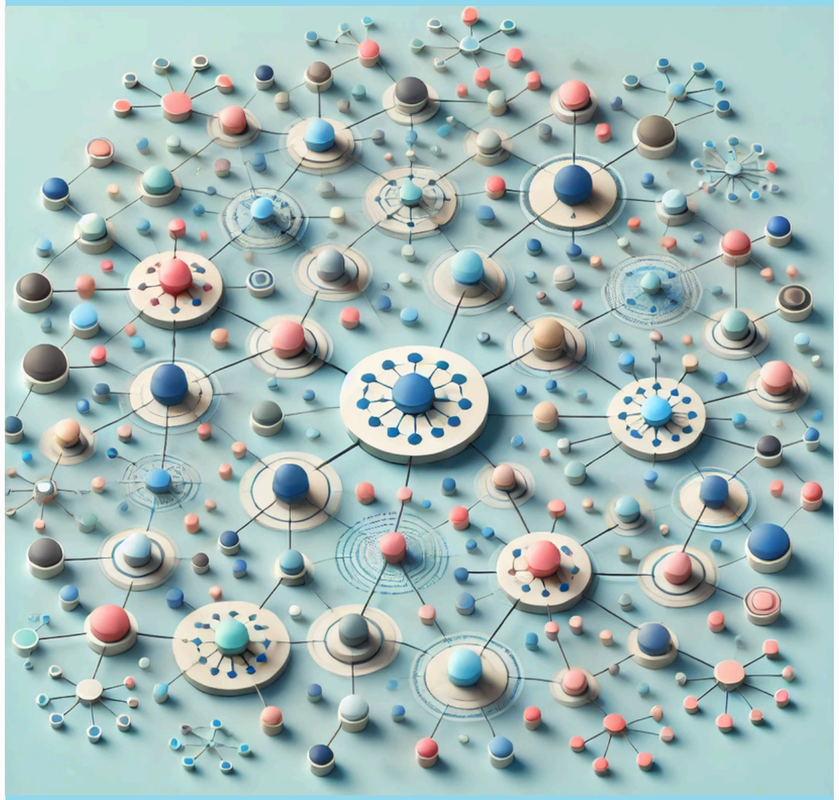


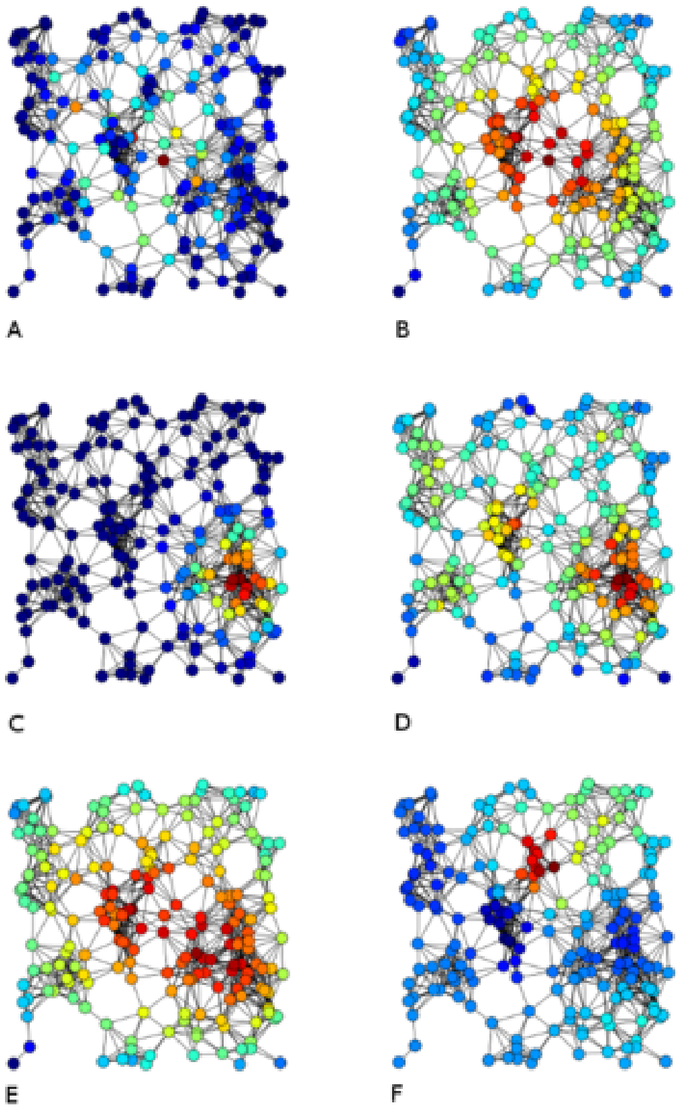
Universität Konstanz

UNIVERSITÄT KONSTANZ

Communities in Networks

Network Science of Socio-Economic Systems

Giordano De Marzo



Recap

The Quest for Online Search Engines

The exponential growth of the number of pages made standard approaches to searches unfeasible

The PageRank

Google introduced the PageRank, focusing of the role of pages within the network

Centrality Measures

Different tasks require different centrality measures

Analyzing Criminal Networks

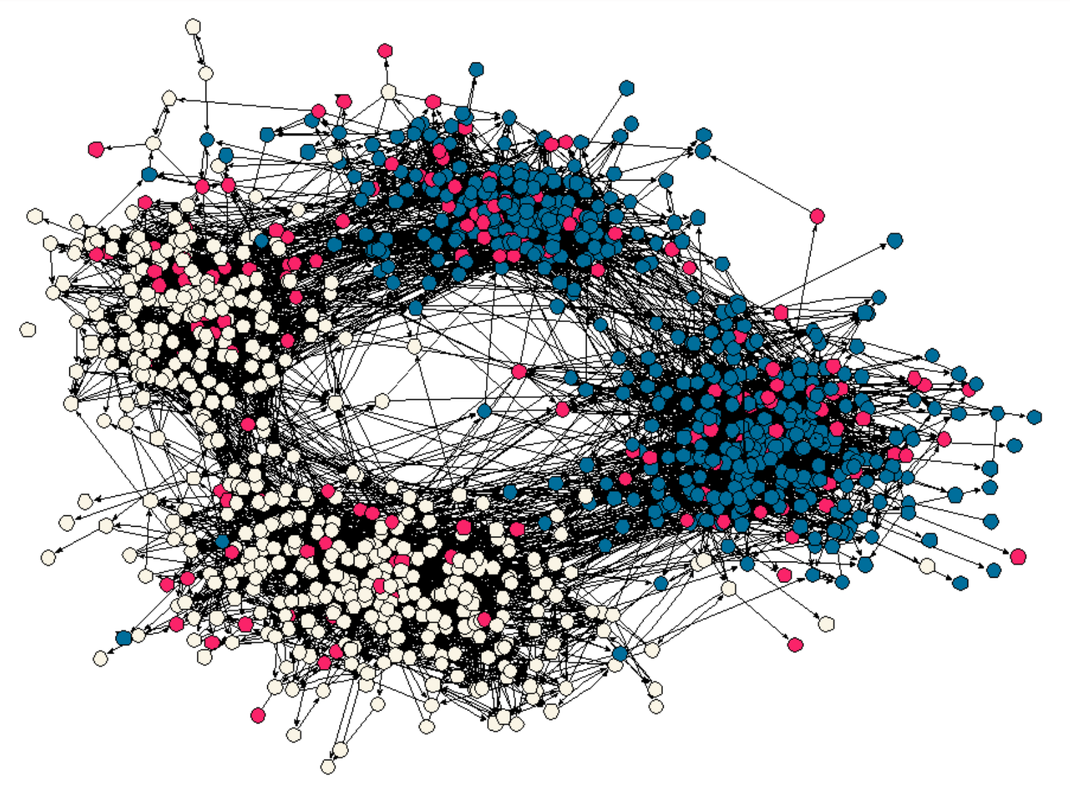
Network science is a useful tool to analyze and target criminal networks

Outline

1. Communities in Networks
2. Community Detection Algorithms
3. Homophily and Communities Formation
4. The Strength of Weak Ties

Communities in Networks

Communities in Networks



Communities are groups of nodes densely connected internally but sparsely connected with the rest of the network. They are the result of

- Clustering: Nodes tend to form tightly-knit groups.
- Homophily: Similar nodes are more likely to connect, reflecting shared characteristics or interests.

