Chapter 8

Community Discovery

Summary

- What’s a Community?
- Communities in static networks
- Evaluation & Benchmarking

Reading

- Chapter 9 of Barabasi’s book
- Fortunato’s survey

Community Discovery

A brief Introduction

Community Discovery

The aim of Community Discovery algorithms is to identify meso-scale topologies hidden within complex network structures

Why Community Discovery?

- "Cluster" homogeneous nodes relying on topological information

Major Problems:

- Community Discovery is an ill posed problem

  Each algorithm models *different* *properties* of communities
- Different approaches comparison
- Context Dependency

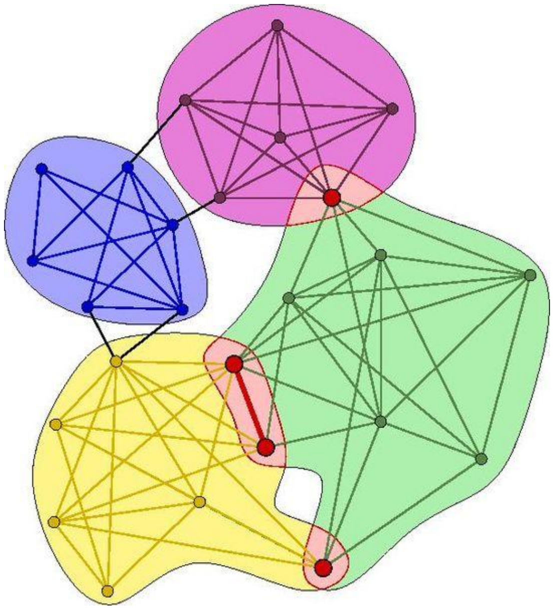
S. Fortunato. *Community detection in graphs.* Physics Reports 486 (2010).

Community Characteristics

Given the complexity of the problem a number of different typologies of approaches where proposed in order to:

Analyze:

- Directed\Undirected graphs
- Weighted\Unweighted graphs
- Multidimensional graphs
- ...



Following:

- Top-Down\Bottom-Up partitioning
- ...

Producing:

- Overlapping Communities
- Fuzzy Communities
- Hierarchical Communities
- Nested Communities
- ...

But...what is it exactly a community?

Unfortunately does not exist a universally shared definition of what a community is...

A *general* idea is that a community should represent:

"A set of entities where each entity is closer, in the network sense, to the other entities within the community than to the entities outside it."

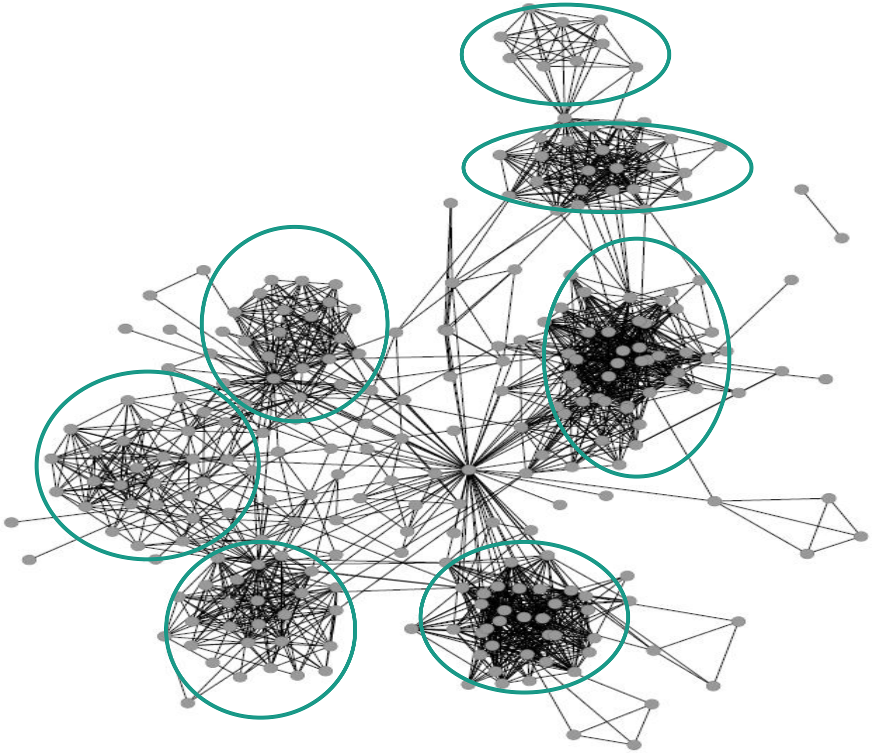
or, equivalently

"A set of nodes tightly connected within each other than with nodes belonging to other sets."

Communities in Complex Networks

In simple, small, networks it is easy identify them by looking at the structure…

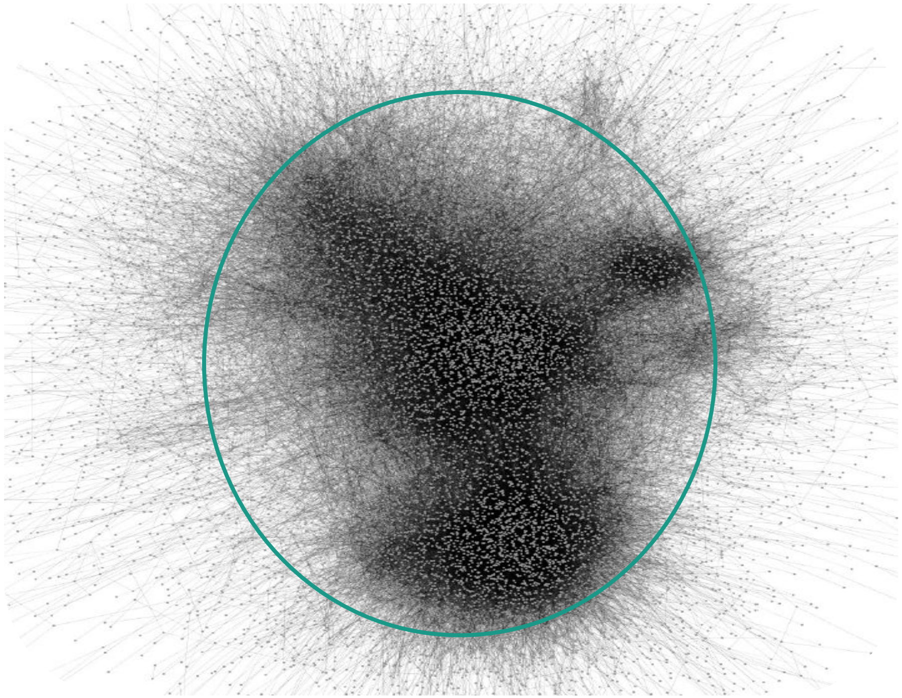
- i.e., using a Force directed layout



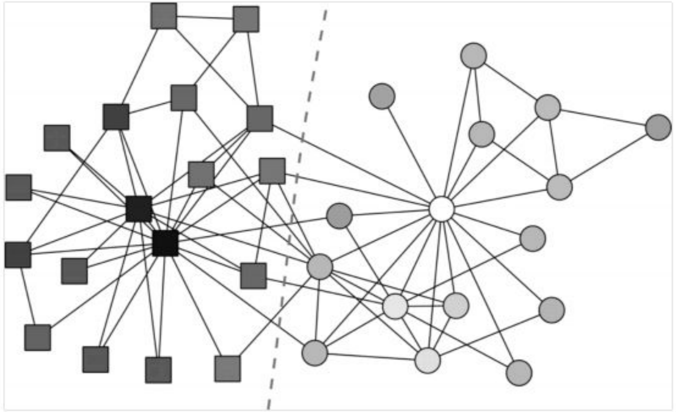
Real world networks? Too complex for visual analysis

We can't easily identify (e.g., visually) different communities

We need automated procedures!



