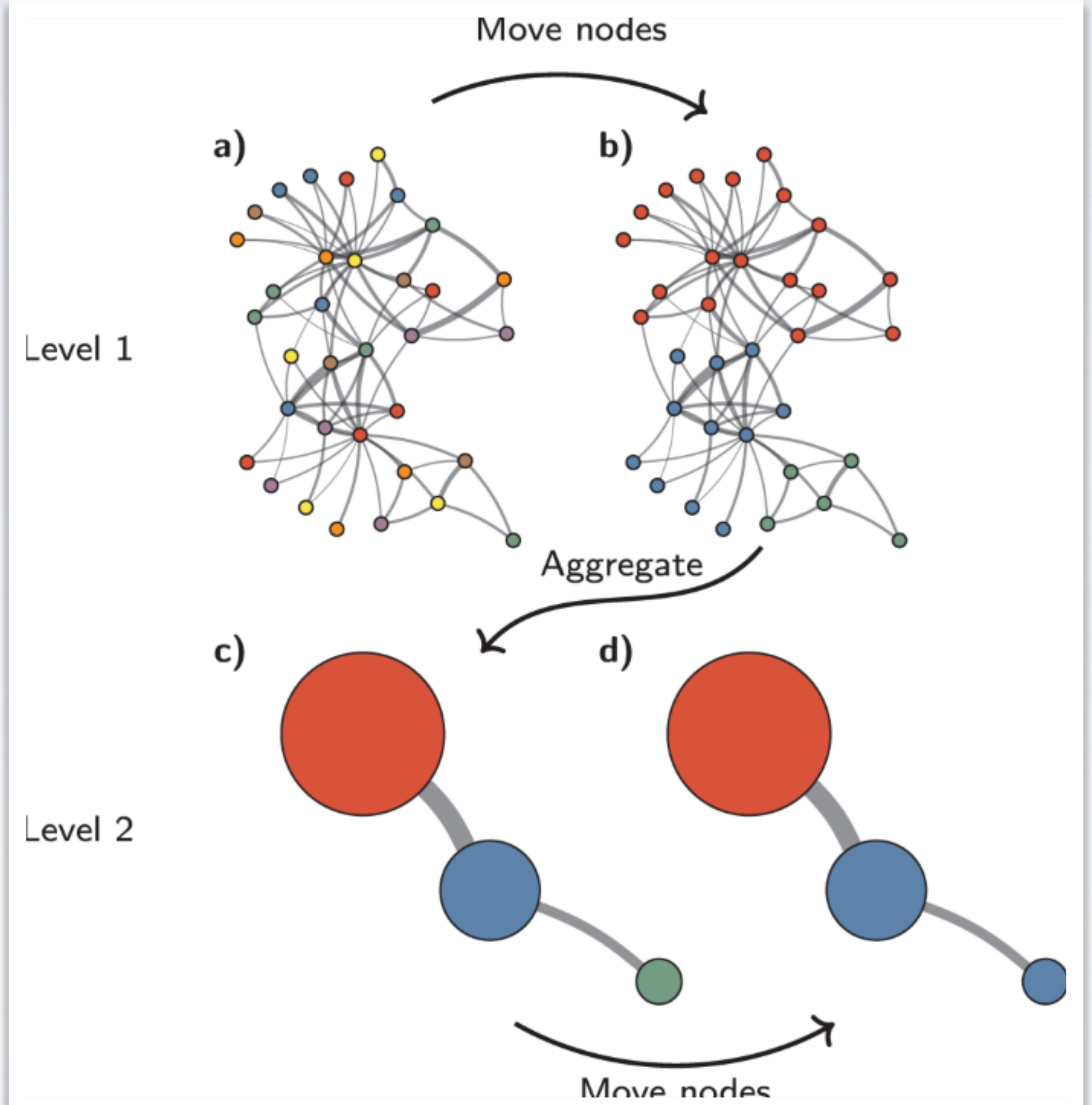
COMMUNITY DETECTION (GRAPH CLUSTERING)

LOUVAIN ALGORITHM



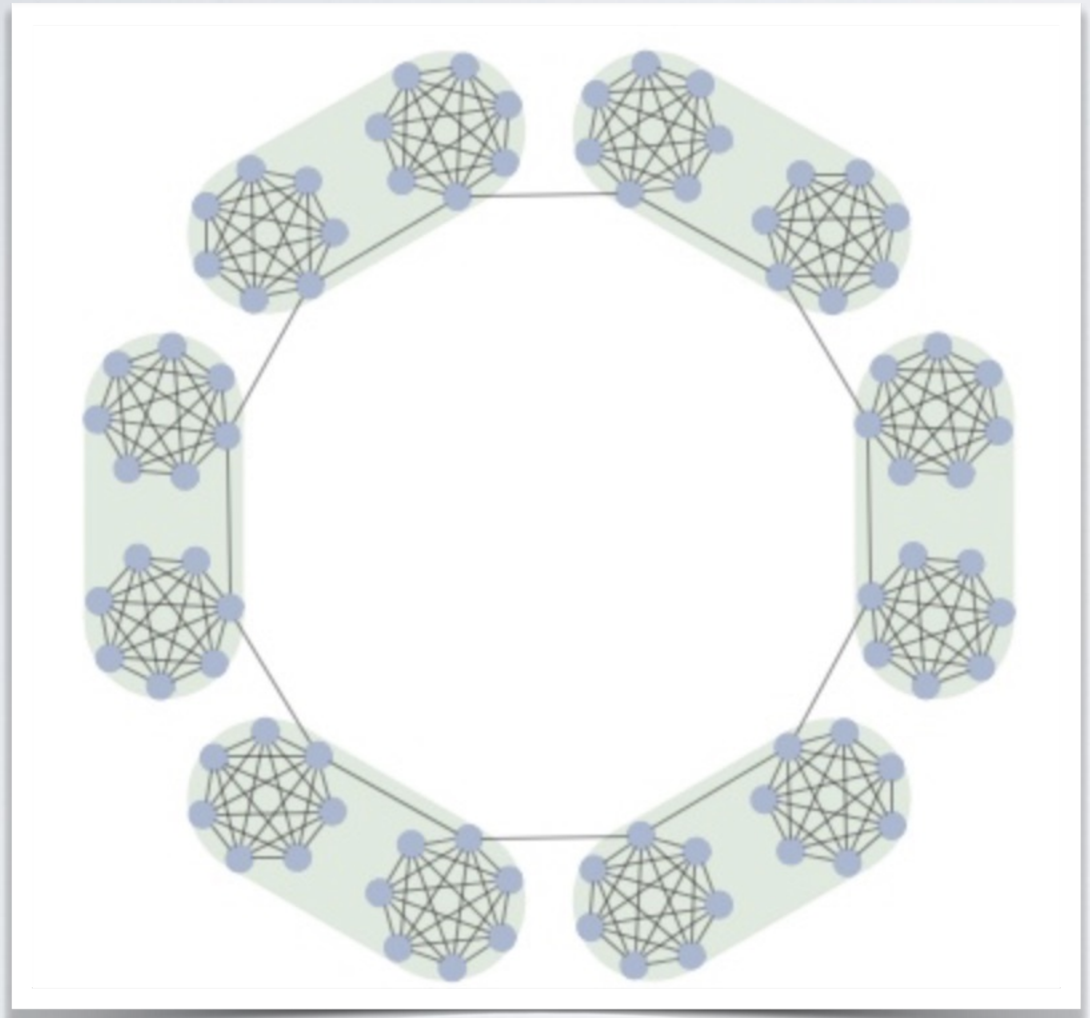
Blondel, Vincent D., et al. "Fast unfolding of communities in large networks." *Journal of statistical mechanics: theory and experiment* 2008.10 (2008): P10008.

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

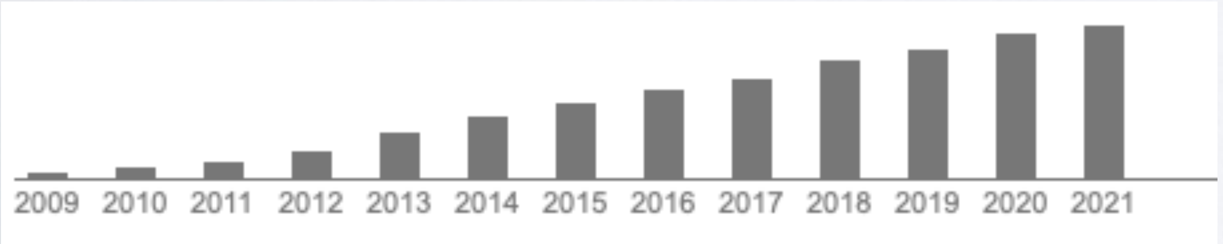
COMMUNITY DETECTION

- 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 ?

