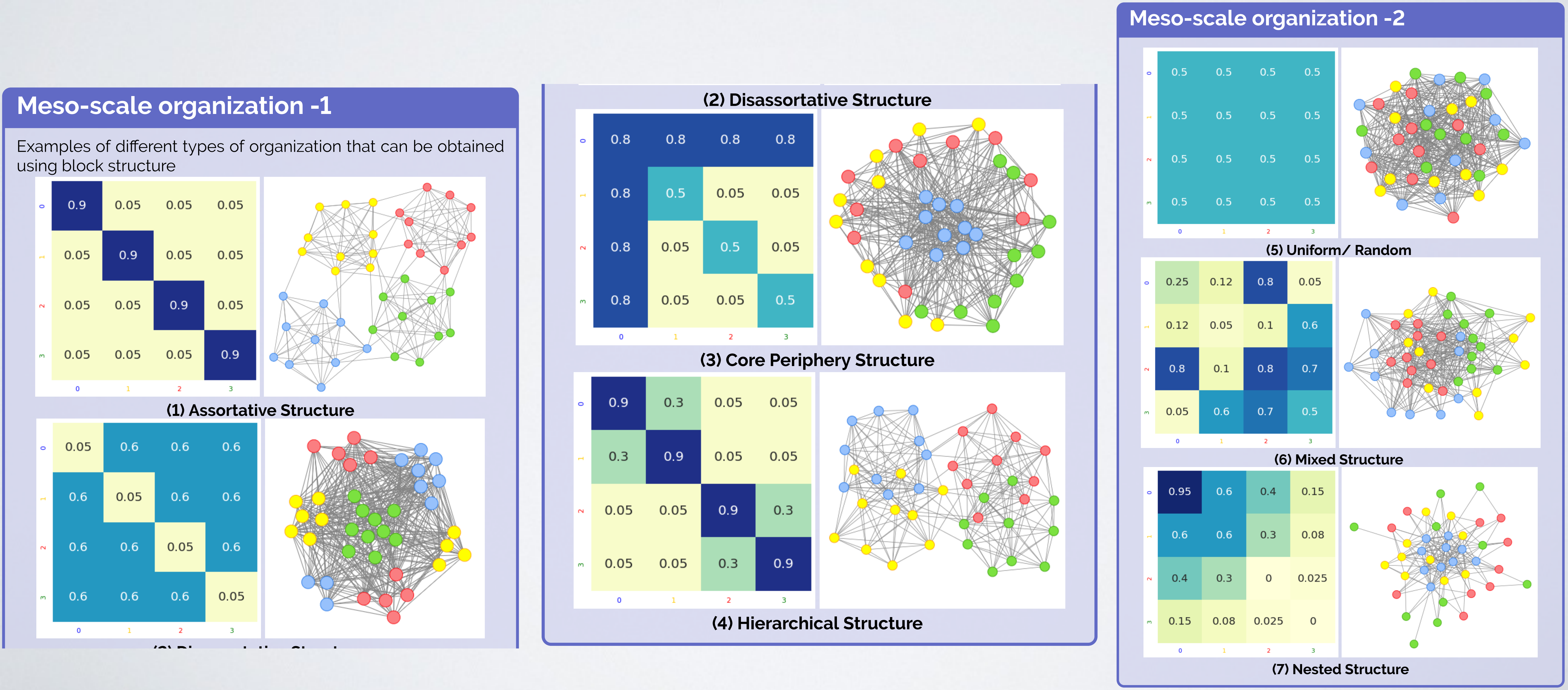
More a patch than a solution…

STOCHASTIC BLOCK MODELS

- Stochastic Block Models (SBM) are based on statistical models of networks
- They are in fact more general than usual communities.
- The model is:
  - Each node belongs to 1 and only 1 community
  - To each pair of communities, there is an associated density (probability of each edge to exist)

STOCHASTIC BLOCK MODELS

- SBM can represent different things:
  - Associative SBM: density inside nodes of a same communities >> density of pairs belonging to different communities.



EVALUATION OF COMMUNITY STRUCTURE

INTRINSIC EVALUATION

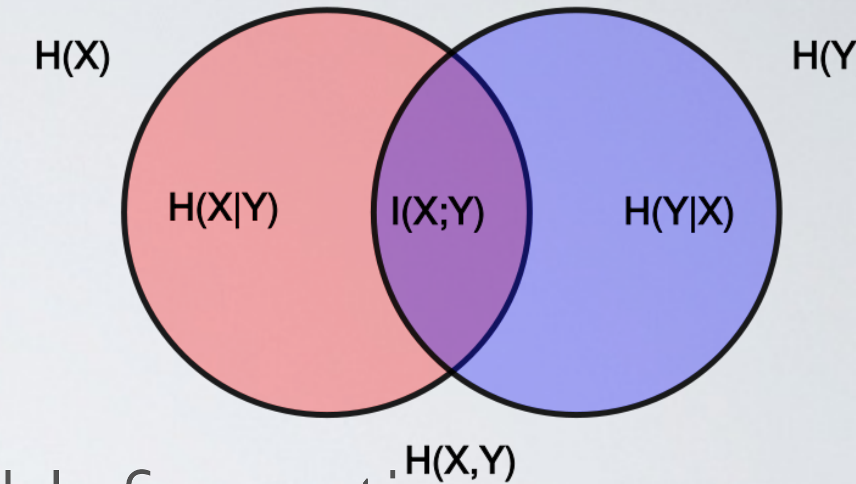
- Partition quality function
  - Already defined: **Modularity**, **graph compression**, etc.
- Quality function for individual community
  - Internal Clustering Coefficient
  - Conductance: $\frac{|E_{out}|}{|E_{out}|+|E_{in}|}$
    - Fraction of external edges

$|E_{in}|, |E_{out}|$:
# of links to nodes inside (respectively, outside) the community

MEASURING PARTITION SIMILARITIES

- Synthetic or GT, we get:
  - Reference communities
  - Communities found by algorithms
- How to measure their similarity ?
  - NMI => AMI
  - ARI
  - …

MEASURING PARTITION SIMILARITIES



- NMI: Normalized Mutual Information

- Classic notion of Information Theory: Mutual Information
  - How much knowing one variable reduces uncertainty about the other
  - Or how much in common between the two variables

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

- Normalized version: NMI
  - 0: independent, 1: identical

- Adjusted for chance: aNMI

$$AMI(U,V) = \frac{MI(U,V) - E\{MI(U,V)\}}{\max\{H(U),H(V)\} - E\{MI(U,V)\}}$$

CORE-PERIPHERY