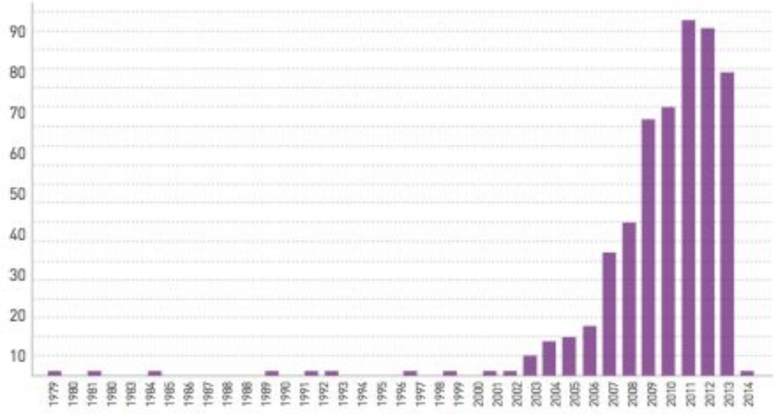
A first example…



Zachary's Karate Club

Communities emerge from the breakup of the Club

Citation history of the Zachary's Karate Club paper

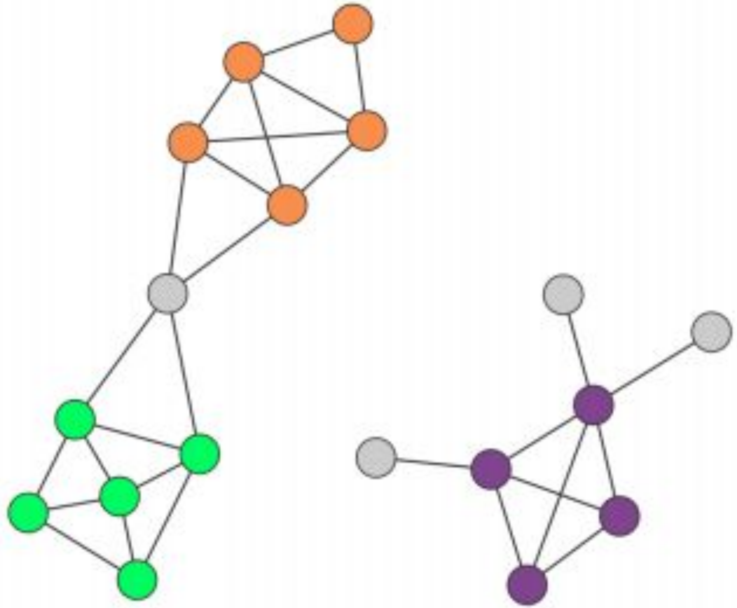


Karate Club Trophy

http://networkkarate.tumblr.com/

Communities:
a few Hypothesis

- **H1**: The community structure is uniquely encoded in the wiring diagram of the overall network
- **H2**: A community corresponds to a connected subgraph
- **H3**: Communities are locally dense neighborhoods of a network



Community Discovery

The nightmare of an ill-posed problem

Algorithms Taxonomy

Community Discovery algorithms can be classified according to:

- the constraints they impose to the meso-scale structures they are searching for
- the way they approach the community retrieval

We can group (*standard*) CD algorithms in the following families:

Internal Density	Bridge Detection
Feature Distance	Percolation
Entity Closeness	Structure Definition
Link Communities	No a priori definition

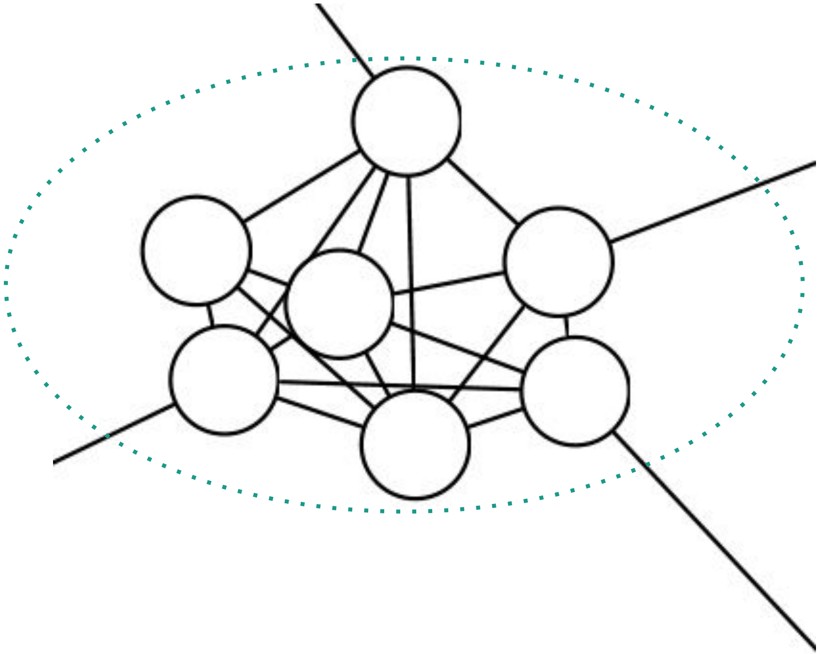
M. Coscia, F. Giannotti, and D. Pedreschi. *A classification for community discovery methods in complex networks.* Statistical Analysis and Data Mining 4, 5 (2011), 512–546.

Taxonomy

Internal Density

“Communities as a sets of densely connected entities”

Each community must have a number of edges significantly higher than what expected in a random graph



Algorithms in this family:

- Greedy Modularity, Louvain, ...

Taxonomy

Internal Density

How to assure high density?

General Idea:

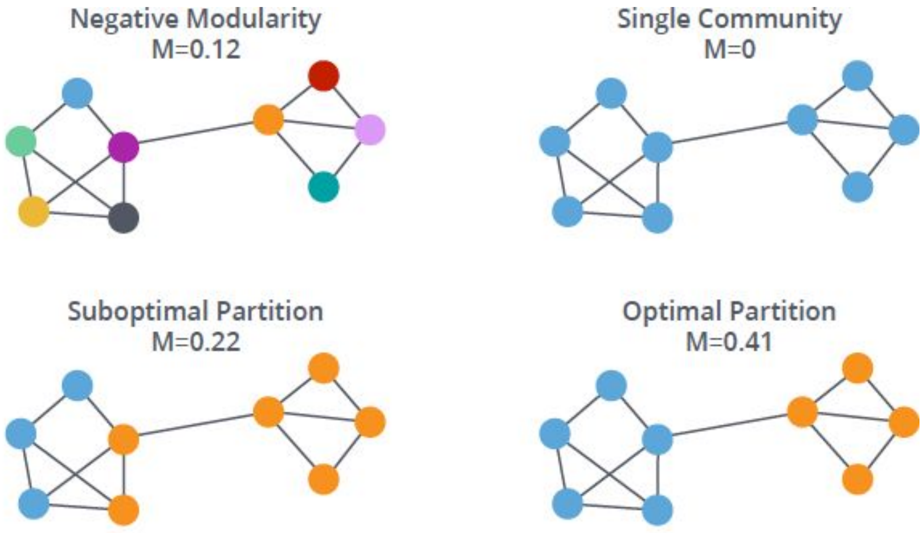
- define a quality function that measures the density of a community and then try to maximize it

Modularity [-1, 1]