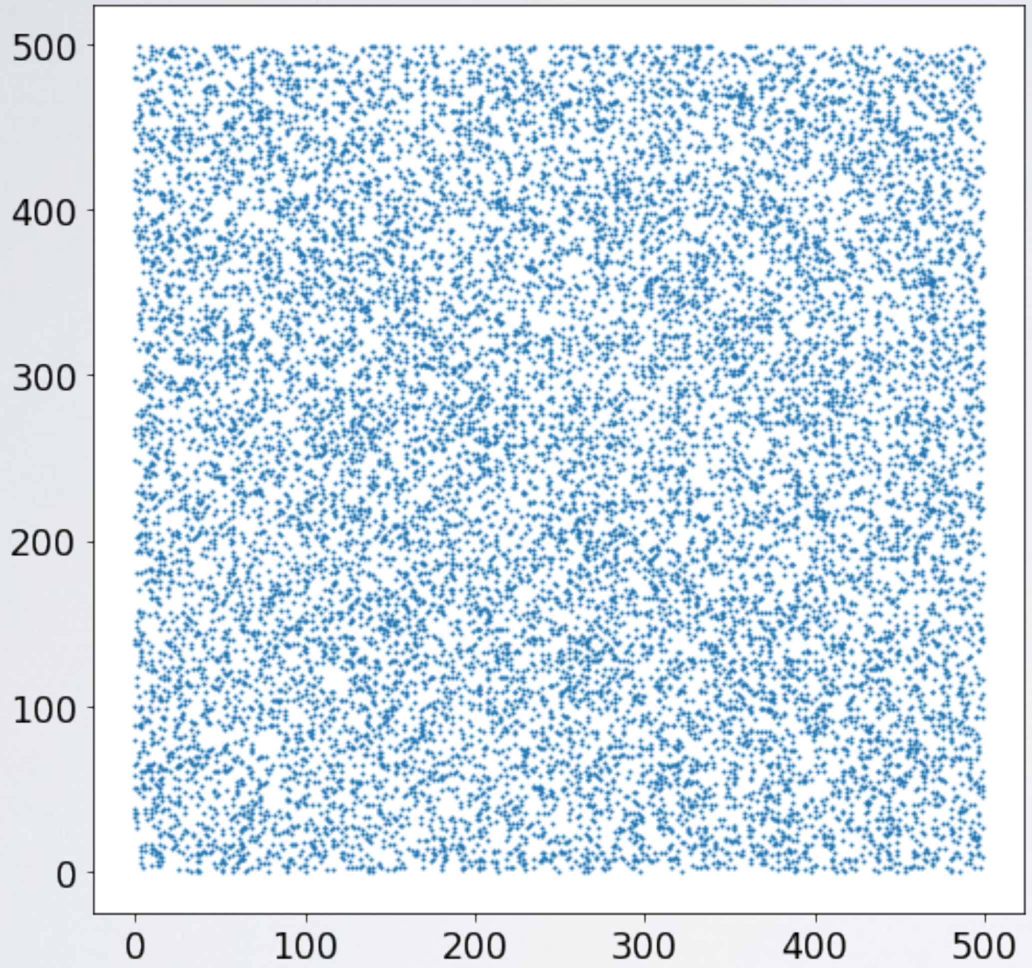
OTHER WEAKNESSES

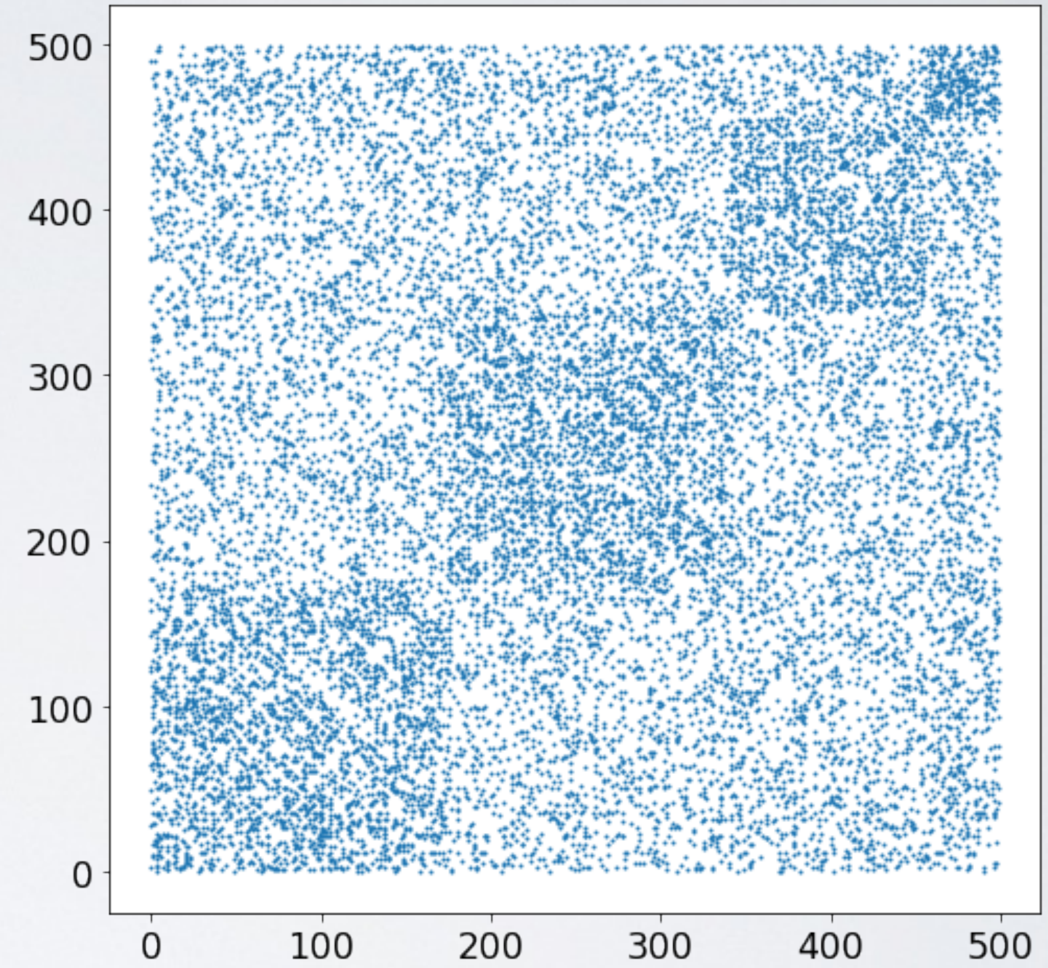
- Modularity has other controversial/not-intuitive properties:
  - Global measure => a difference in one side of the network can change communities at the other end (imagine a growing clique ring…)
  - Unable to find no community:
    - Network without community structure: Max modularity for partitions driven by random noise

- To this day, Louvain and modularity remain most used methods
  - Results are usually "good"/useful
  - Some newer methods gain popularity (SBM, Leiden,…)



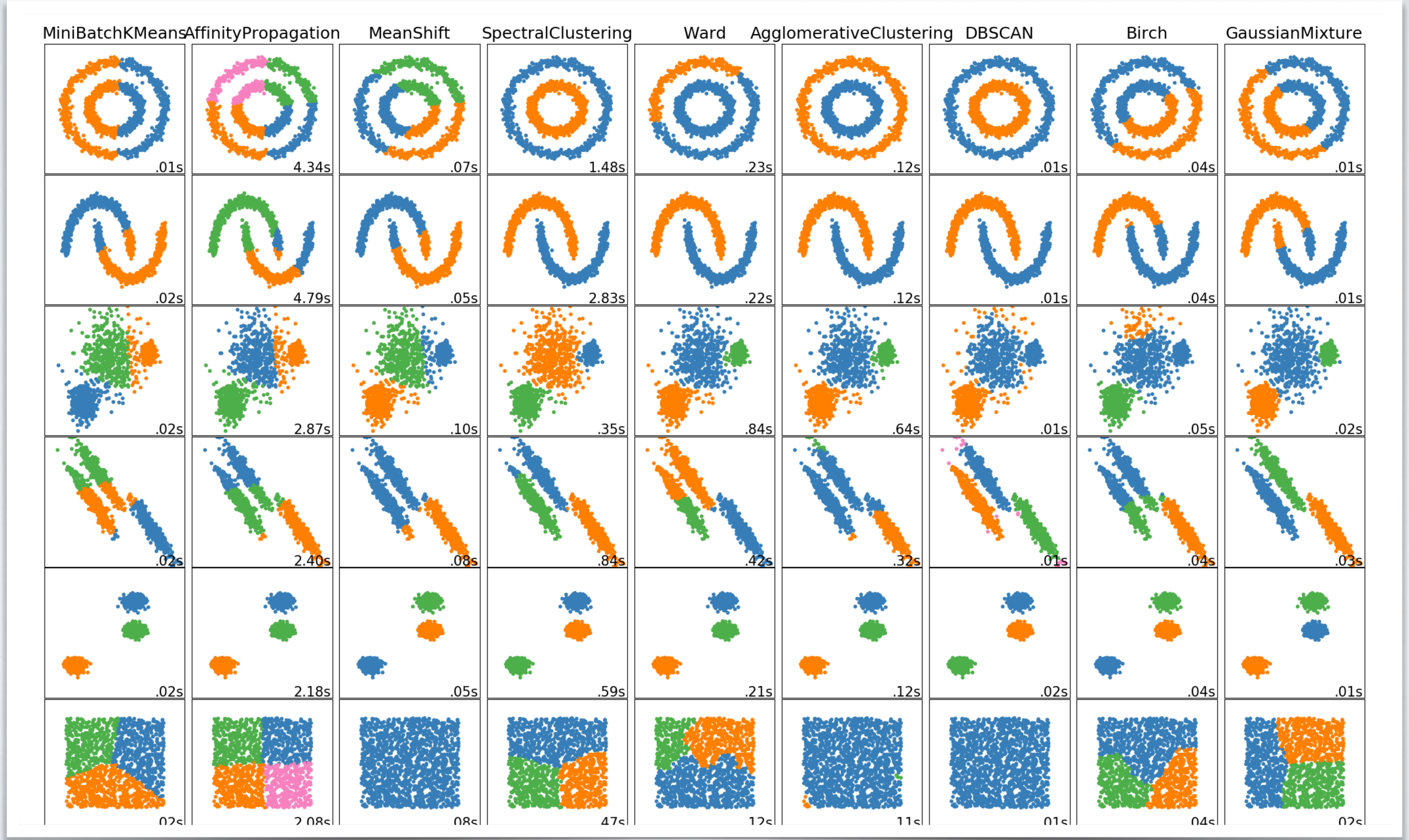
OTHER WEAKNESSES





A of a random network with nodes ordered Randomly (left) or according to Modularity maximization (right)