Examples include:

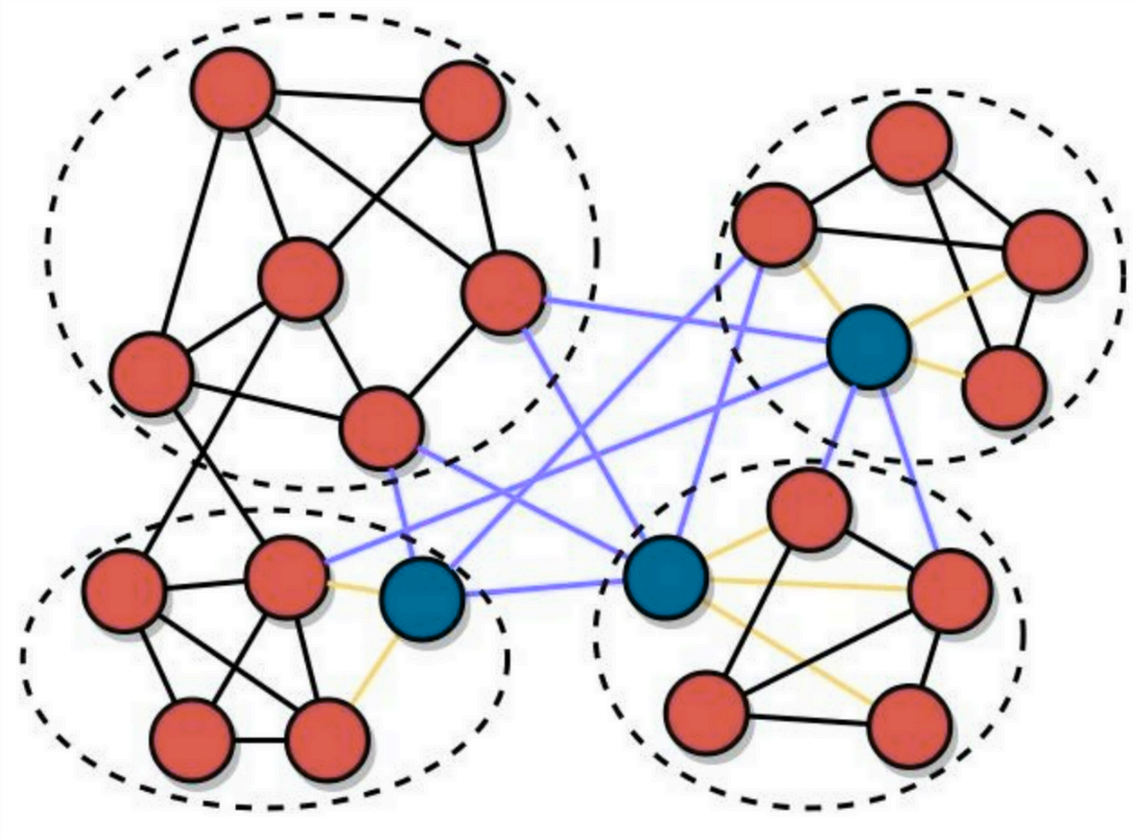
- Social circles in friendship networks.
- Research groups in co-authorship networks.
- Modules in biological networks

Strong vs Weak Communities

Communities can be strong or weak

- in a strong community each node has more connections within the community than with nodes outside its community
- in a weak community, only the total number of connections within a community is larger than the number of connections outside the community

The communities with the blue nodes are weak

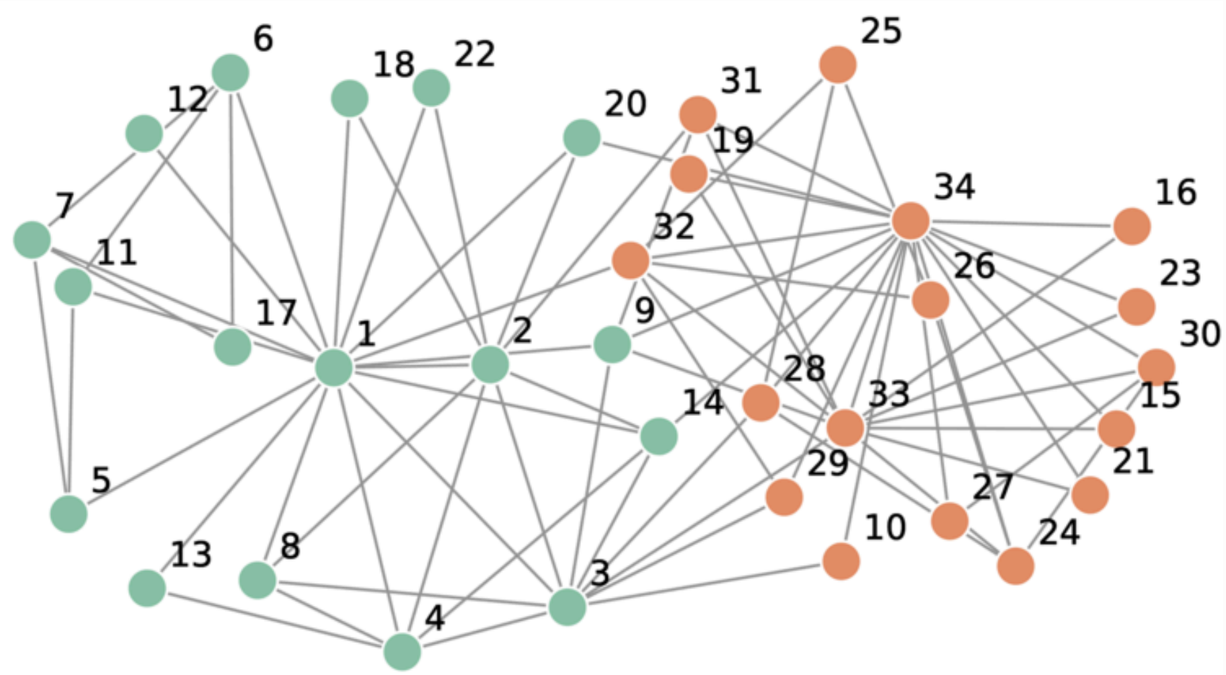


Community Detection

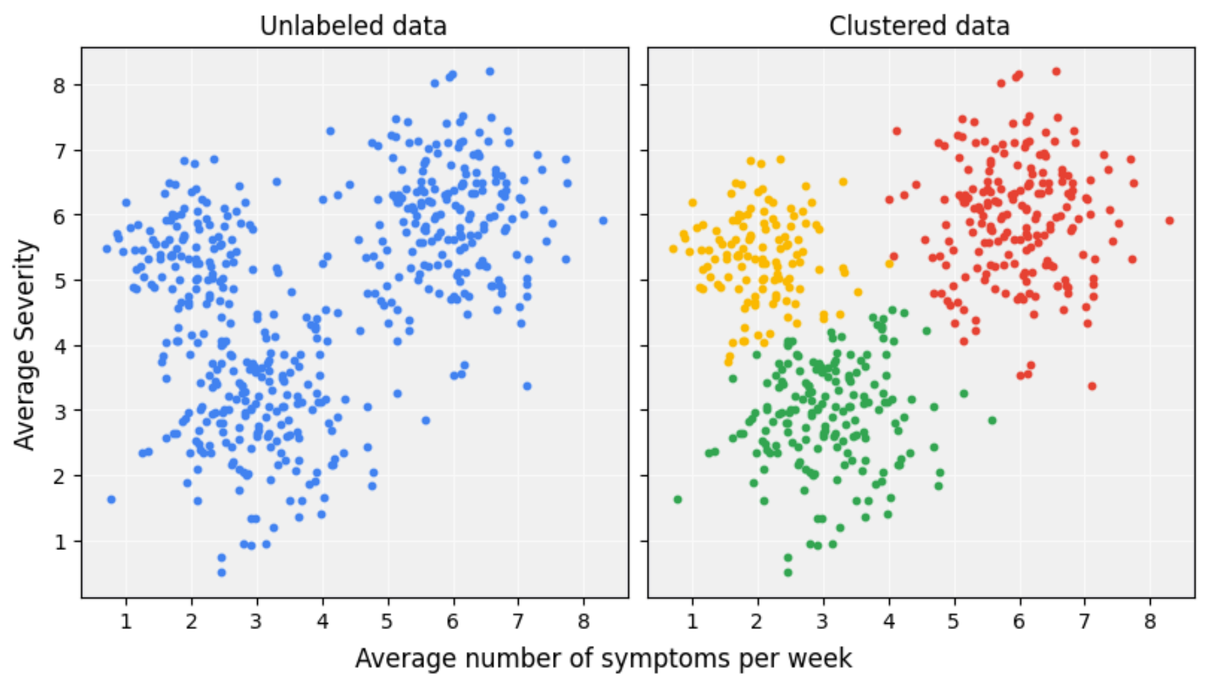
Community Detection consists in automatically partition a network into communities based on its structure

This task can be very challenging

- Communities can overlap (e.g., in social networks, people belong to multiple groups).
- Networks often lack clear boundaries.
- Real-world networks are large and complex (e.g., millions of nodes, billions of edges).



Community Detection vs Clustering



Community detection is conceptually similar to clustering

- Community Detection:
  - Operates on networks where relationships are non-Euclidean.
  - Focuses on identifying densely connected groups of nodes.
- Clustering:
  - Operates on metrical spaces.
  - Groups points based on distance or similarity metrics.

What is the Quality of a Partition?

A network can be partitioned in many ways, but not all partitions are meaningful. We need methods to evaluate the goodness of a partition
There are several approaches to measure quality:

1. Edge Density:
   - Compare the number of edges inside communities to those between communities.
   - Stronger communities have higher internal edge density.
2. Cut-based Metrics:
   - Evaluate partitions based on the number of edges cut between communities.
3. Comparison with Null Models:
   - Assess partitions by comparing them to random network models.
   - Null Hypothesis: Communities arise purely by chance.
4. Objective Functions:
   - Define and optimize a mathematical function (e.g., modularity) to find the "best" partition.