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

Null Model expected density

1 if i,j in same community, 0 otherwise



Taxonomy

Louvain

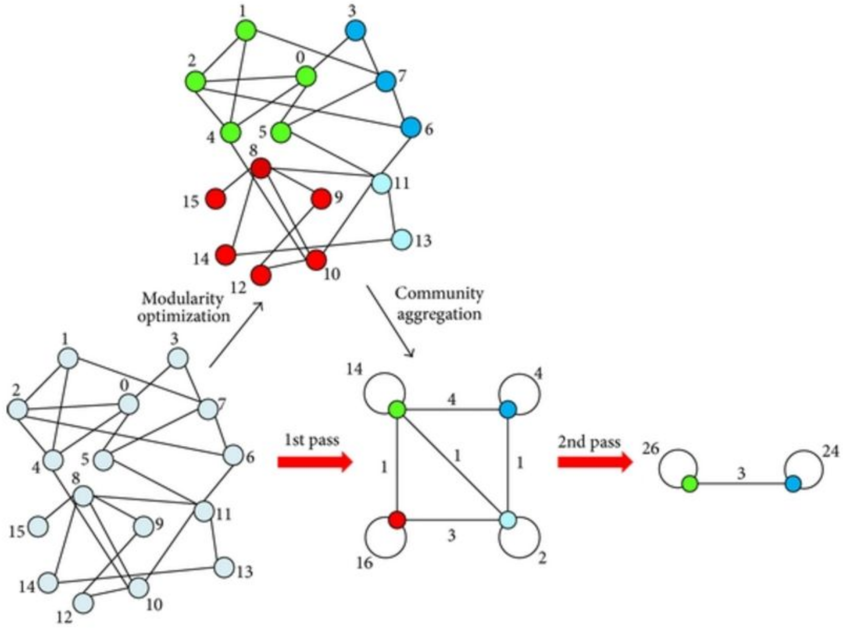
In order to maximize this value efficiently, the Louvain Method has two phases that are repeated iteratively.

Initialization:
Each node in the network is assigned to its own community.

- Phase 1:
Each node is then moved into the adjacent community that guarantee the greatest modularity increase.

- Phase 2:
A new graph is created: its nodes are the updated communities and weighted links connect them accounting for bridges in the original graph.

Phases 1 and 2 are repeated until modularity is maximized



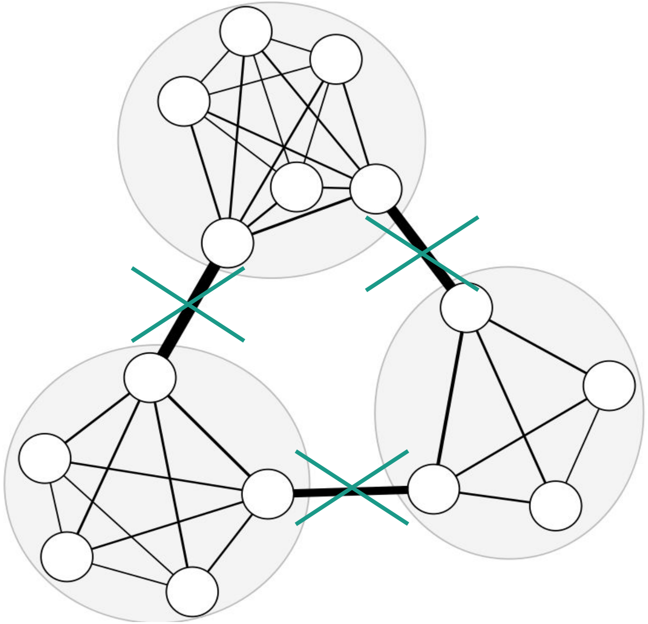
VD Blondel, et al. *Fast unfolding of communities in large networks.* Journal of statistical mechanics: theory and experiment (2008)

Taxonomy

Bridge Detection

"Communities as components of the network obtained by removing bridges"

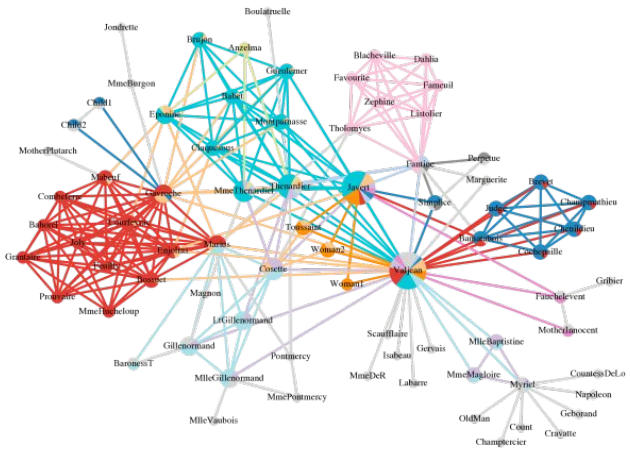
Partitioning, usually top-down, approaches



Algorithms in this family:
- Girvan Newman (edge betweenness), ...

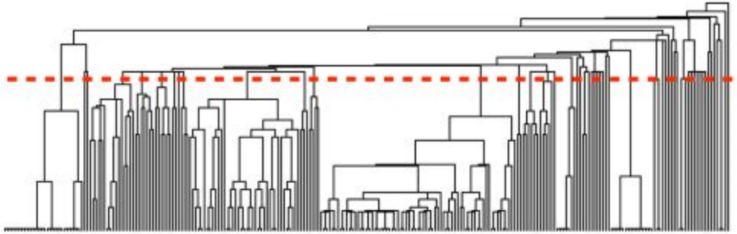
Taxonomy

Girvan-Newman



Steps

1. Compute the betweenness of all existing edges in the network;
2. Remove the edge(s) with the highest betweenness;
3. Recompute the betweenness for all edges;
4. Repeat steps 2 and 3 until no edges remain.



The end result of the Girvan–Newman algorithm is a dendrogram.

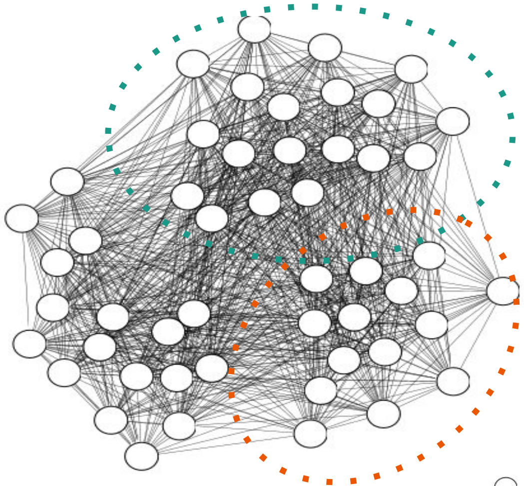
The leaves of the dendrogram are individual nodes.

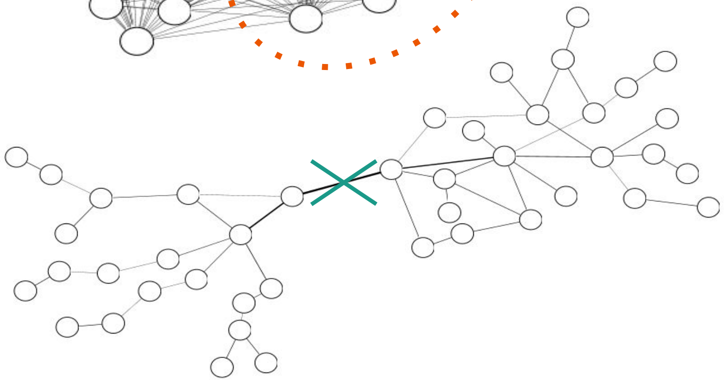
Taxonomy

Density Vs. Bridges

These two definitions seems very similar...
Are they equivalent?