COMMUNITY DETECTION



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…

ALTERNATIVES

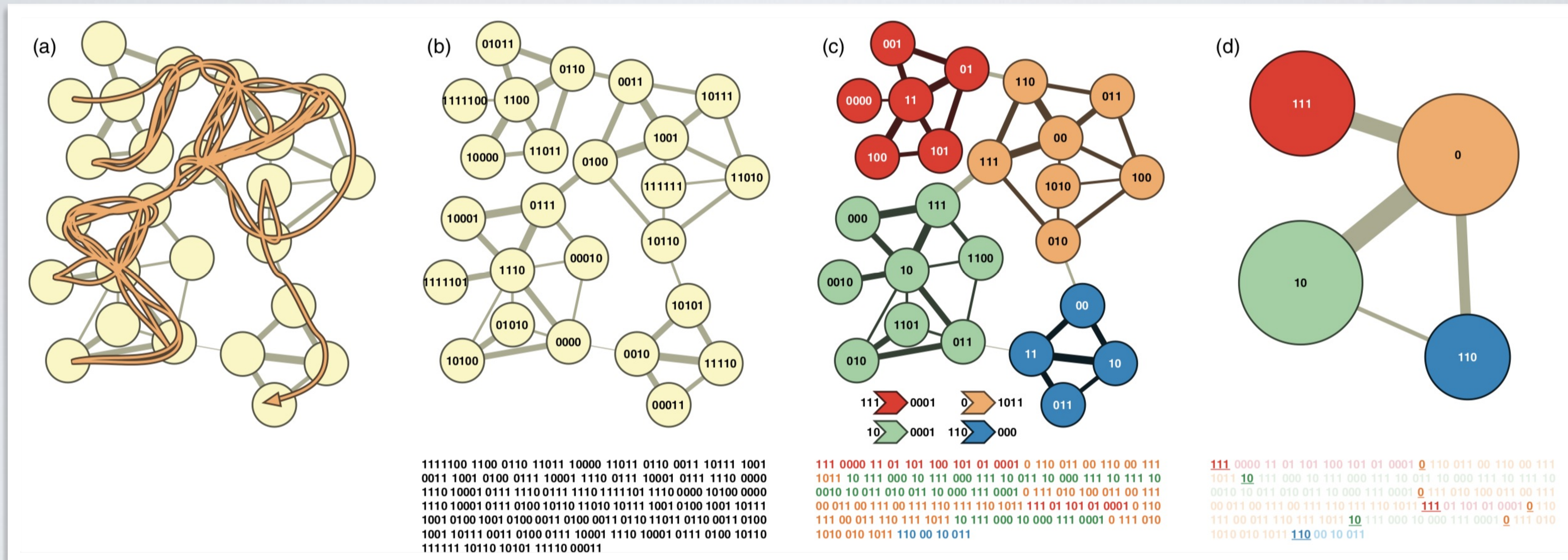
- Most serious alternatives (in my opinion)
  - Infomap (based on information theory —compression)
  - Stochastic block models (bayesian inference)
- These methods have a clear definition of what are good communities. Theoretically grounded

INFOMAP

- [Rosvall & Bergstrom 2009]
- Find the partition minimizing the *description* of any *random walk* on the network
- We want to *compress* the description of random walks

Rosvall, Martin, and Carl T. Bergstrom. "Maps of random walks on complex networks reveal community structure." *Proceedings of the National Academy of Sciences* 105.4 (2008): 1118-1123.

INFOMAP



Random walk

Description Without Communities

With communities

Huffman coding: short codes for frequent items
Prefix free: no code is a prefix of another one (avoid fix length/separators)

The Infomap method

Finding the optimal partition M:

- Shannon's source coding theorem (Shannon's entropy)

for a probability distribution P = {p_i}

$$L(\mathcal{P}) = H(\mathcal{P}) \equiv -\sum_i p_i \log p_i$$

- Minimise the expected description length of the random walk

Sum of Shannon entropies of multiple codebooks weighted by the rate of usage

probability of between modules movements of a RW

probability of within modules movements of a RW

$$L(\mathbf{M}) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} p^{i}_{\circlearrowright} H(\mathcal{P}^{i})$$

Expected decryption length of partition M

Cost of movement between modules, i.e. the frequency weighted average length of codewords

Cost of movements inside the module

Algorithm

1. Compute the fraction of time each node is visited by the random walker (Power-method on adjacency matrix)
2. Explore the space of possible partitions (deterministic greedy search algorithm - similar to Louvain but here we join nodes if they decrease the description length)
3. Refine the results with simulated annealing (heat-bath algorithm)

INFOMAP

- To sum up:
  - Infomap defines a *quality function* for a partition different than modularity
  - Any algorithm can be used to optimize it (like Modularity)

- Advantage:
  - Infomap can recognize random networks (no communities)
  - Good results in practice, fast.

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

Stochastic Block Models (SBM)

A stochastic block model is a random graph model defined by:

k	number of blocks
b	$n \times 1$ vector such as b_i describes the index of the block of node i.
E	$k \times k$ **stochastic block matrix**, such as E_{ij} gives the number of edges between blocks i and j (or the probability to observe an edge between any pair of nodes chosen with one node in each of the two blocks).

Generating networks

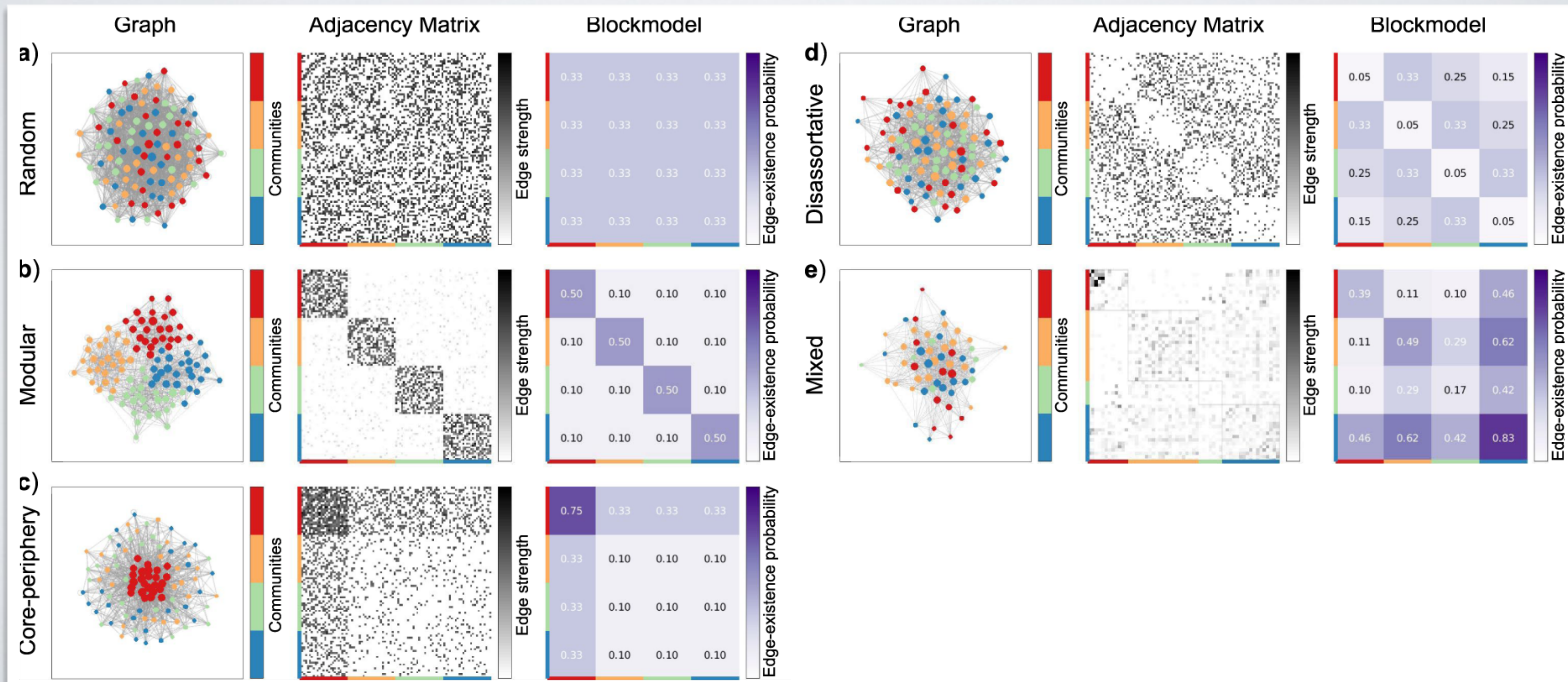
1. Take N disconnected nodes
2. Connect each $u,v \in V$ nodes with probability $E_{b(u),b(v)}$

Properties:

- Every vertices in a same module are statistically equivalent
- Vertices in a module are connected by a random graph
- Emergent degree distribution is a combination of Poisson distributions

STOCHASTIC BLOCK MODELS

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



STOCHASTIC BLOCK MODELS

- Being able to represent any block preference is powerful and potentially relevant
- Problem: Often hard to interpret in real situations.
  - SBM can be "constrained": we impose that intra d.>inter d.