Newman's Modularity

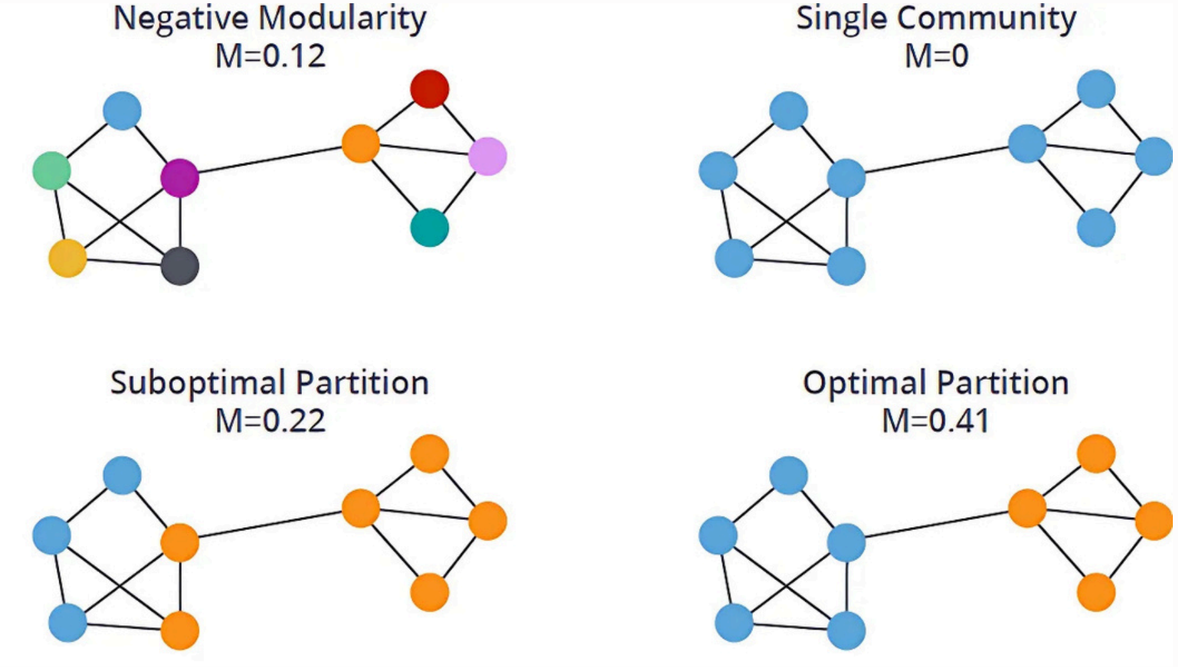
Modularity Q is a measure of the quality of a division of a network into communities

- Evaluates how well the network is partitioned
- Compares the density of edges within communities to what would be expected in a random network

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

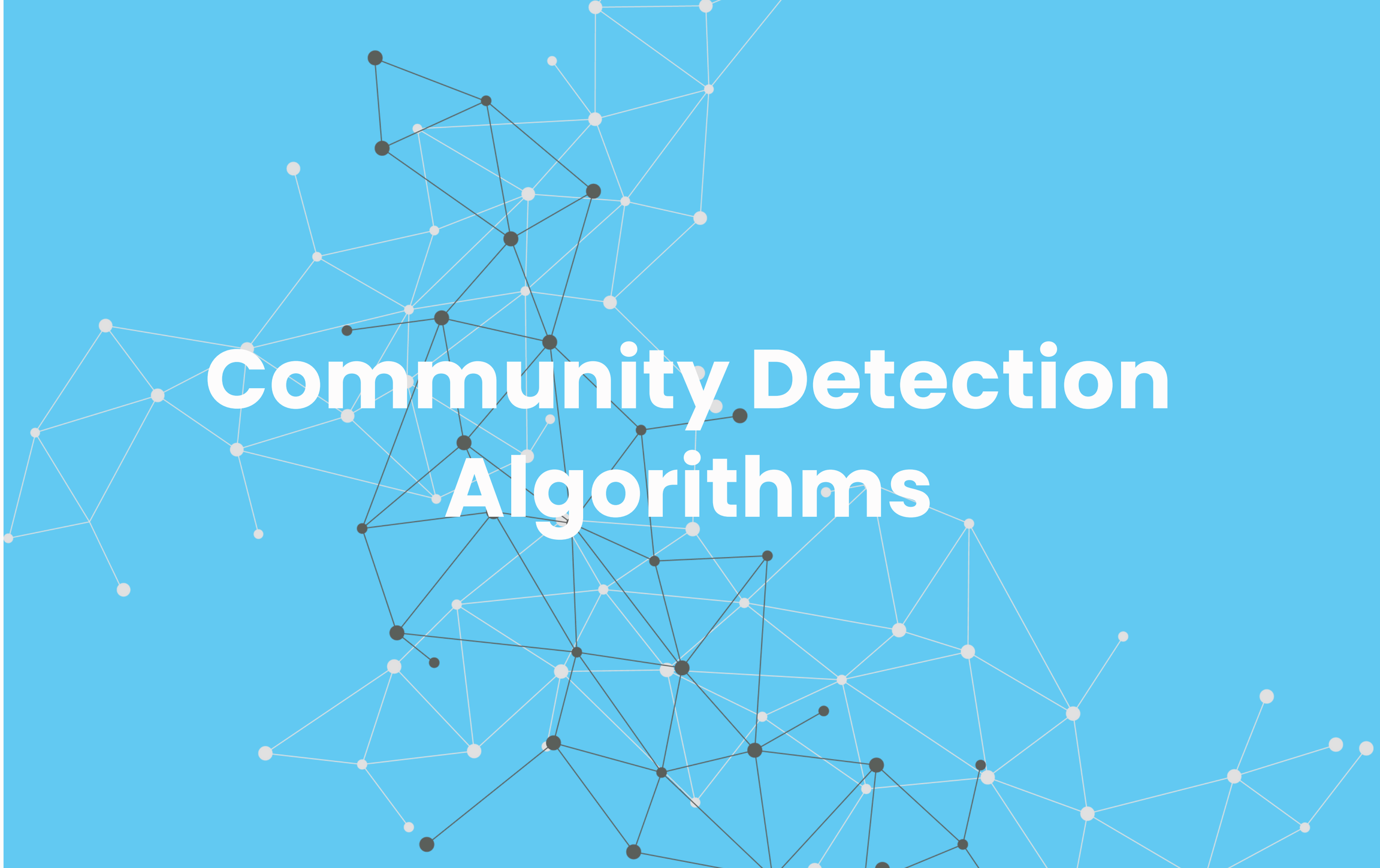
- A_{ij}: Adjacency matrix (1 if nodes i and j are connected, 0 otherwise).
- k_i: Degree of node i ($k_i = \sum_j A_{ij}$).
- m: Total number of edges in the network ($m = \frac{1}{2} \sum_{ij} A_{ij}$).
- c_i, c_j: Community assignments of nodes i and j.
- $\delta(c_i, c_j)$: Kronecker delta (1 if $c_i = c_j$, 0 otherwise).

Understanding the Modularity



Higher values of Q indicate better-defined community structures.

1. Negative Modularity (Q<0)
   - Poor partitioning where nodes are incorrectly grouped
2. Single Community (Q=0):
   - Entire network treated as one community
3. Suboptimal Partition (Q>0):
   - A reasonable division, but non optimal
4. Optimal Partition (Q is max):
   - A clear and well-defined community structure