- In some networks yes;
- In dense network there are no clear bridges.
- For very sparse networks a density definition will fail, even if we can detect some bridges





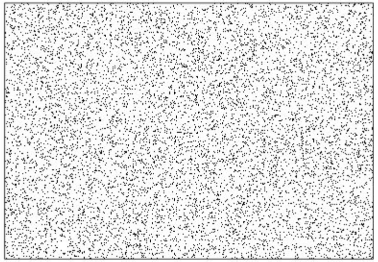
Taxonomy

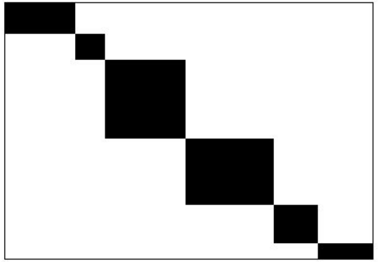
Feature Distance

“Communities as set of entities that share a precise set of features”

Once defined a distance measure based on the values of the selected node features.

The entities within a community are more similar to each other, than the ones outside the community.





Clustering approach

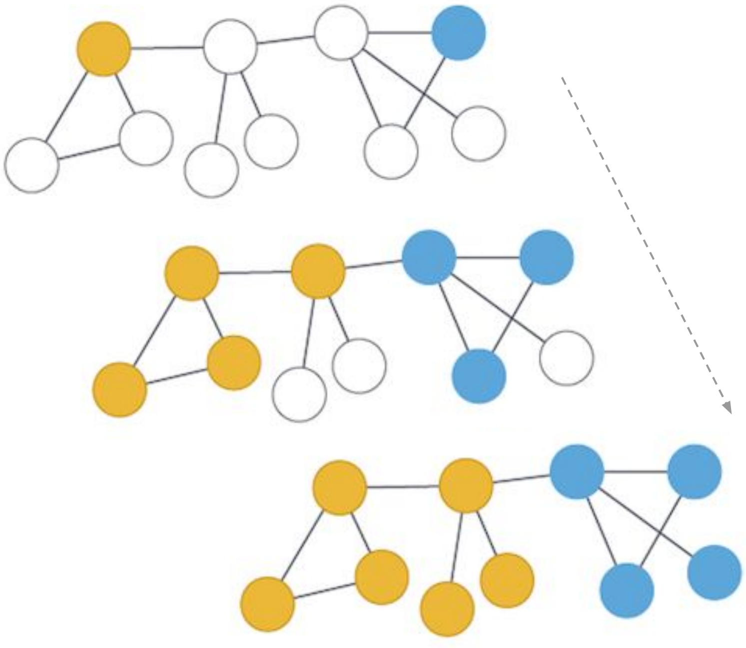
- It considers any kind of vertex features, not only their adjacencies (in the latter case we can map this definition in the density one).

Taxonomy

Percolation

"Communities as sets of nodes grouped together by the propagation of a same property, action or information"

Usually percolation approaches do not optimize an explicit quality function.



Algorithms in this family:

- Label Propagation
- Demon, Angel
- ...

Taxonomy

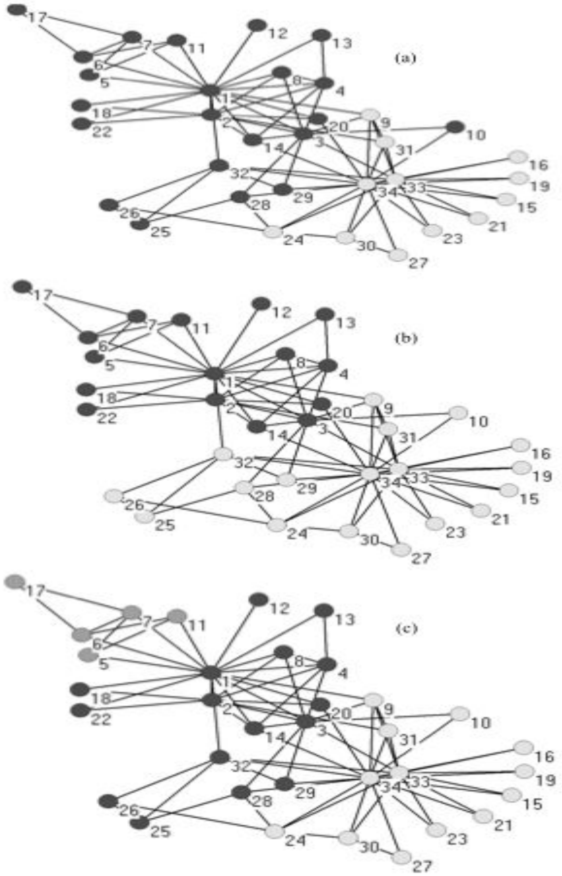
Label Propagation

1. Each node has an unique label (i.e. its id)

2. In the first (setup) iteration each node, with probability α, change its label to one of the labels of its neighbors;

3. At each subsequent iteration each node adopt as label the one shared *(at the end of the previous iteration)* by the majority of its neighbors;

4. We iterate until consensus is reached.



Taxonomy

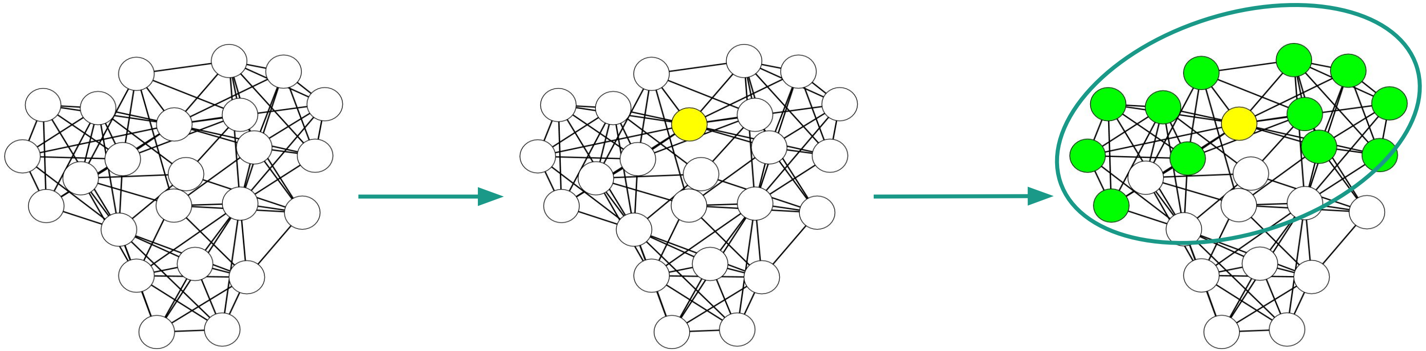
Demon/Angel

Assumptions

- Locally, each node is able to identify its communities
- Globally, we are tangled in complex overlaps

Idea:

- node-centric bottom-up approach



Taxonomy

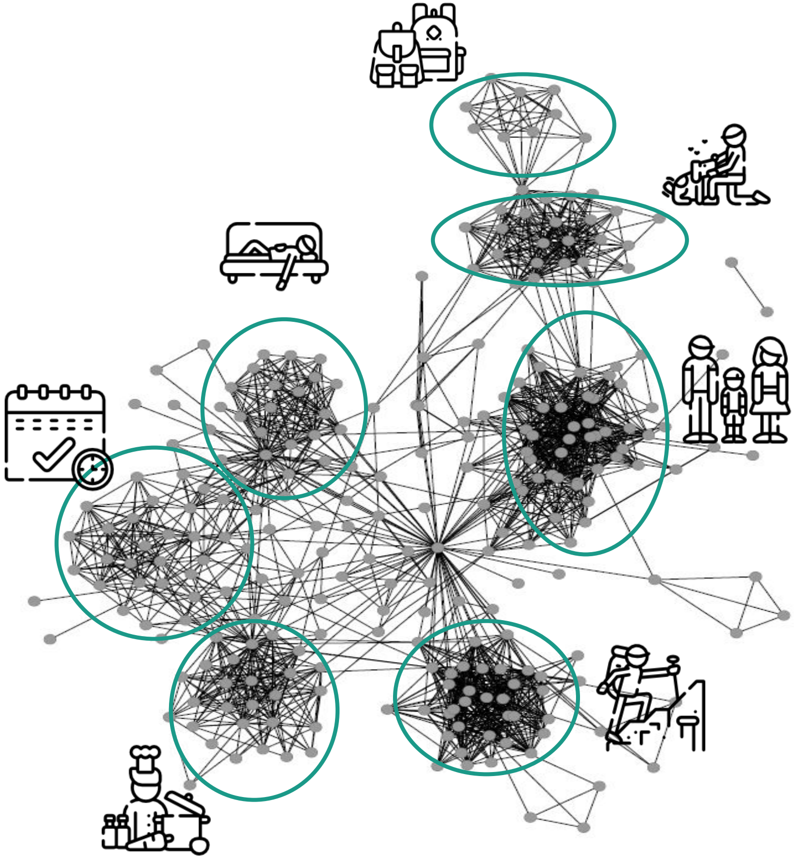
Demon/Angel

Real Networks are Complex Objects

- Can we make them “simpler”?