STOCHASTIC BLOCK MODELS

- General idea of SBM community detection:
  - Specify the desired number of cluster
  - Find parameters to optimize the maximum likelihood
    - Principle: The best parameters are those that allow to generate the observed network with the highest probability
- Main weakness of this approach
  - Number of clusters must be specified (avoid trivial solution)

STOCHASTIC BLOCK MODELS

- Solution to the number of blocks problem:
- [2016 Peixoto]
  - Minimum Description Length (MDL) (Occam's razor)
  - We minimise the cost of encoding
    - The model (its parameters)
    - The graph knowing the model

STOCHASTIC BLOCK MODELS

Information Theoretic Formulation

A: adjacency matrix
k: degree sequence
e: Matrix of edges between blocks
b: partitions

Model cost (bits) $S + L$

$$S = -\,log_2 P(A \,|\, k, e, b)$$

bits necessary to encode the graph knowing the model

$$L = -\,log_2 P(k, e, b)$$

bits necessary to encode the model

Objective = maximize the graph compression.
-Too many communities: over-complexifying the model
-Too few communities: Harder to encode the graph, since the model provides few useful information
Occam's razor

Peixoto, Tiago P. "Bayesian stochastic blockmodeling." *arXiv preprint arXiv:1705.10225* (2017).

STOCHASTIC BLOCK MODELS

- To sum up:
  - SBM have a convincing definition of communities
  - In practice, inference usually slower than louvain/infomap
  - But more powerful
  - Can also say if there is no community
  - And also suffer from a form of resolution limit
- Less often used, but regain popularity since works by Peixoto.
  - Variants: degree-corrected, overlapping, corrected for clustering…

EVALUATION OF COMMUNITY STRUCTURE

EVALUATION

- We intuitively "know" what are good communities
- But we have:
  - Several mathematical formulations
  - Several optimisation (greedy…) algorithms that might introduce biases.
- How to know which method to use ?

EVALUATION

- Two main approaches:
  - Intrinsic/Internal evaluation
    - Partition quality function
    - Individual Community quality function
  - Comparison of observed communities and expected communities
    - Synthetic networks with community structure
    - Real networks with Ground Truth

INTRINSIC EVALUATION

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

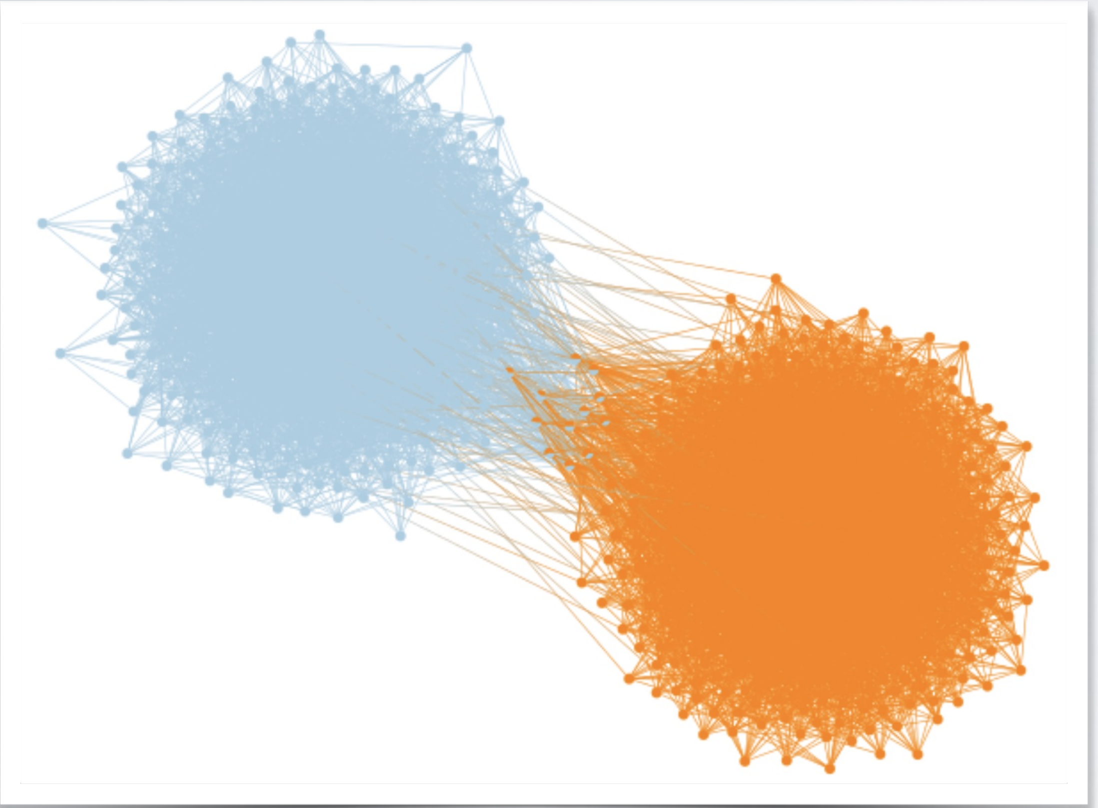
$|E_{in}|, |E_{out}|$:
# of links to nodes inside (respectively, outside) the community

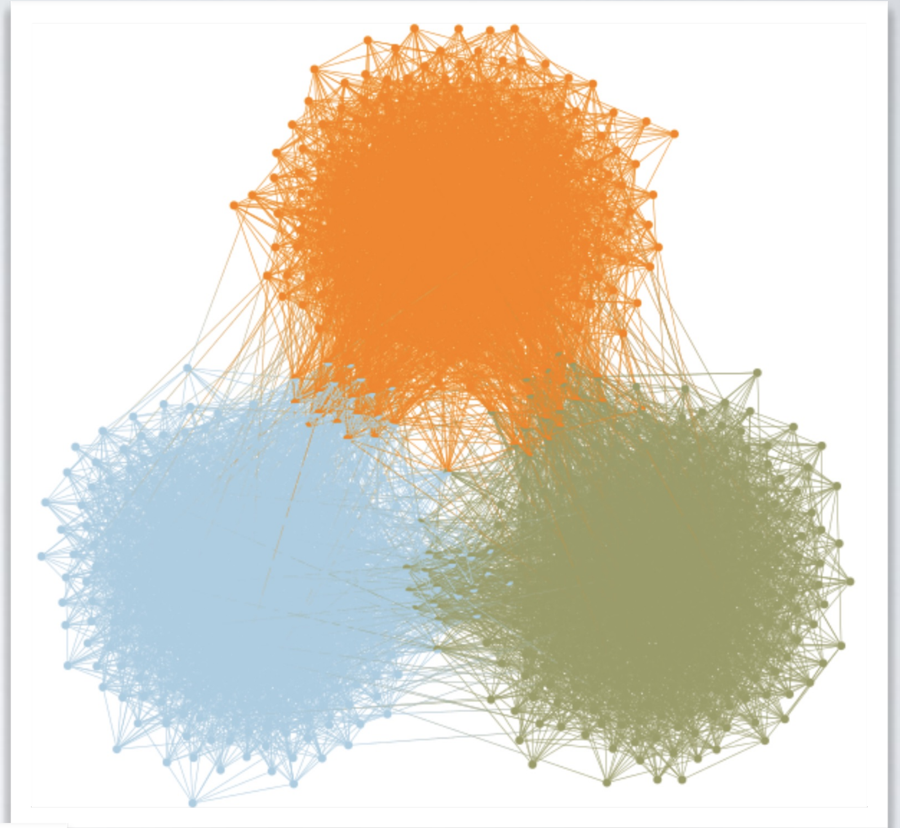
COMPARISON WITH GROUND TRUTH

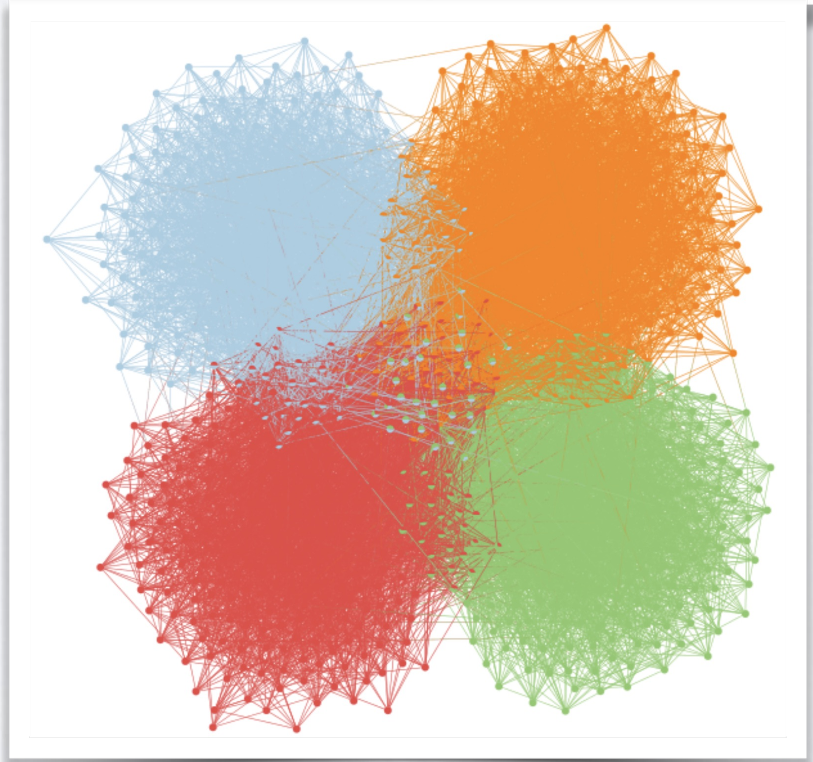
SYNTHETIC NETWORKS

- Planted Partition models:
  - Another name for SBM with manually chosen parameters
    - Assign degrees to nodes
    - Assign nodes to communities
    - Assign density to pairs of communities
    - Attribute randomly edges
  - Problem: how to choose parameters?
    - Either oversimplifying (all nodes same degrees, all communities same #nodes, all intern densities equals…)
    - Or ad-hoc process (sample values from distributions)