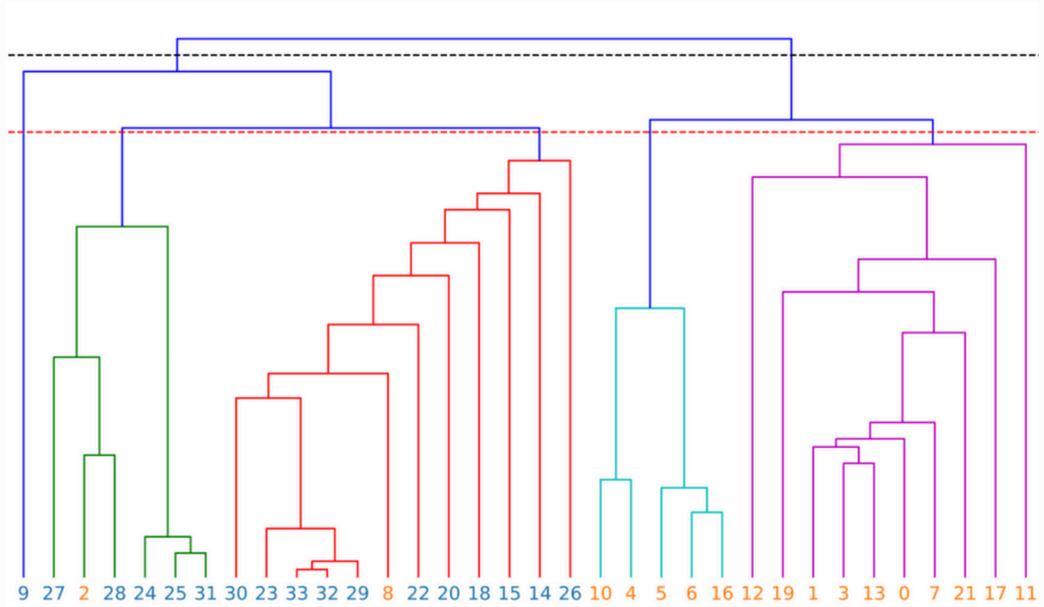


Community Detection Algorithms

Girvan-Newman Algorithm

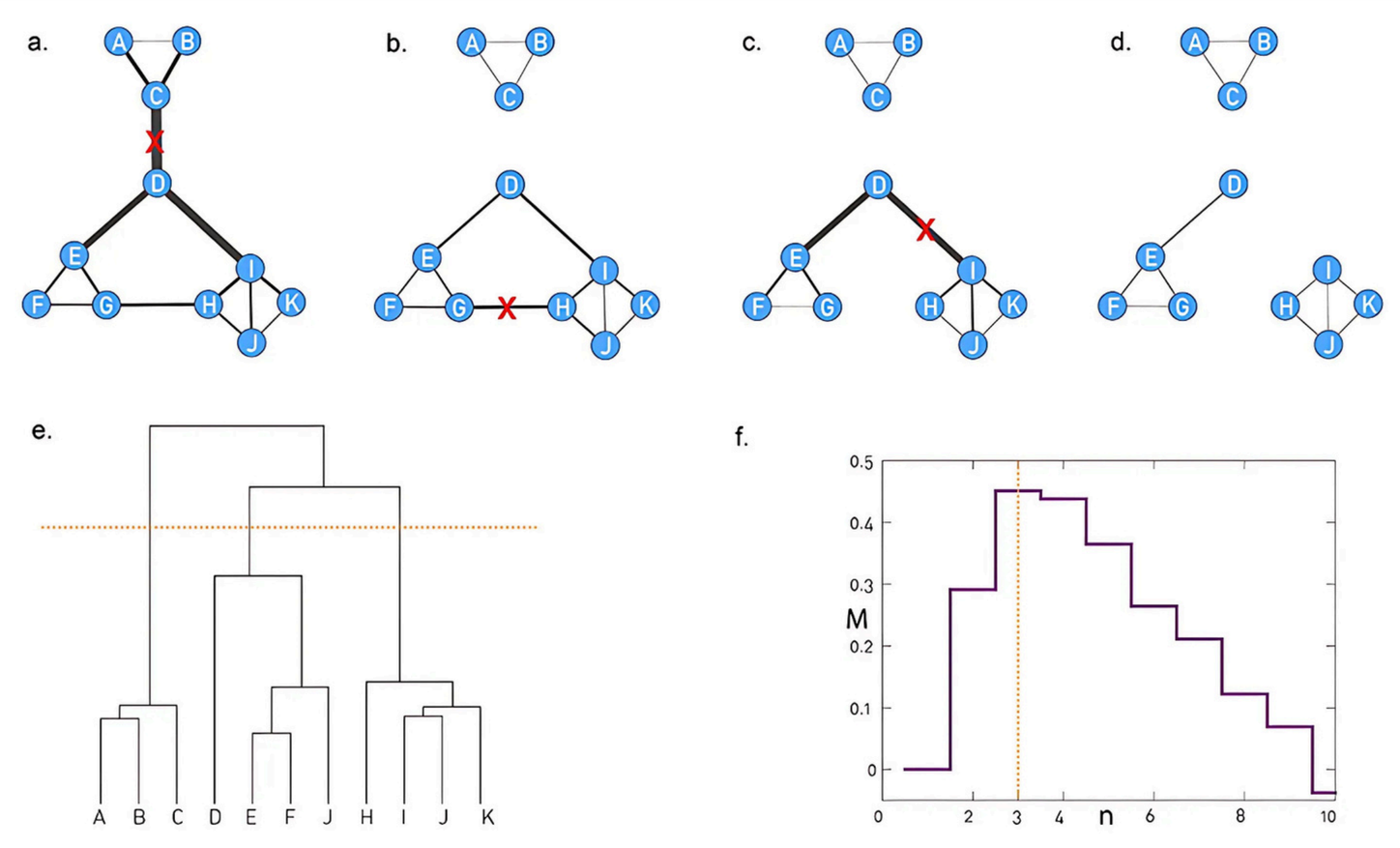
The Girvan-Newman algorithm is a popular method for community detection.

- It identifies communities by iteratively removing edges with the highest betweenness centrality
  a. Compute the betweenness centrality for all edges in the network.
  b. Remove the edge with the highest betweenness.
  c. Recompute betweenness and repeat until the network breaks into communities.
- The process returns a dendrogram



Girvan-Newman Visualized

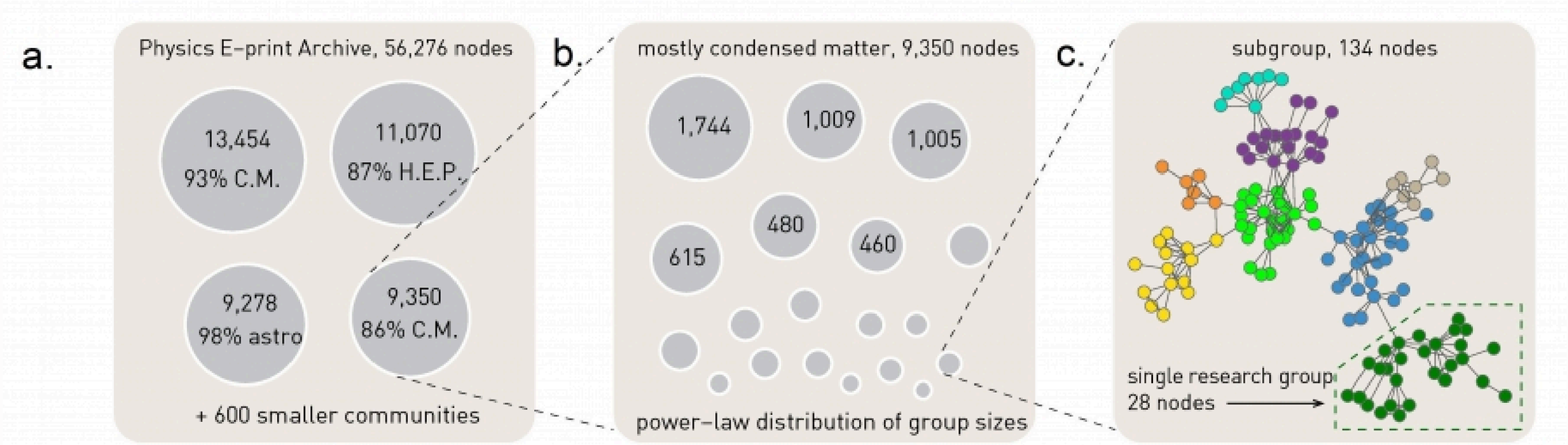
We can use modularity to cut the dendrogram getting the best partition



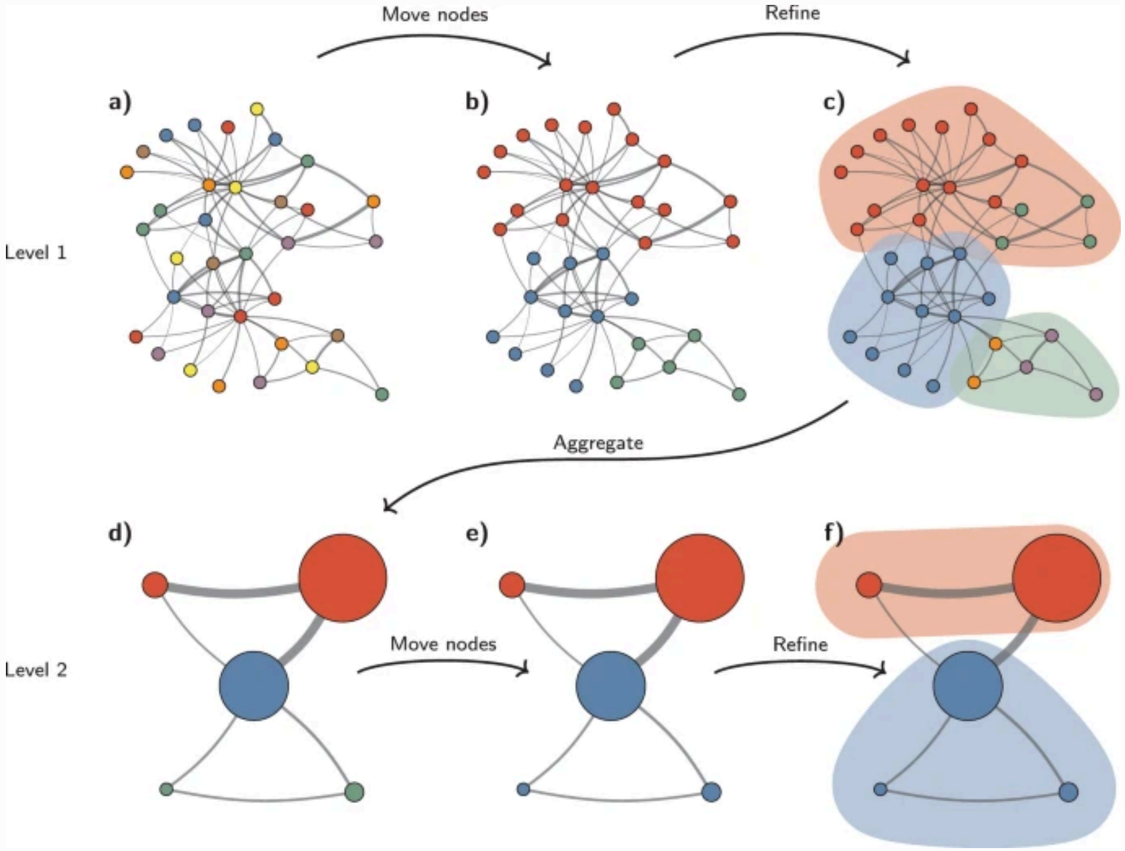
Greedy Modularity Maximization

The Greed Modularity Maximization algorithm starts with each node as its own community and iteratively merge communities to improve modularity.

1. Assign each node to its own community.
2. Merge the pair of communities that results in the largest increase in modularity (ΔM).
3. Repeat until no further improvement in modularity is possible.



Louvain and Leiden Algorithms



Louvain clustering is a widely used method for community detection that works by iteratively optimizing modularity. N

- Nodes are moved between communities to maximize modularity
- The communities are aggregated into "meta-nodes"
- This process repeats until no further modularity improvement is possible.

Leiden clustering builds on Louvain by introducing refinement steps to improve the quality of the partition.