Ego-Networks

(networks builded upon a focal node , the "*ego*", and the nodes to whom *ego* is directly connected to plus the ties, if any, among the alters)



Taxonomy

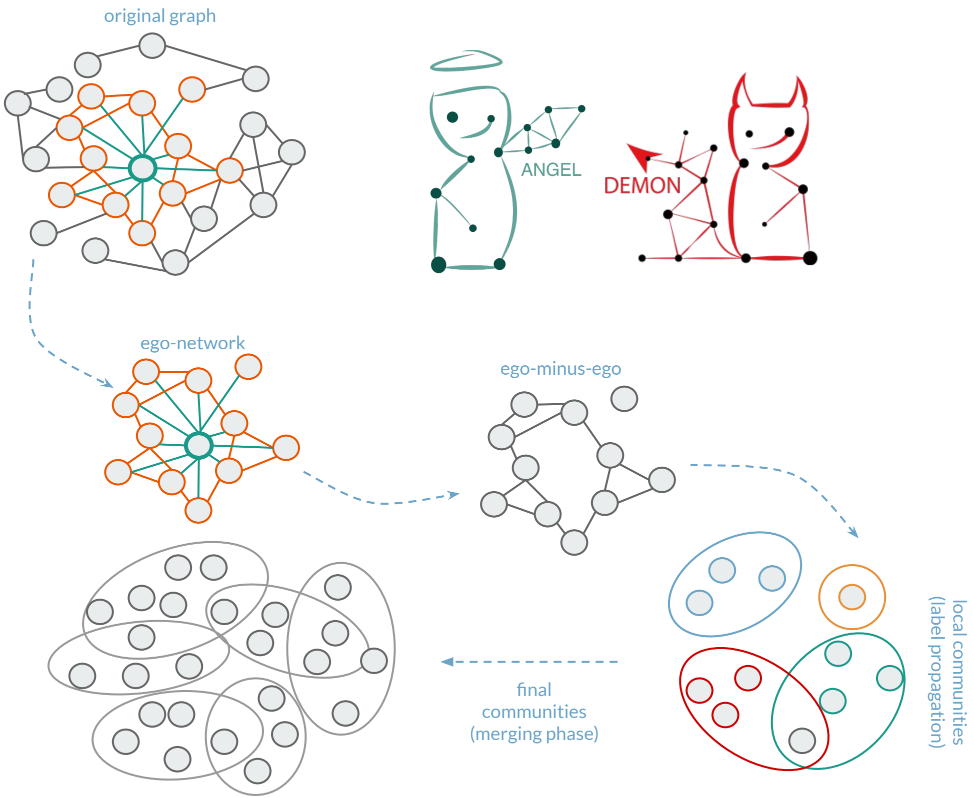
Demon/Angel

For each node n:

1. Extract the Ego Network of n
2. Remove n from the Ego Network
3. Perform a Label Propagation
4. Insert n in each community found
5. Update the raw community set C

For each local community c in C

6. Merge with “similar” ones in the set (given a threshold)
(i.e. merge iff at most the ε% of the smaller one is not included in the bigger one)



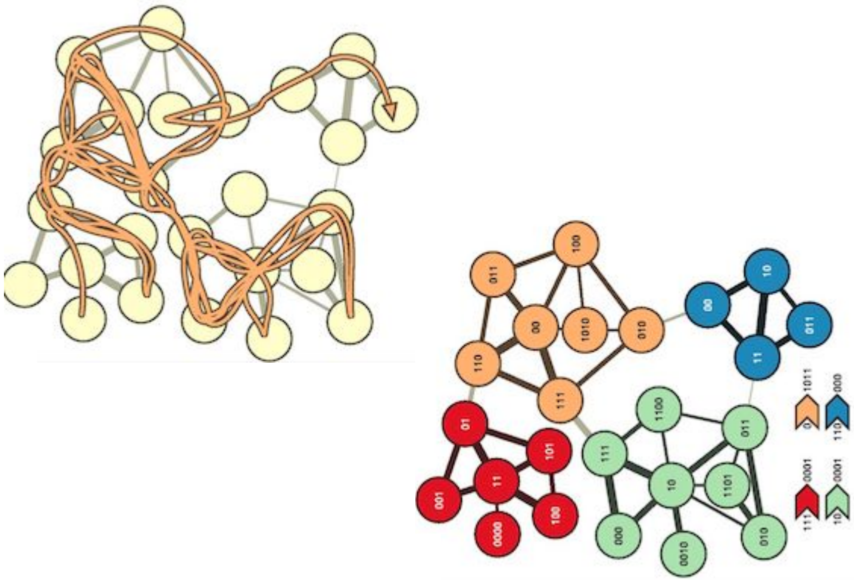
Taxonomy

Entity Closeness

“Communities as sets of nodes that can reach any member of their group crossing a very low number of edges, significantly lower than the average shortest path in the network”

Idea:

Minimize the distances among nodes, implicitly avoiding the presence of bridges within communities



Algorithms in this family:

- Infomap (Conductance Optimization)
- ...

Taxonomy

Infomap

The core of the algorithm follows closely the Louvain method:

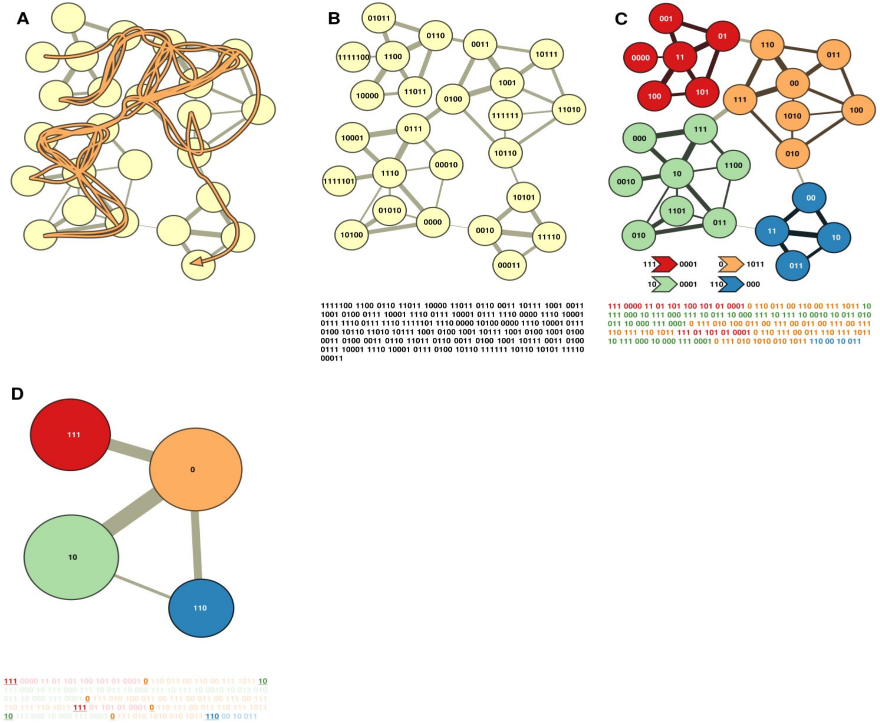
- Phase 1:
  Each node is moved to the neighboring module that results in the largest decrease of the map equation.