SYNTHETIC NETWORKS



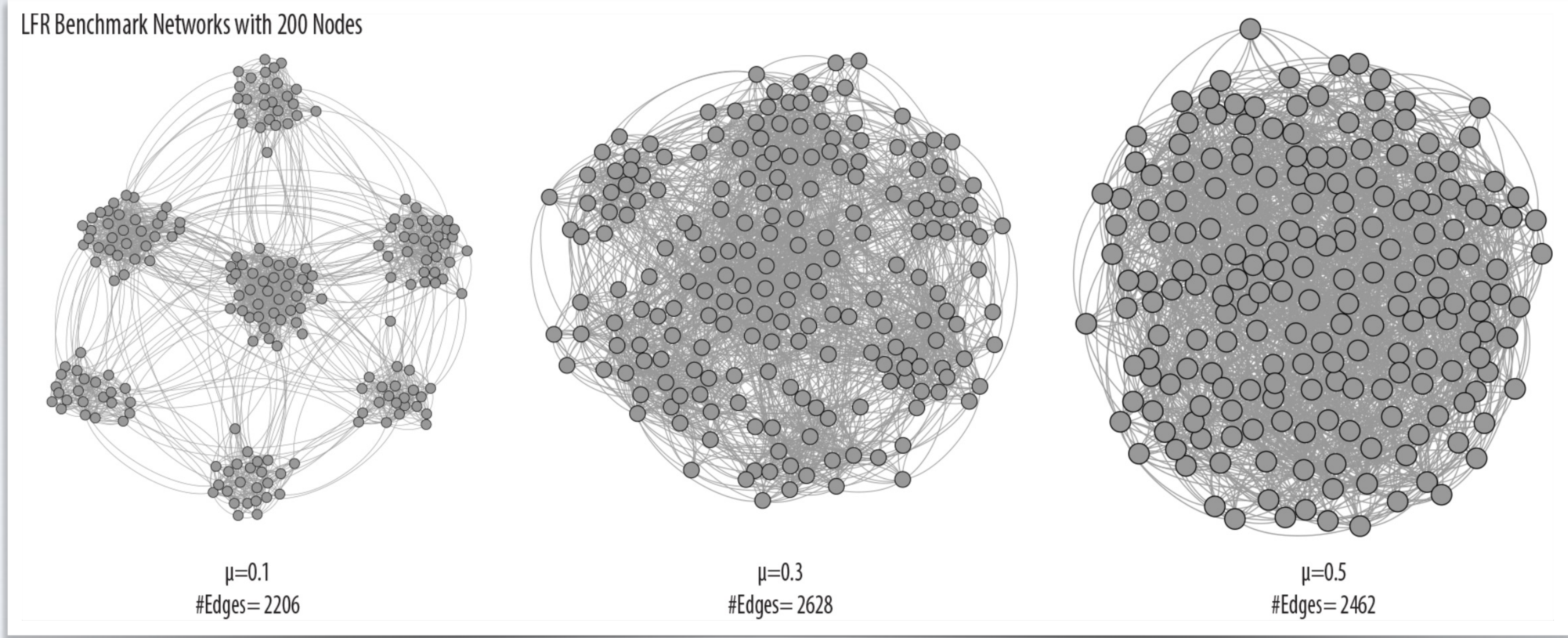




SYNTHETIC NETWORKS

- LFR Benchmark [Lancichinetti 2008]
  - High level parameters:
    - Slope of the power law distribution of degrees/community sizes
    - Avg Degree, Avg community size
    - Mixing parameter: fraction of external edges of each node
  - Varying the mixing parameter makes community more or less well defined
- Currently the most popular

SYNTHETIC NETWORKS



SYNTHETIC NETWORKS

- Pros of synthetic generators:
  - We know for sure the communities we should find
  - We can control finely the parameters to check robustness of methods
    - For instance, resolution limit…
- Cons:
  - Generated networks are not realistic: simpler than real networks
    - LFR: High CC, scale free, but all nodes have the same mixing coefficient, no overlap, …
    - SBM: depend a lot on parameters, random generation might lead to unexpected ground truth (it is *possible* to have a node with no connections to other nodes of its own community…)

REAL NETWORKS WITH GT

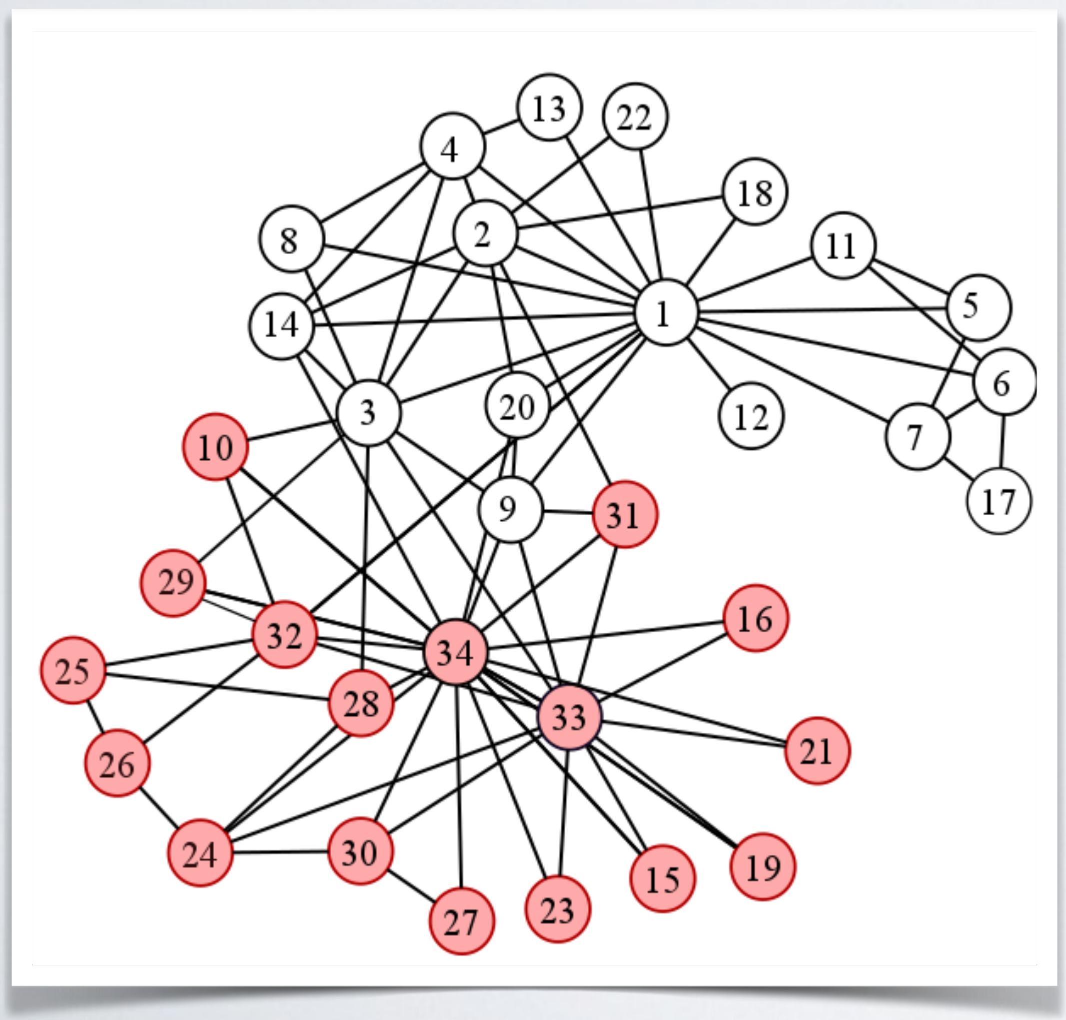
- In some networks, **ground truth** communities are known:
  - Social networks, people belong to groups (Facebook, Friendsters, Orkut, students in classes…)
  - Products, belonging to categories (Amazon, music…)
  - Other resources with defined groups (Wikipedia articles, Political groups for vote data…)
- Some websites have collected such datasets, e.g.
  - http://snap.stanford.edu/data/index.html

REAL NETWORKS WITH GT

- Pros of GT communities:
  - Retain the full complexity of networks and communities
- Cons:
  - No guarantee that communities are **topological** communities.
  - In fact, they are not: some GT communities are not even a single connected component…
- Currently, controversial topic
  - Some authors say it is non-sense to use them for validation
  - Some others consider it necessary