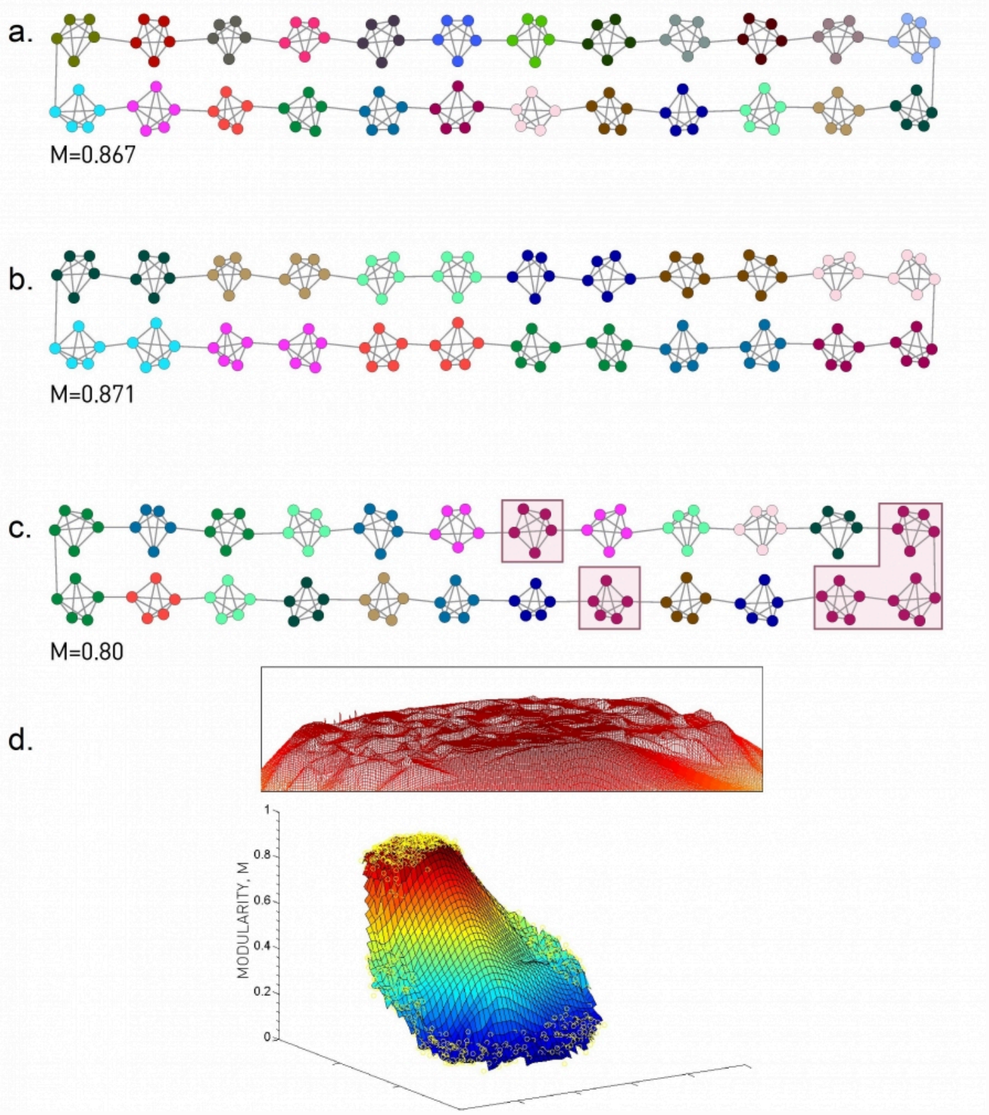
- It ensures that communities are well-connected and locally optimal at every step

Limits of Modularity

Networks with clear community structures should have an optimal partition with maximal modularity

- Other partitions should be distinguishable from this maximum
- In reality many partitions have modularity values close to the maximum.
- Even random partitions or random networks can have high modularity values

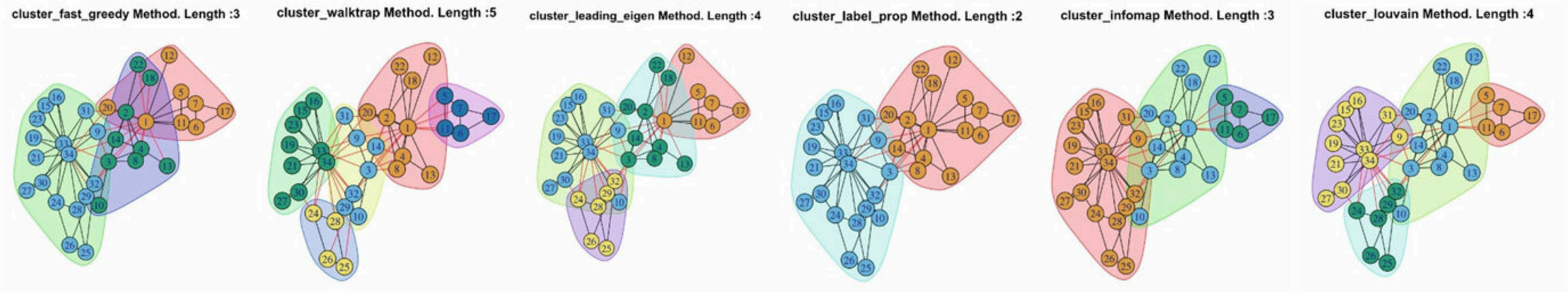
Modularity offers a useful framework for understanding community structure but has also notable limitations



Comparison of Different Algorithms

Different community detection algorithms tend to produce different partiotions

- in many cases there are strong similarities
- however there can be strong variations, also in the number of communities
- often nodes that are treated differently from different algorithms act as bridges
- it’s good practice to test more than a single algorithm

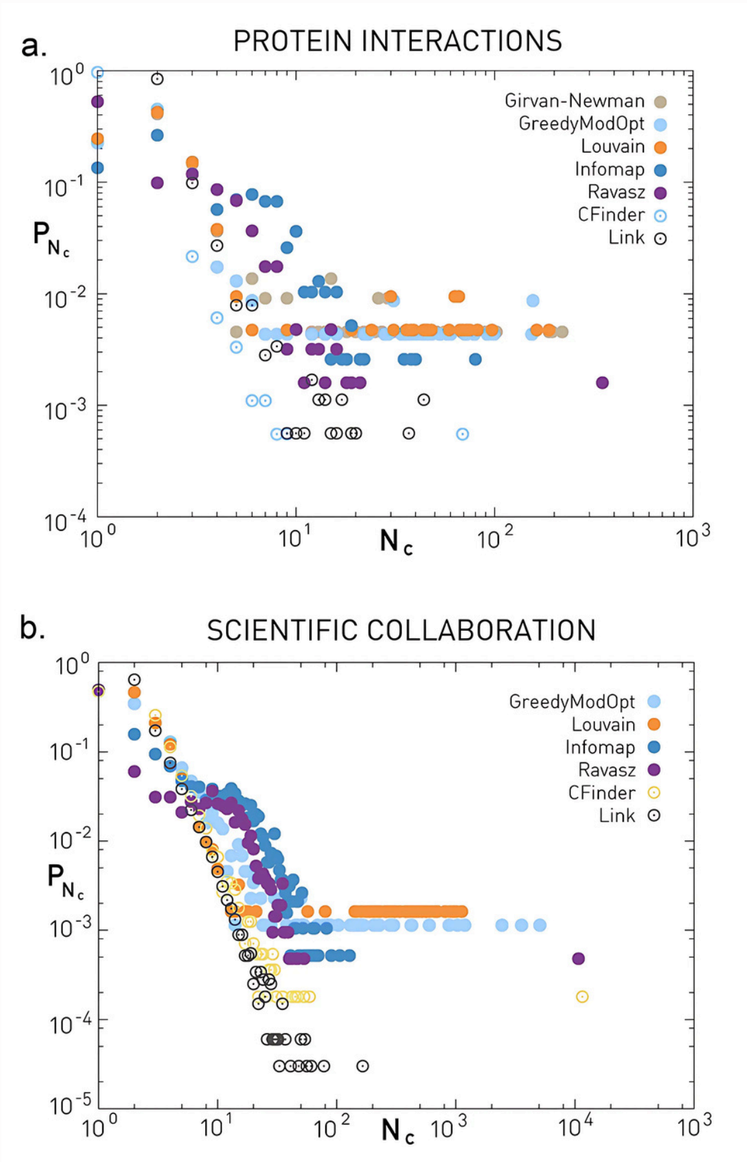


Community Size Distribution

Many networks exhibit a power law distribution of community sizes

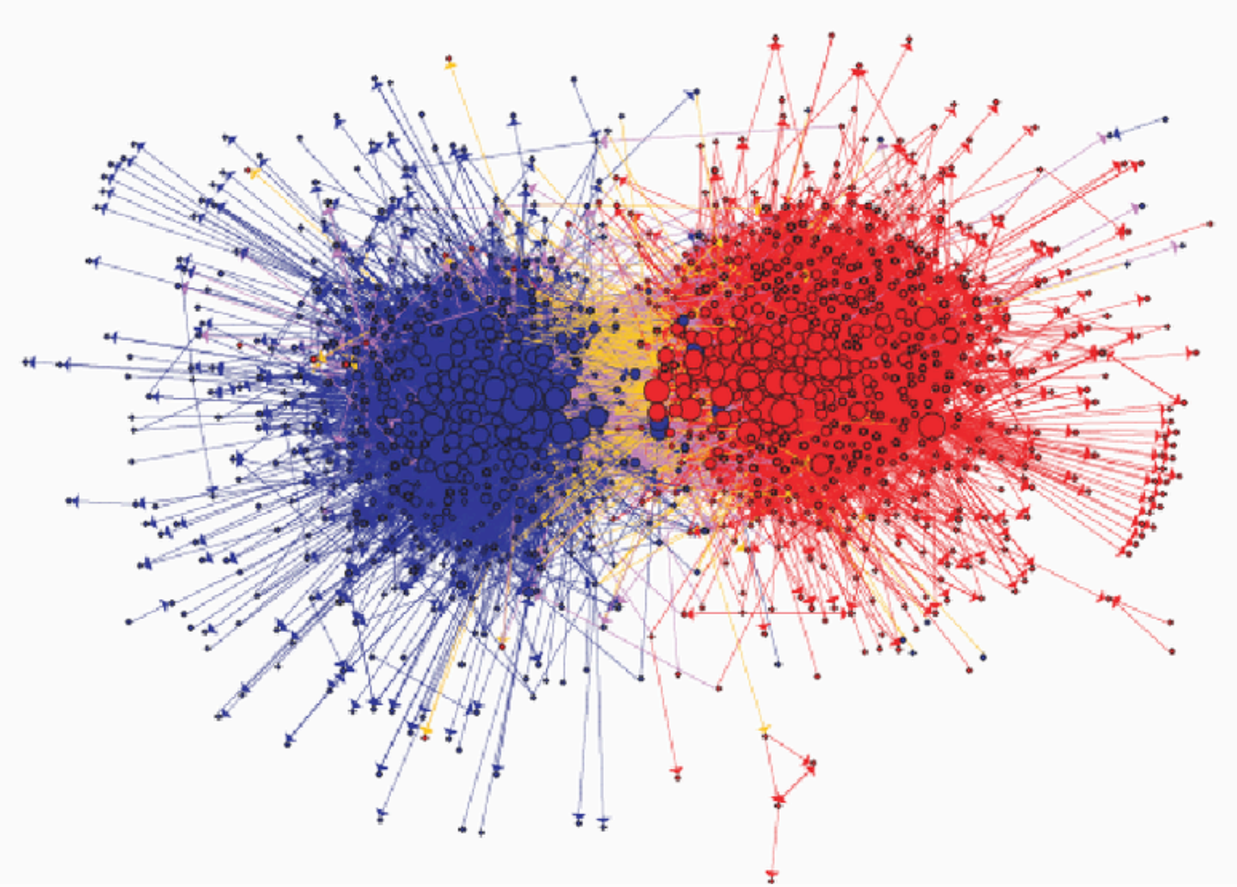
- Protein Interaction Network
- Science Collaboration Network