- Phase 2:
  The network is rebuilt, with the modules of the last level forming the nodes at this level.
  This hierarchical rebuilding of the network is repeated until the map equation cannot be reduced further.

Implicit optimization of the Conductance measure: $\phi(G) = \min_{S \subseteq V} \varphi(S)$

Where:

- $\varphi(S) = \dfrac{\sum_{i \in S, j \in \bar{S}} a_{ij}}{\min(a(S), a(\bar{S}))}$ is the conductance for a cut
- $(S, \bar{S})$ is a cut, and
- $a(S) = \sum_{i \in S} \sum_{j \in V} a_{ij}$



Taxonomy

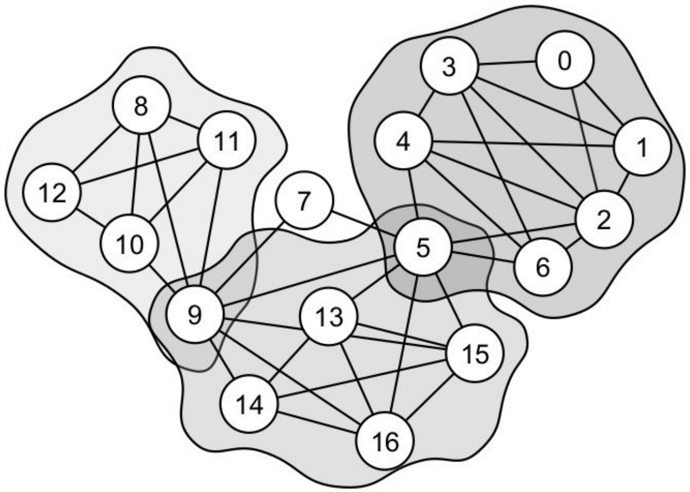
Structure Definition

“Communities as sets of nodes having a precise number of edges among them, distributed in a precise topology defined by a number of rules”

Idea:

Identify precise patterns within a network

(e.g., cliques, quasi-cliques, ...)



Algorithms in this family:

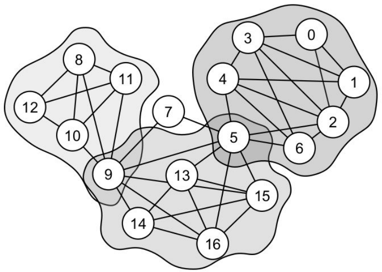
- k-cliques, ...

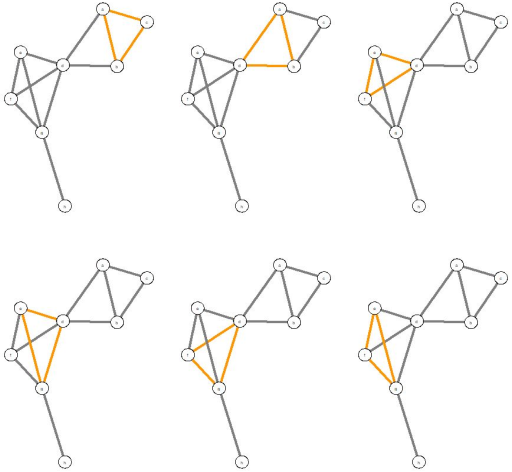
Taxonomy

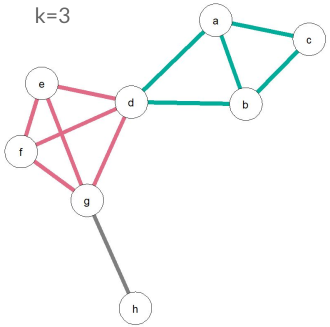
k-Cliques

A very popular algorithm: **k-cliques**

- Also this case is different from the density definition: node 7 is in some sense “dense” (is in a triangle), but outside of any community







Algorithm steps:

1. Identify *k-cliques*, which are fully connected networks with k nodes. (The smallest possible k would be k = 3. Otherwise, the cliques would be only edges.)

2. A community is defined as a set of adjacent k-cliques, that is, k-cliques that share exactly *k-1* nodes.
   With k = 3, two 3-cliques are adjacent if they share exactly two nodes (equivalent to an edge).

Taxonomy

Link Communities

“Communities as sets of links clustered together since they belong to a particular relational environment”

Links and their relations are used to identify communities:

- the links endpoints identify the induced nodes communities

Taxonomy

No a priori definition

“Communities as sets of nodes that shares a particular set of features (not necessarily topology related) as defined by an analyst”

Category often used to group approaches that leverage specific peculiarities of complex networks instances

(e.g., time, multi-layers, high-order,...)

Community Discovery

Peculiar Topologies and Explicit Semantics

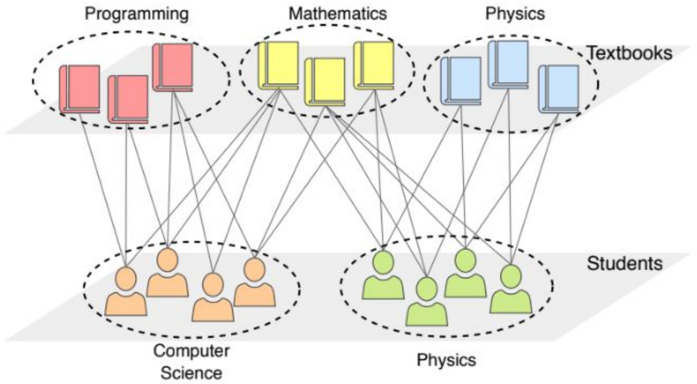
Bipartite & Directed Networks

So far we assumed networks to be simple, undirected and (mostly) unweighted.

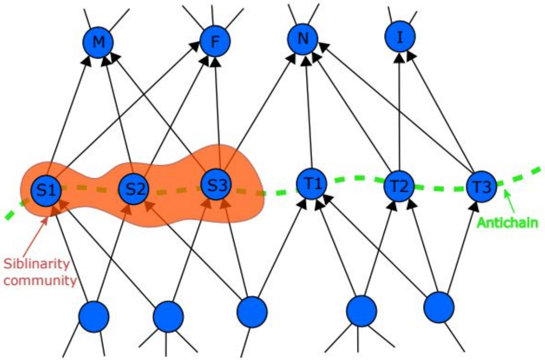
In presence of more complex topologies alternative strategies can be applied and communities become something different

Examples:

- Antichains, Sibilinary Communities (DAG)
- One-to-One, Many-to-One (bipartite)



Taguchi, Hibiki, and Tsuyoshi Murata. *"BiMLPA : Community Detection in Bipartite Networks by Multi-Label Propagation."* NetSciX (2020).