REAL NETWORKS WITH GT

- Example: the most famous of all networks: Zackary Karate Club

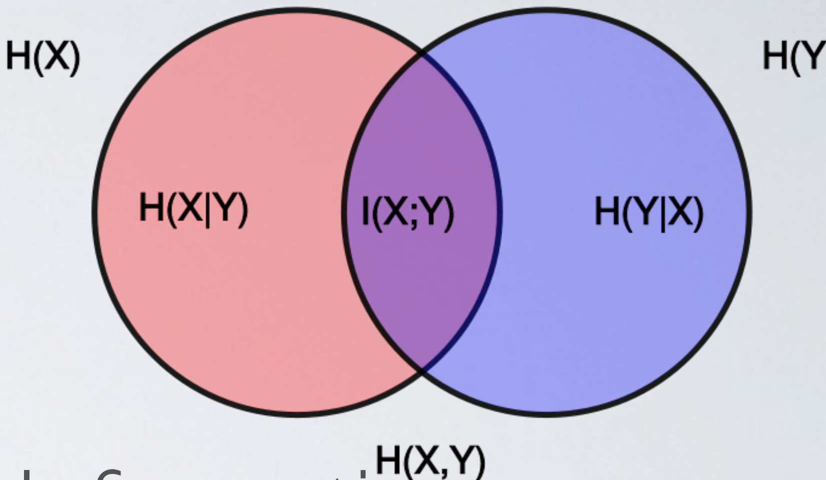


If your algorithm find the right communities,
Then it is wrong…

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

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

MEASURING PARTITION SIMILARITIES

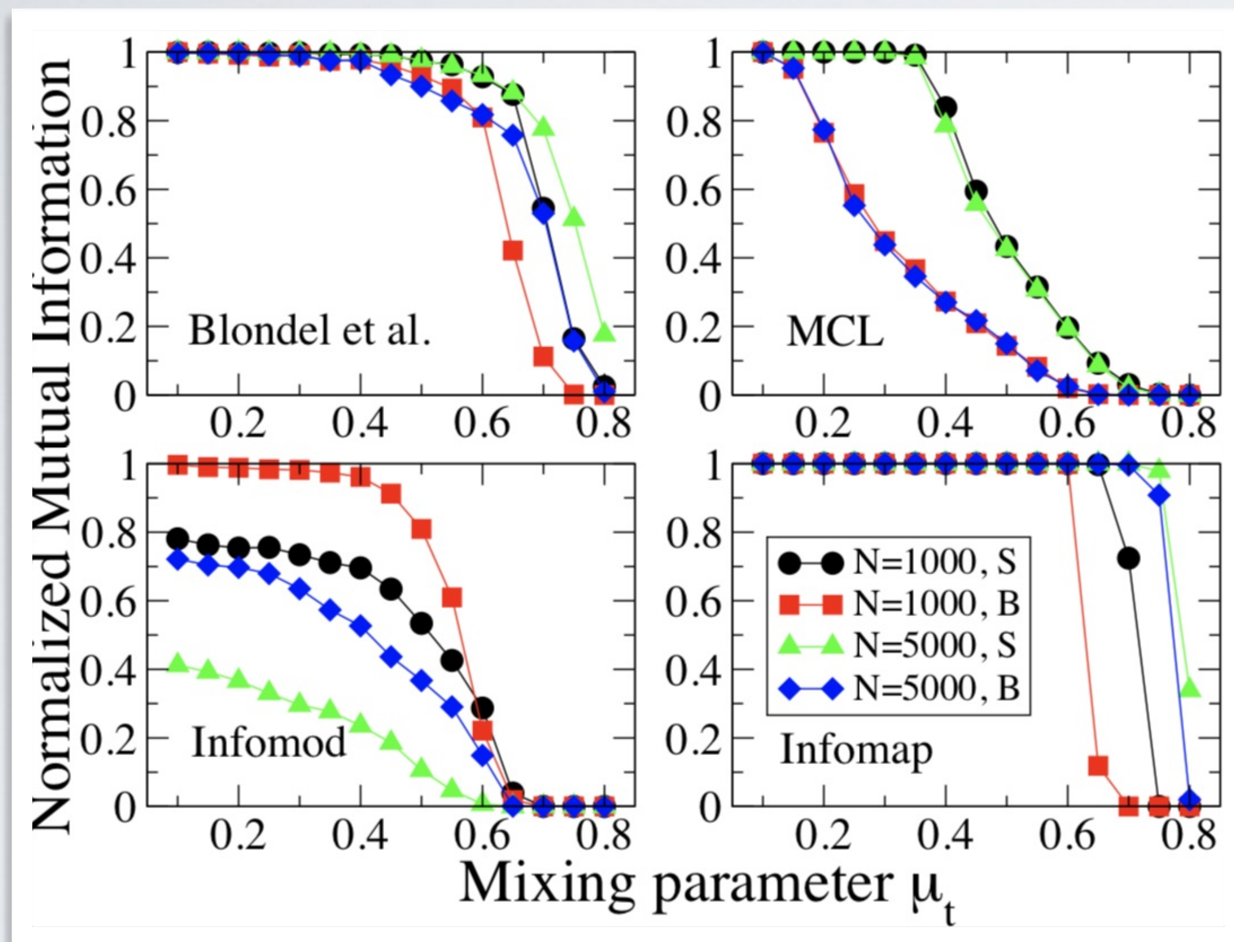
$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

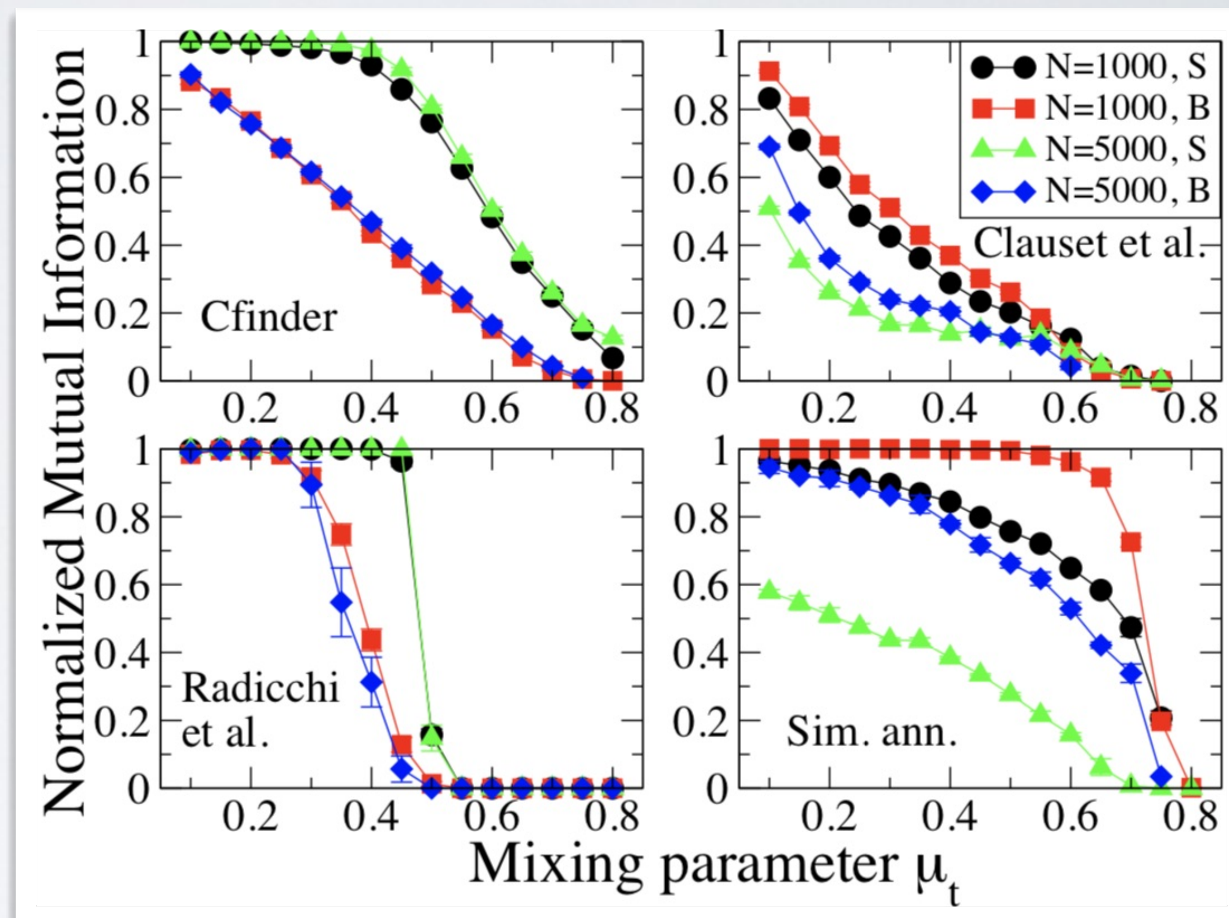
For all pairs of clusters taken in different partitions