Fat-tailed distributions are not algorithm artifacts, but rather an inherent property of certain networks



Homophily and Communities Formation

Why Networks have Communities?



Now that we know how to detect communities, a natural question arises: why do networks form communities in the first place?

- Homophily and Clustering play an important role
- We need models to link these ideas to the observed structures in real-world networks

How do local interactions lead to global community structures? Can we replicate community formation using simple rules?

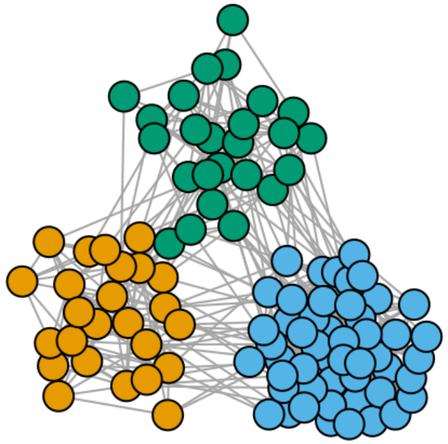
Stochastic Block Model

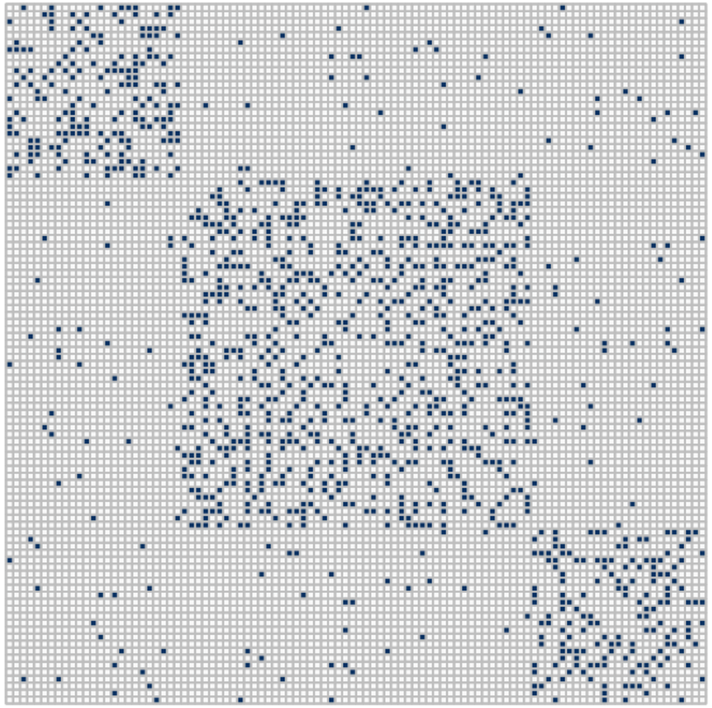
The Stochastic Block Model is one of the simplest generative model that creates communities

- similar to the random network model
- works by defining probabilities of connections within and between groups

Nodes are assigned to predefined groups and then linked at random

- Within-group: High probability of connections
- Between-group: Low probability of connections



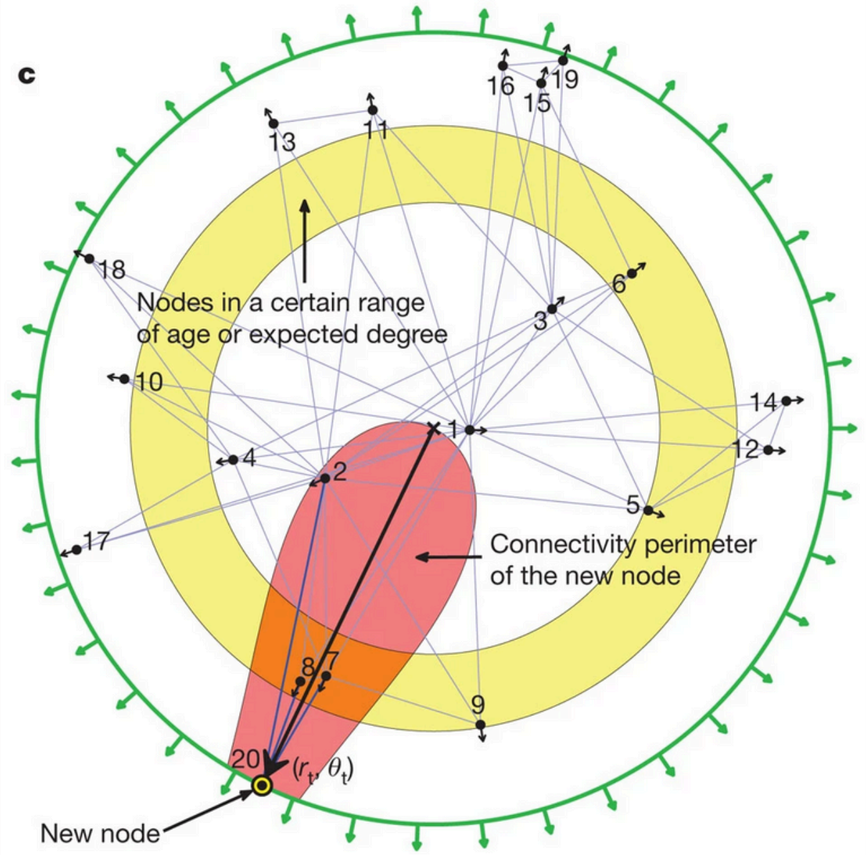


Homophilic Preferential Attachment

In many real-world networks, connections are driven by a combination of popularity and similarity or homophily

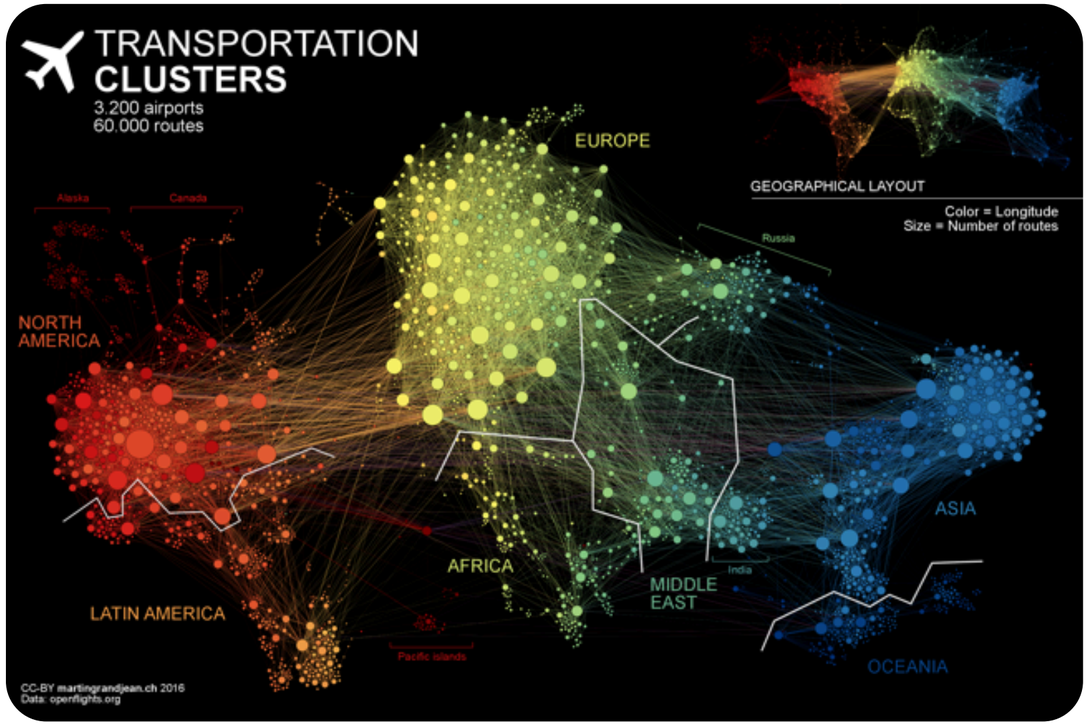
- The classic Barabási-Albert model focuses only on popularity
- We can modify the linking probability by introducing a similarity between nodes
- new nodes are more likely to link to similar and popular nodes

Also this model produces scale free networks but with a community structure



Papadopoulos, Fragkiskos, et al. "*Popularity versus similarity in growing networks.*" Nature 489.7417 (2012): 537-540.