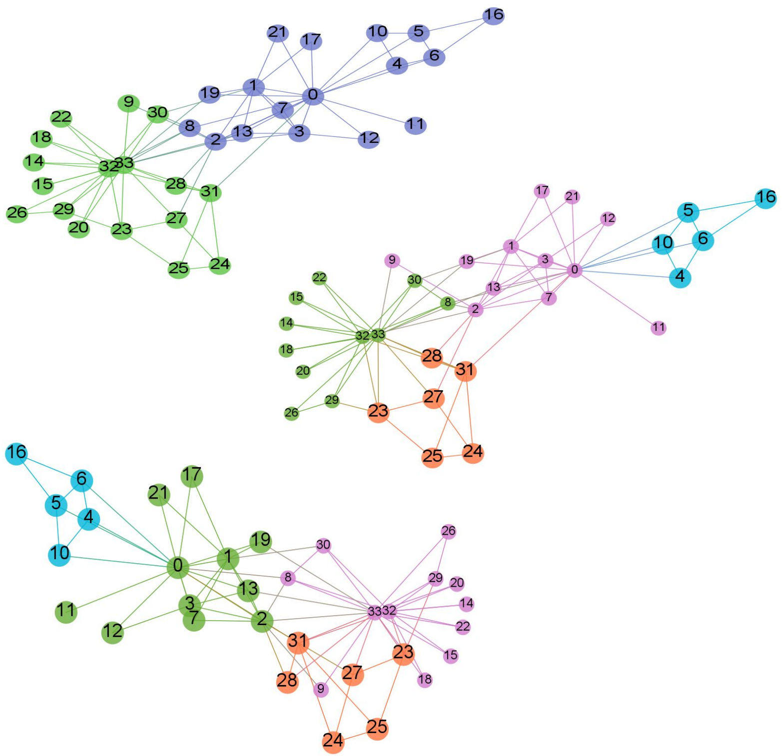
Vasiliauskaite, Vaiva, and Tim S. Evans. "*Making Communities Show Respect for Order.*" *arXiv preprint arXiv:1908.11818* (2019).

Attributed Networks

Nodes and edges can be characterized by additional semantic layers

- e.g., age, nationality, education...

A meaningful partition (segmentation) needs to be both topologically and semantically consistent.



Citraro, Salvatore, and Giulio Rossetti. *"Eva: Attribute-Aware Network Segmentation." Complex Networks and Their Applications* (2019).

Community Discovery

Evaluation strategies

Strategies

Internal Evaluation

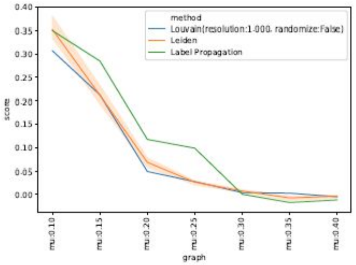
- Partition quality function
  (i.e., modularity, conductance, density…)

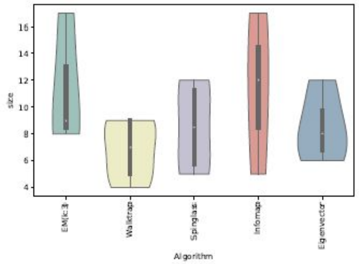
- Community characterization
  (i.e., size distribution, overlap distribution…)

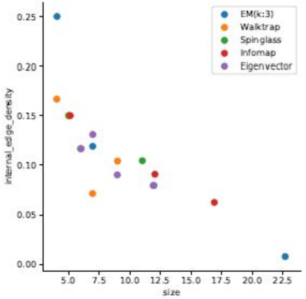
- Execution time and Complexity

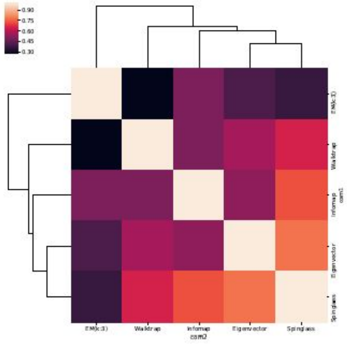
External Evaluation

- Ground truth testing (or partitions comparison)







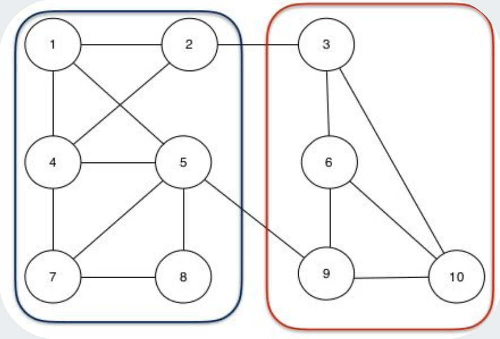


Internal Evaluation

Quality Functions

Several fitness functions can be defined to assess the quality of a partition.

Usually, the best partition is the one that maximize (or minimize) a given fitness function in its worst case scenario (i.e., when computed on the worst community identified)



Approx. formulae
(for exercise only)

Internal Edge Density

$$\frac{2|E_C|}{|V_C|(|V_C|-1)}$$

E_C edges within C
V_C nodes within C

Worst case: **min**

Best-worst case: **max**

Average Node Degree

$$\frac{1}{|V_C|}\sum_{i \in C} d_i$$

d_i degree of node i

Worst case: **min**

Best-worst case: **max**

Modularity

$$\left(\frac{|V_c|}{|E|} - \frac{\deg_C}{2|E|}\right)^2$$

$\deg_C$ sum of degrees within C

$$\deg_C = \sum_{i \in C} d_i$$

Worst case: **min**

Best-worst case: **max**

Conductance

$$\frac{2|E_{OC}|}{2|E_C| + |E_{OC}|}$$

E_{OC} edges out of C

Worst case: **max**

Best-worst case: **min**

Yang, Jaewon, and Jure Leskovec. "*Defining and evaluating network communities based on ground-truth.*" Knowledge and Information Systems 42.1 (2015): 181-213.

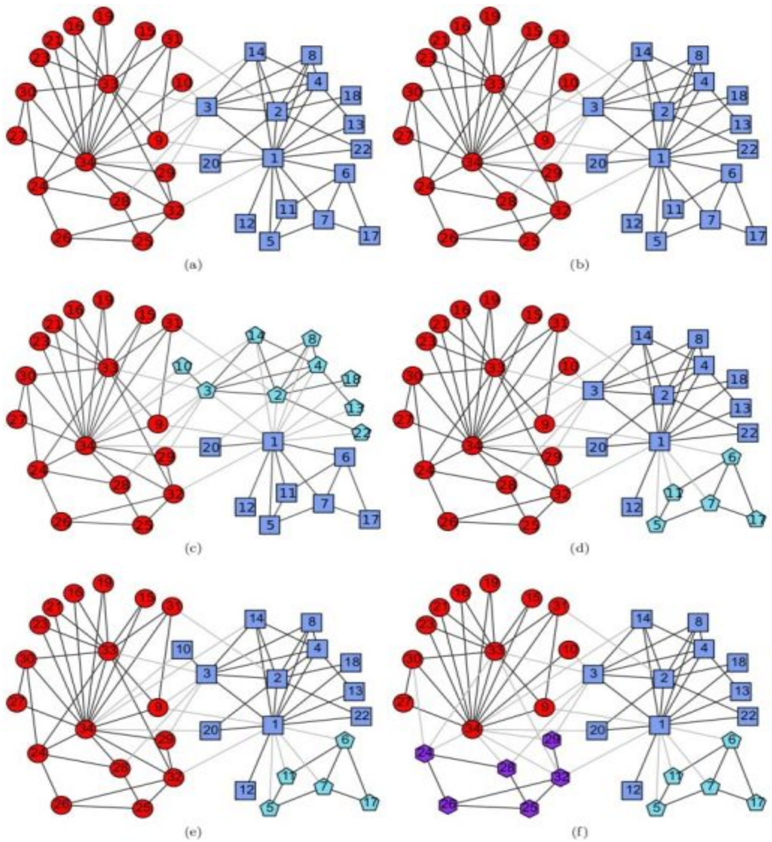
External Evaluation

Ground truth testing

Given a graph G, a ground truth partition P(G) and the set of identified communities C estimate the resemblance the latter has with P(G).