Probability for a node picked at random to belong to both x and y

Probably for a node picked at random to belong to x

ALGORITHMS COMPARATIVE ANALYSIS

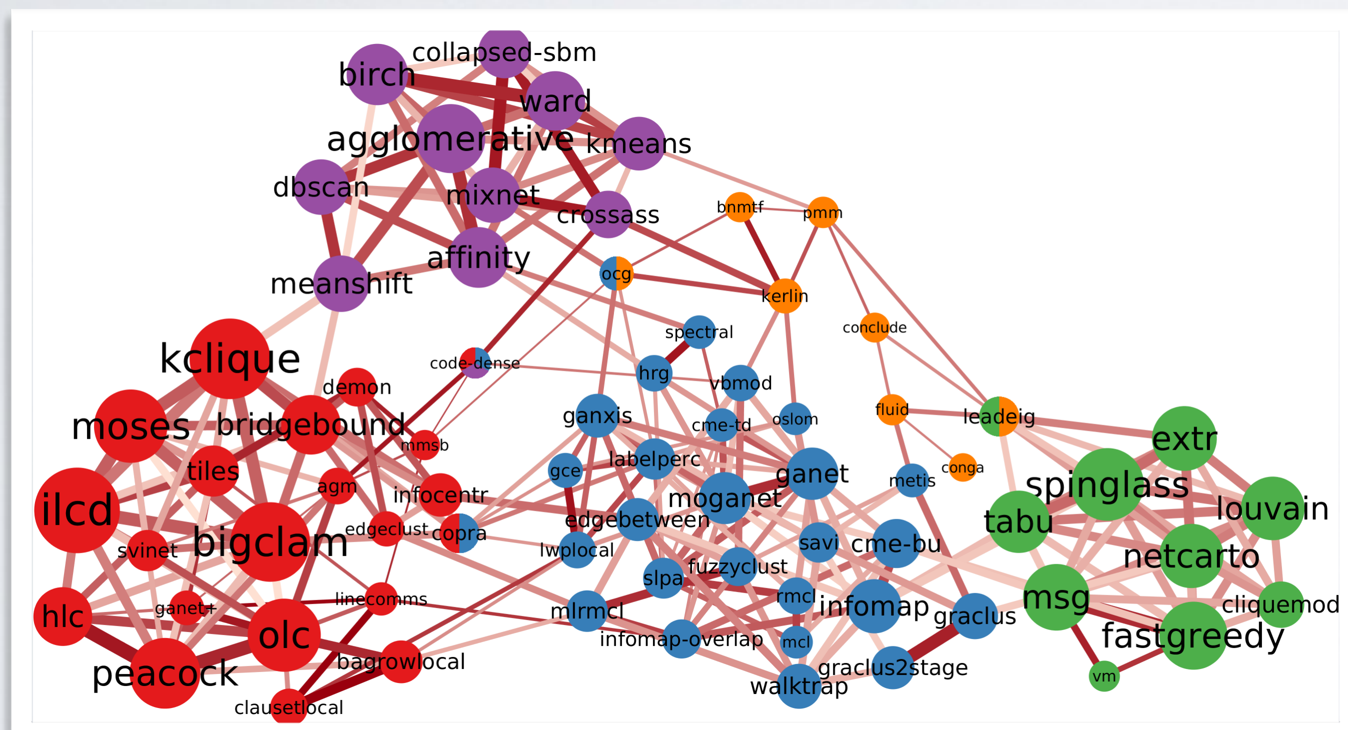




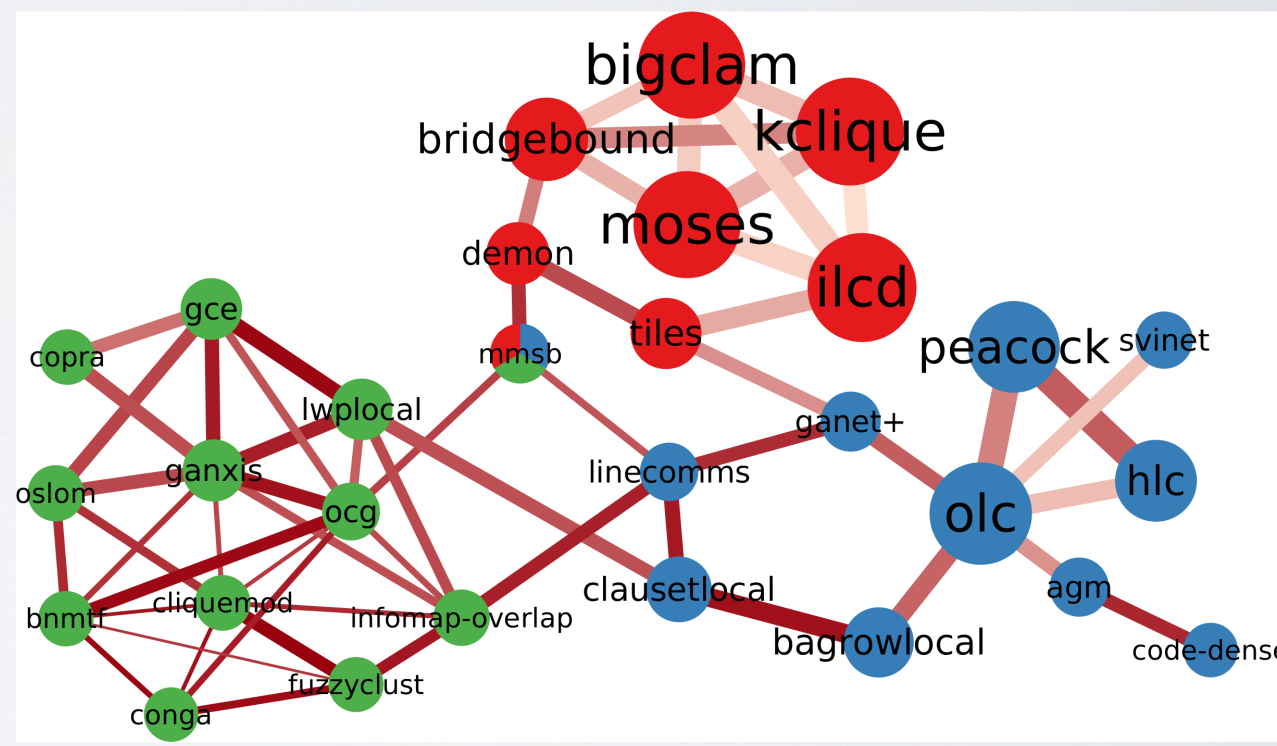
Lancichinetti, Andrea, and Santo Fortunato. "Community detection algorithms: a comparative analysis." *Physical review E* 80.5 (2009): 056117.

ALGORITHMS COMPARATIVE ANALYSIS

Rank	Algorithm	oNMI MAX
1	linecomms	165
2	oslom	73
3	infomap-overlap	64
4	savi	62
5	labelperc	57
6	rmcl	54
7	edgebetween	41
7	leadeig	41
7	vbmod	41
10	gce	32



All methods



Overlapping only

Coscia, Michele. "Discovering Communities of Community Discovery." *arXiv preprint arXiv:1907.02277* (2019).

OTHER MESO-SCALE ORGANIZATIONS

MESO-SCALE

- MACRO properties of networks:
  - degree distribution, density, average shortest path…
- MICRO properties of networks:
  - Centralities
- MESO-scale: what is in-between
  - Community structure
  - Overlapping Community Structure
  - Core-Periphery
  - Spatial Organization (another class)

CORE-PERIPHERY

- Already introduced in the first class, k-cores, etc.