Spatial Networks



In many real-world scenarios, networks are not abstract but are embedded in space

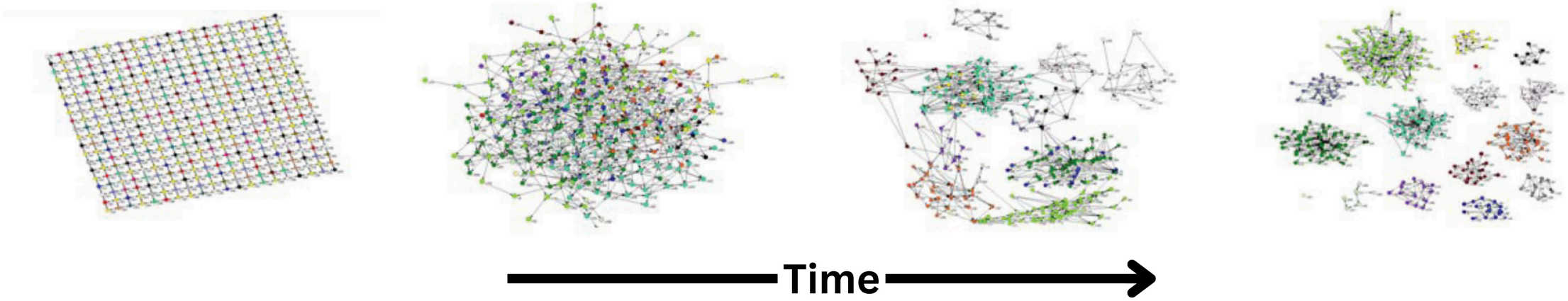
- Nodes that are closer in space are more likely to connect
- Proximity plays the role of similarity
- Examples of Spatial Networks
  - Airport Networks
  - Road Networks
  - Power Grids

We can simulate such a situation by placing nodes on a plane and creating links with probability proportional to proximity and degree

Coevolution Models

Are individuals drawn to those who are already similar to them (similarity bias)or does joining a community make its members more similar over time?

- The reality is likely more complex: a feedback loop between similarity and connections exists
- Similarity Drives Connections and Connections Strengthen Similarity:
- Once individuals join a community, interactions can lead to convergence



Centola, Damon, et al. "*Homophily, cultural drift, and the co-evolution of cultural groups.*" Journal of Conflict Resolution 51.6 (2007): 905-929.

Affiliation Networks

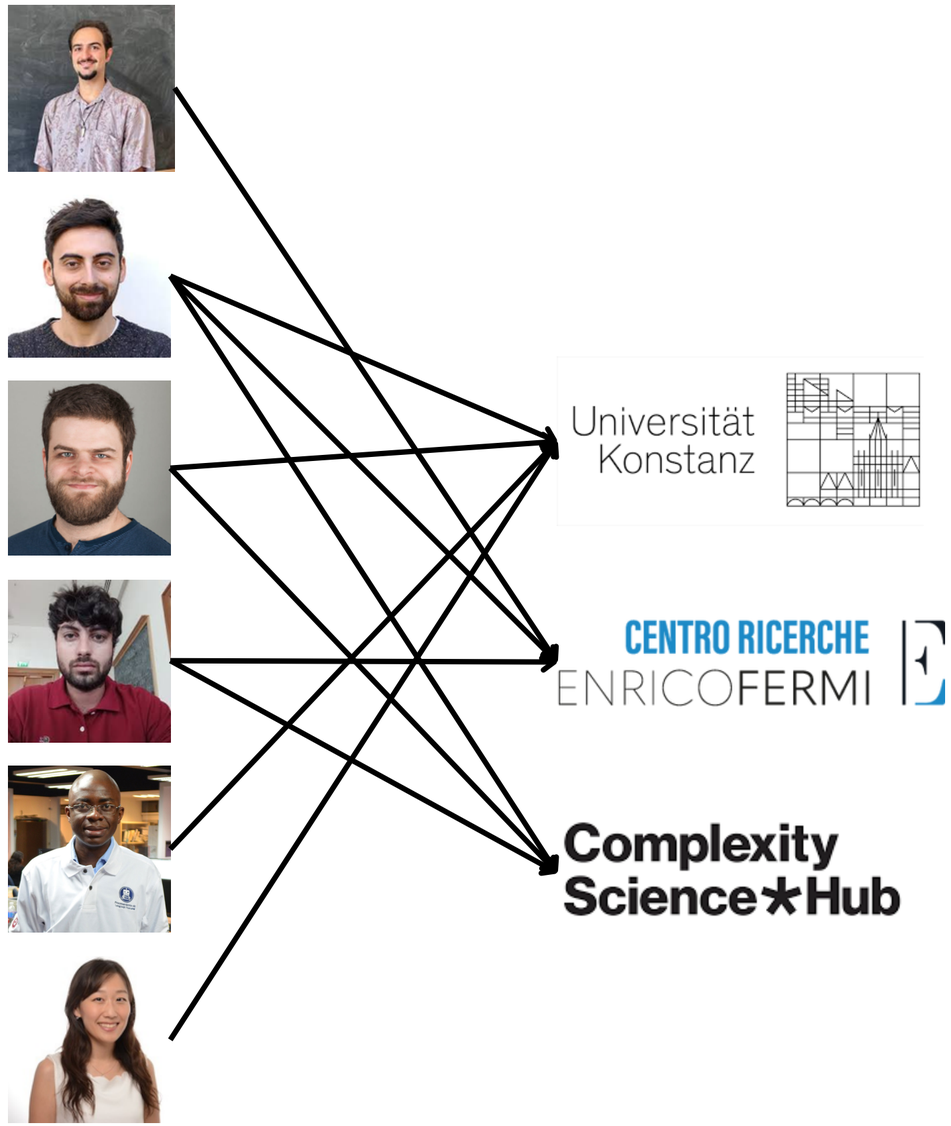
Connections between individuals may derive from shared affiliations/activities

- work place
- school
- hobbies

We can describe this as a bipartite network

- we have two classes of node
  - researchers
  - research centers
- no links among nodes of the same type
- links mean affiliation

Bipartite networks are very common!

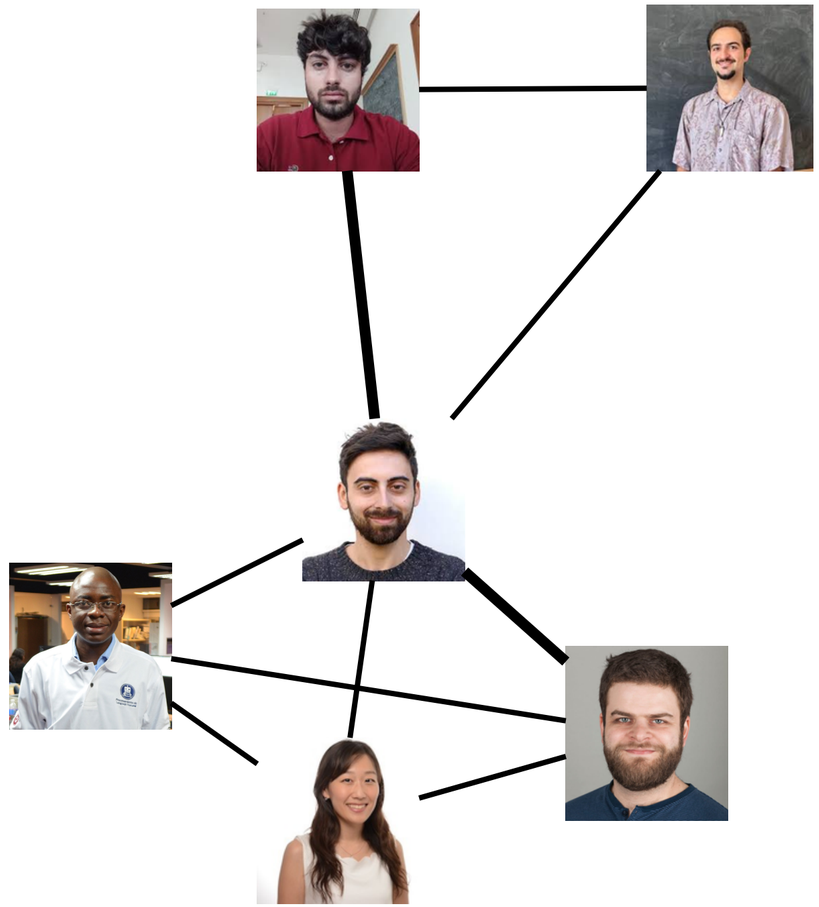


Bipartite Networks Projection

Starting from a bipartite network we can get a monopartite network by performing a projection

- two possible projections (one per layer) are possible
- let's consider researchers
  - the simplest approach is to link people sharing the same affiliation
  - more affiliations shared = stronger links
- in the same way one could build a network of research centers

We will consider the projection of bipartite networks in more detail in the upcoming lectures



The Strength of Weak Ties

How to Find a Job

Often the best way to find a new job is through friends

- personal contacts allow people to access information
- however the most useful contacts are not close friends
- acquaintances are generally more useful in this context?