General Criticism(s)

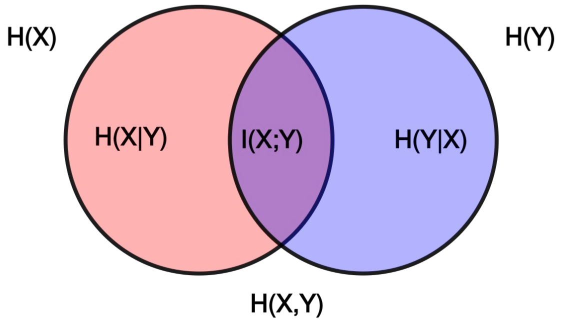
- Different approaches generates communities following different criteria (“ill posed” problem”)
- It is not necessarily true that the ground truth represent the only valid semantic\topologic partition for the analyzed graph.



Peel, et a. "*The ground truth about metadata and community detection in networks.*" *Science advances* 3.5 (2017): e1602548.

External Evaluation

NMI



Normalized Mutual Information is a measure of *similarity* borrowed from information theory:

$$NMI(X,Y) = \frac{H(X)+H(Y)-H(X,Y)}{\frac{H(X)+H(Y)}{2}} \in [0,1]$$

- *H(X)* is the entropy of the random variable *X* associated to an identified community,
- *H(Y)* is the entropy of the random variable *Y* associated to a ground truth community,
- *H(X,Y)* is the joint entropy.

The higher the NMI the more similar the compared partitions are

Advantages

- Extensively used in literature

Drawbacks

- Computational complexity ~ $O(|C|^2)$ (where C is the community set)
- Needs to be approximated in case of overlapping partitions

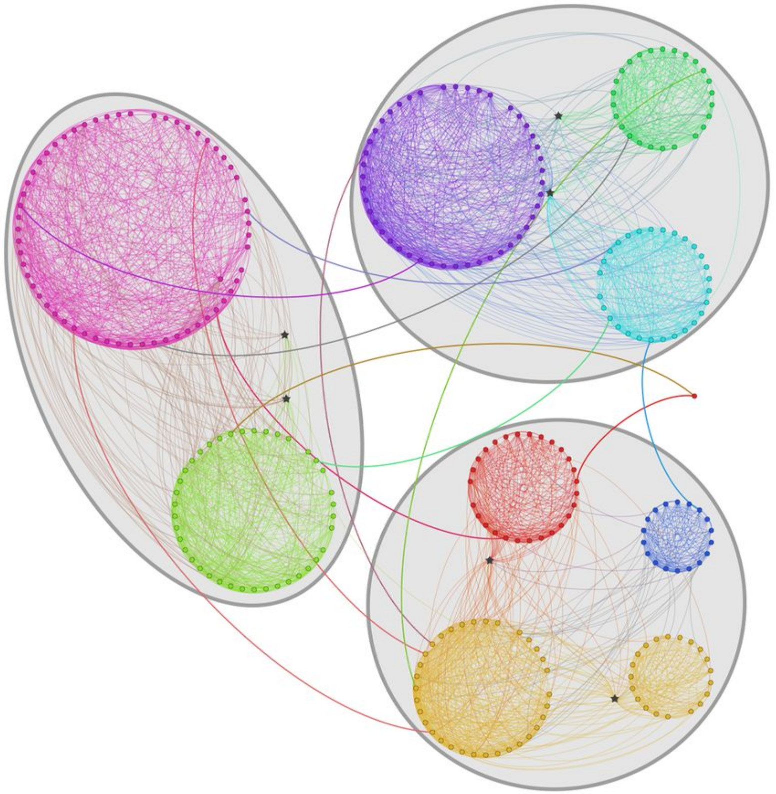
External Evaluation

Synthetic Benchmarks

Testing against topological ground truths

Synthetic graphs with embedded community structure (e.g., LFR)

- More stable than semantic ground truth partitions
- Community structure depends on the fitness function optimized by the chosen model
- Approximation of real world networks



Lancichinetti, Andrea, Santo Fortunato, and Filippo Radicchi. "*Benchmark graphs for testing community detection algorithms.*" *Physical review E* 78.4 (2008): 046110.

Summarizing

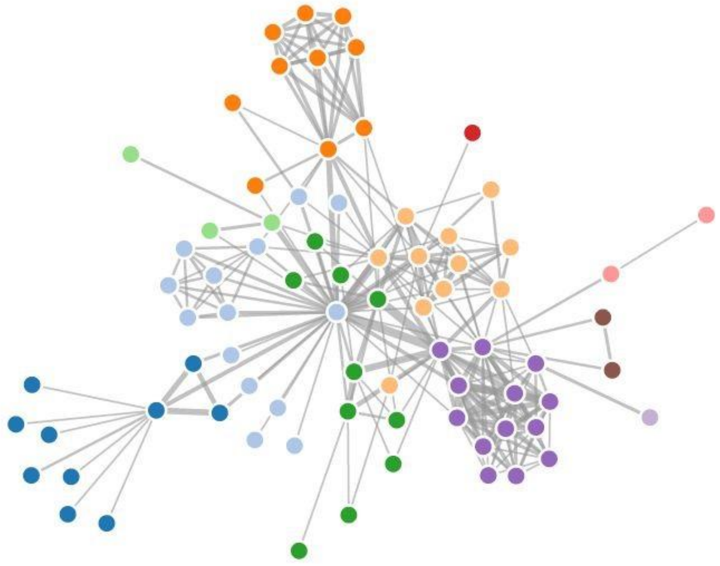
Community Discovery is, perhaps, the hottest topic in complex network analysis

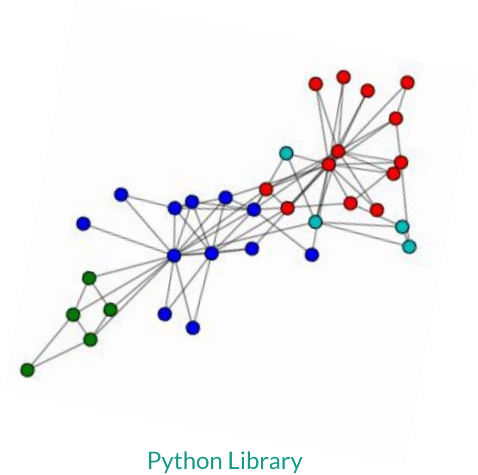
Major issues:

- Problem definition
- Community evaluation

Problem specializations:

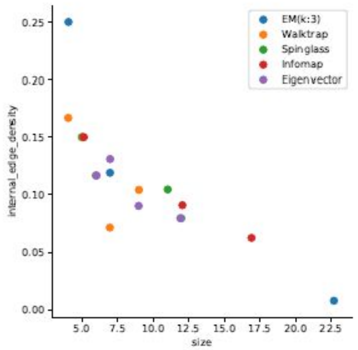
- Evolutionary Community Discovery
  (How do communities evolve in dynamic networks?)
- Multidimensional Community Discovery
- …





CD·LIB

"One Library to Rule them All"



Python Library

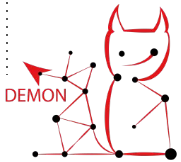
python™ Package Index

pip install cdlib

Algorithms

48

Crisp, Overlapping, Fuzzy, Attributed, Bipartite Community Discovery



Evaluation

40

Clustering quality & Comparison functions

https://cdlib.rtfd.io

Rossetti et al. "*CDLIB: a python library to extract, compare and evaluate communities from complex networks.*" *Applied Network Science* (2019).

https://bit.ly/2OO9tK7

Chapter 8

Conclusion

Take Away Messages

1. Complex networks are composed by hidden meso-scale structures
2. Identify them is not a trivial task
3. Evaluate them is not a trivial task
4. Knowing them is fundamental to identify functional modules of a system

Suggested Readings

- Chapter 9 of Barabasi's book
- Fortunato's survey
- Cited papers

What's Next

Chapter 9:
Dynamic Of Networks