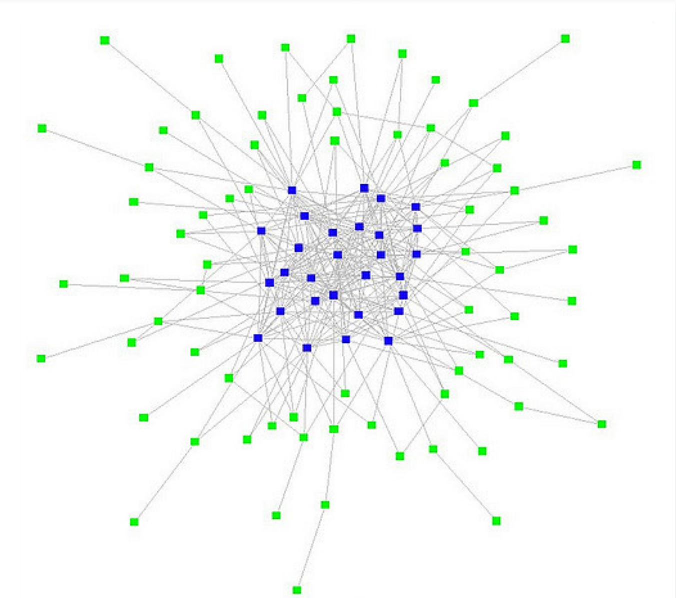
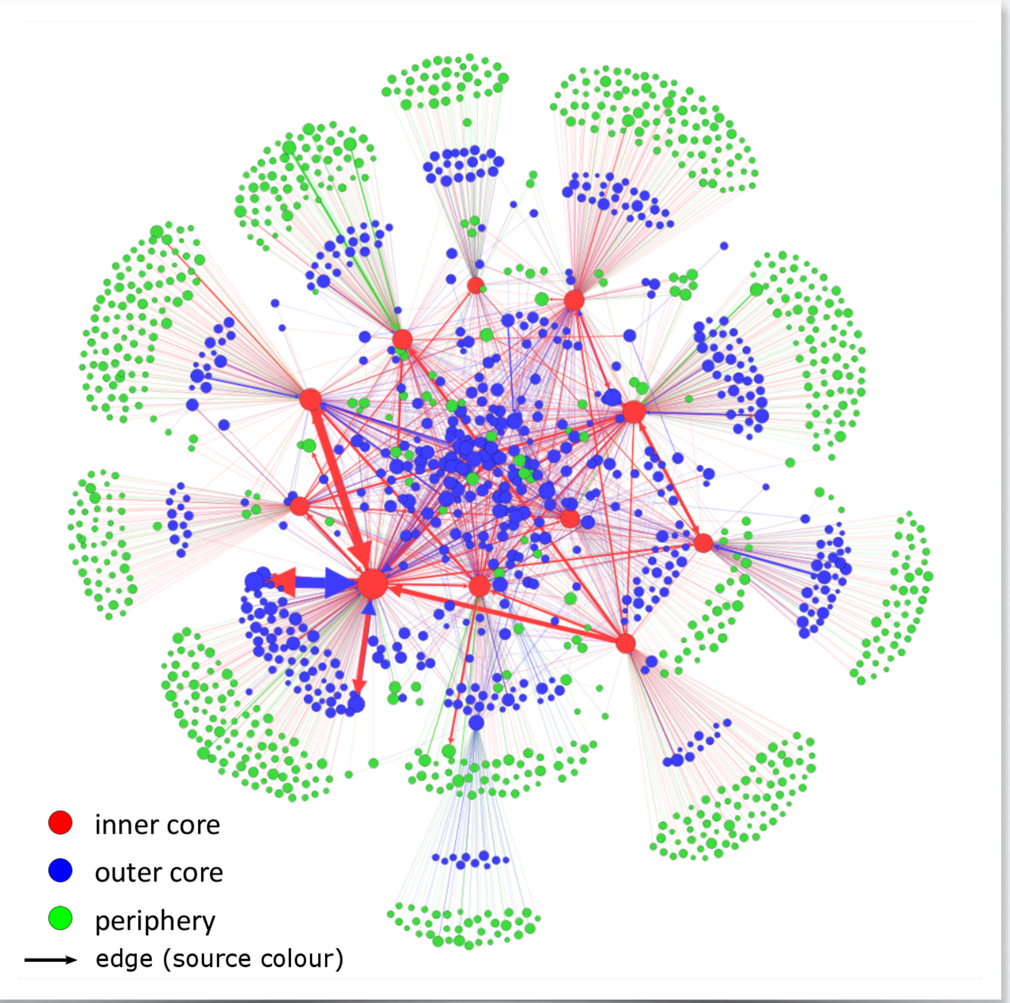
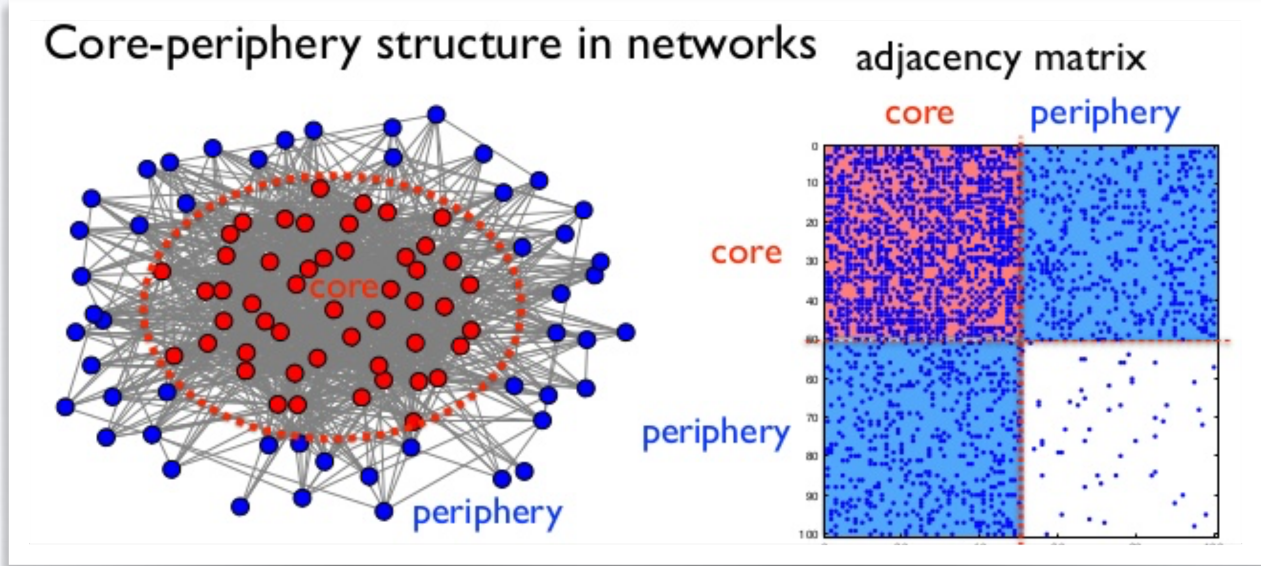


Figure 4 – Core/Periphery Network





OVERLAPPING COMMUNITIES

- In real networks, communities are often overlapping
  - Some of your High-School friends might be also University Friends
  - A colleague might be a member of your family
  - …
- Overlapping community detection is considered much harder
  - And is not well defined
- Difference between "attributes" and overlapping communities ?
  - Community of Women, Community of 17-19yo, Community of fans of…

OVERLAPPING COMMUNITIES

- Many algorithms
  - Adaptations of modularity, random walks, label propagations…
  - Original methods
  - Many local methods (local criterium), unlike global optimization for non-overlapping methods.

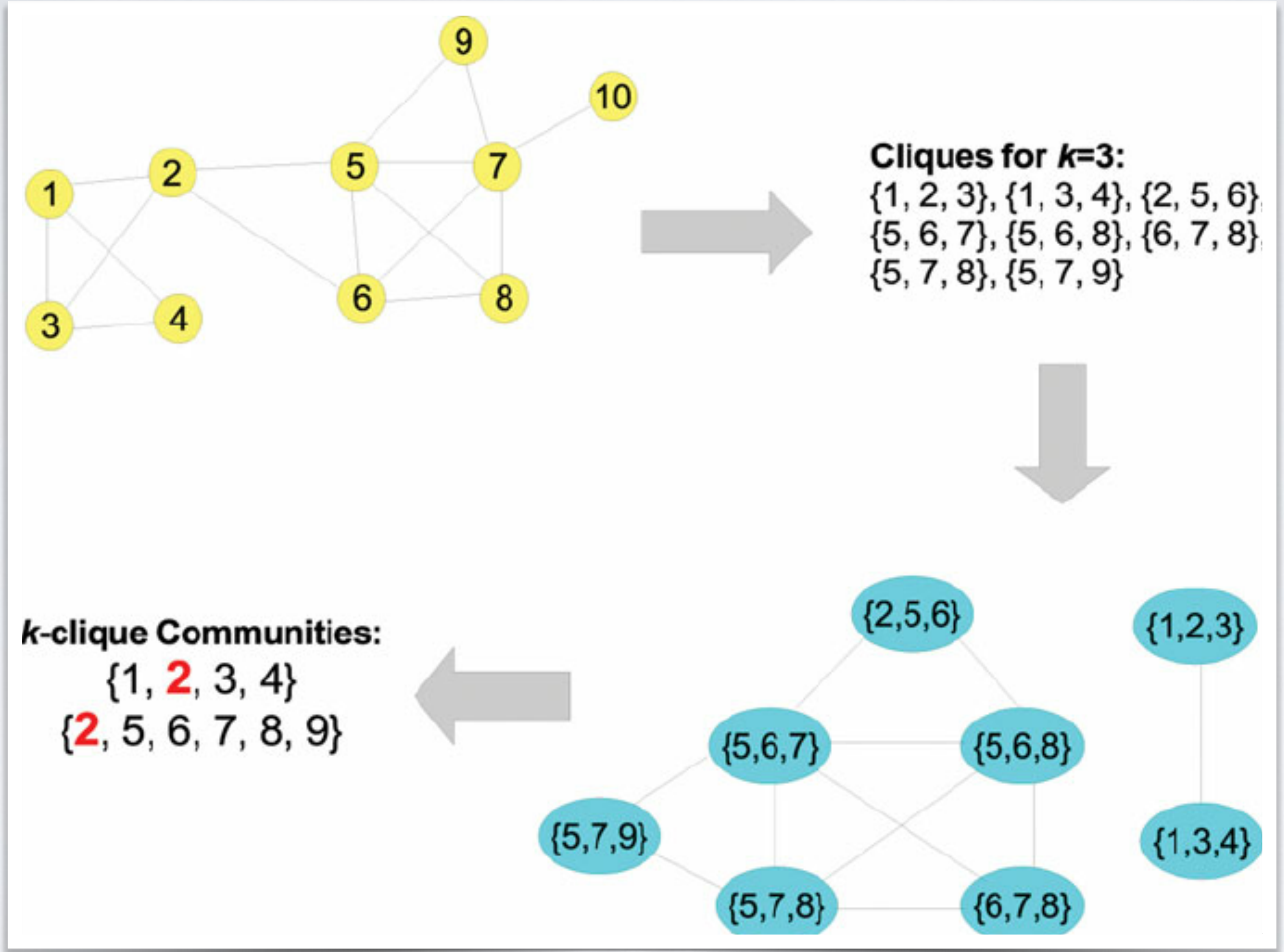
OVERLAPPING COMMUNITIES

- Motif-based definitions:
  - Cliques
    - Of a given size
    - Maximal cliques
  - N-cliques
    - Set of nodes such as there is at least a path of length <=N between them
    - Generalization of cliques for N>1
    - Computationally expensive

K-CLIQUE PERCOLATION

- (Other name: CPM, C-finder)
- Parameter: size k of atomic cliques
- 1)Find all cliques of size k
- 2)merge iteratively all cliques having k-1 nodes in common

K-CLIQUE PERCOLATION



HIERARCHICAL COMMUNITIES

Lancichinetti, Andrea, et al. "Finding statistically significant communities in networks." *PloS one* 6.4 (2011): e18961.

NESTEDNESS