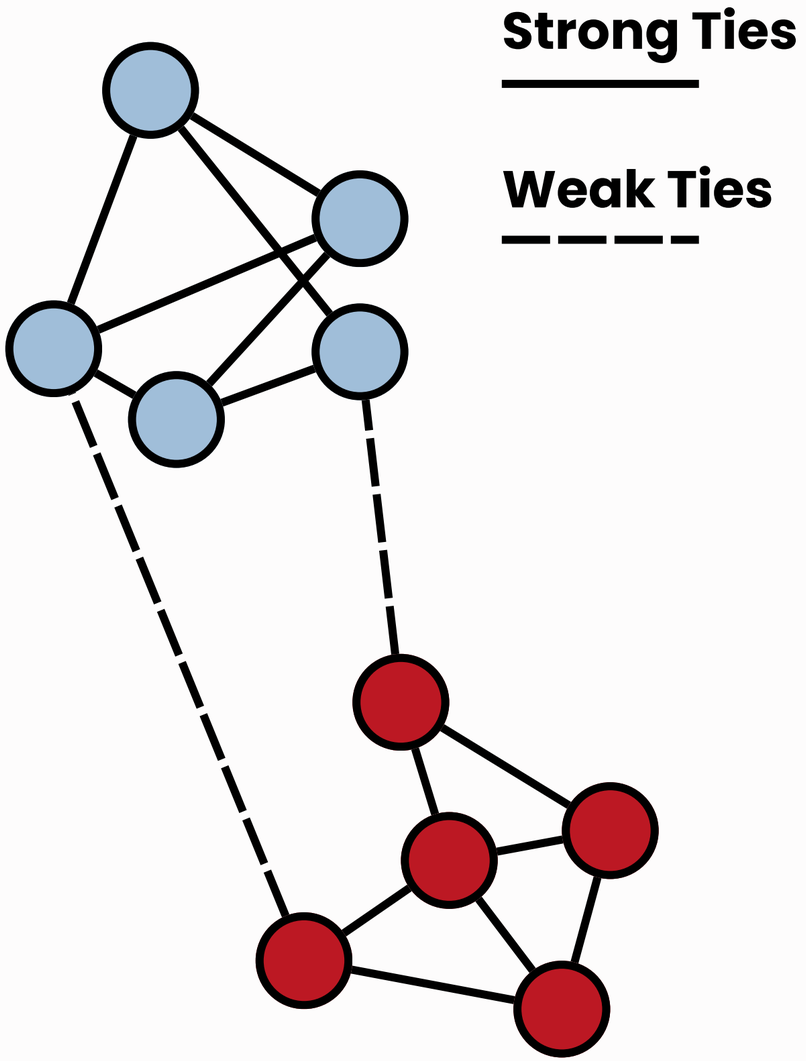
What is going on? Our friends should be more helpful than random acquaintances

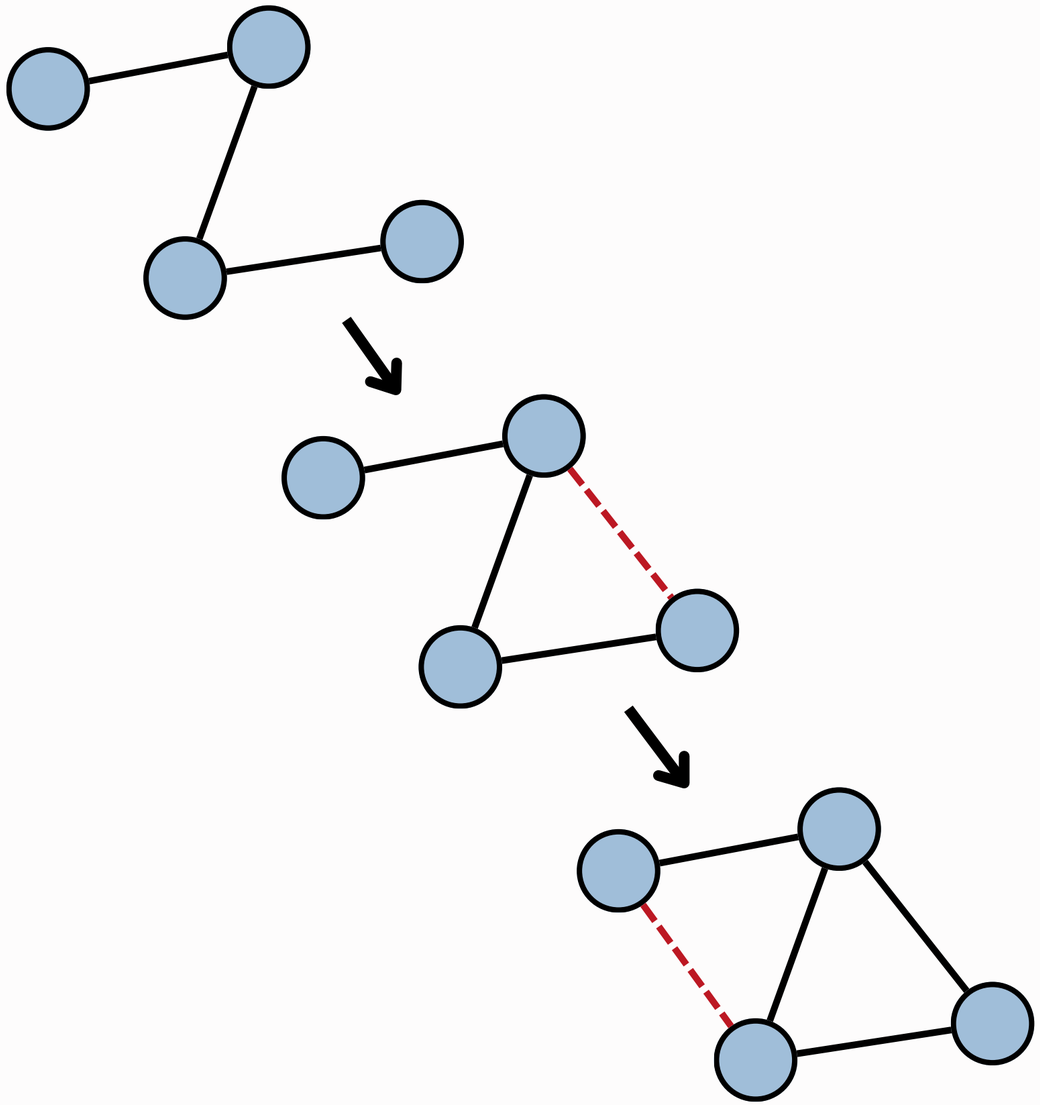
The Strength of Weak Ties

Mark Granovetter addressed this paradox in his seminal paper “The Strength of Weak Ties”

- social networks are characterized by strong and weak ties
  - links with our friends are strong but also redundant
  - links with acquaintances are much weaker but not redundant
- weak ties are more fragile, but they give access to precious information

In practice, your close friends can’t help because they have the same information as you do





Structurally Embedded Links

Strong friendships (links) tend to be also structurally embedded

- these links are formed in a region already full of links
- triadic closure plays an important role in this context
- friends of friends tend to become friends
- this create stronger links, but also a redundant structure

Edge Overlap

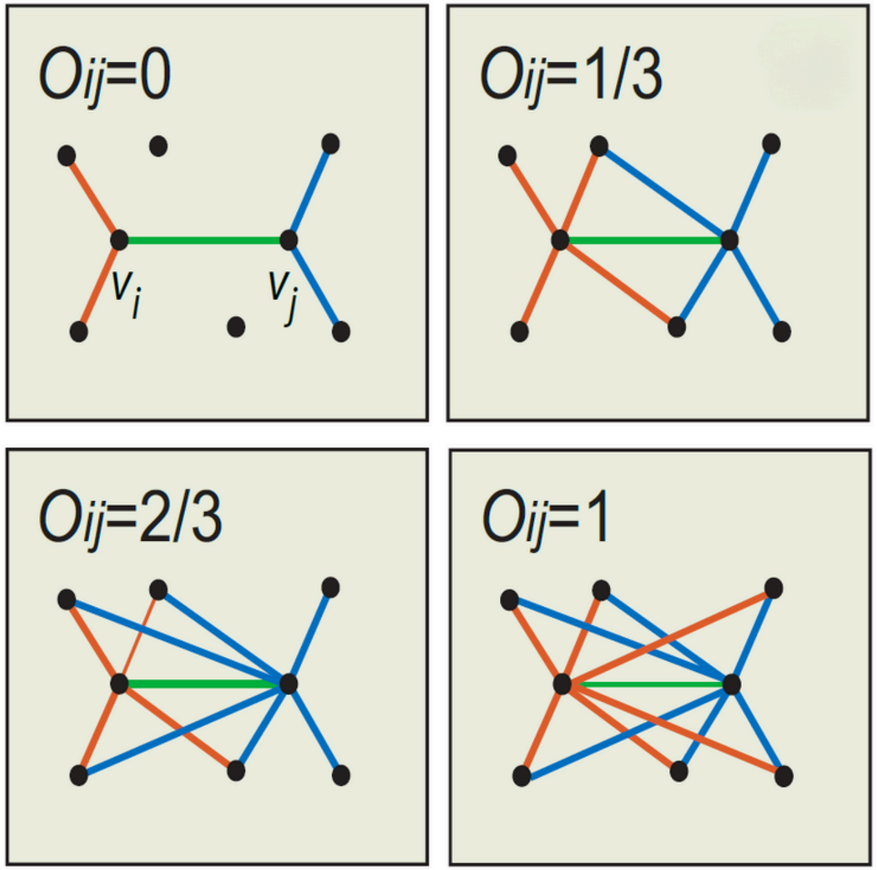
We saw that we can use the (local) clustering coefficient to measure the number of triangles a node forms

- to quantify redundancy we need a similar measure also for edges
- we can do so using the edge overlap

Given two nodes i, j, the edge overlap O_{ij} of the link connecting them is

$$O_{ij} = \frac{|(N(i) \cap N(j)) \backslash \{i, j\}|}{|(N(i) \cup N(j)) \backslash \{i, j\}|}$$

In simple terms, the overlap is the number of common friends of i, j divided by the total number of friends i and j have



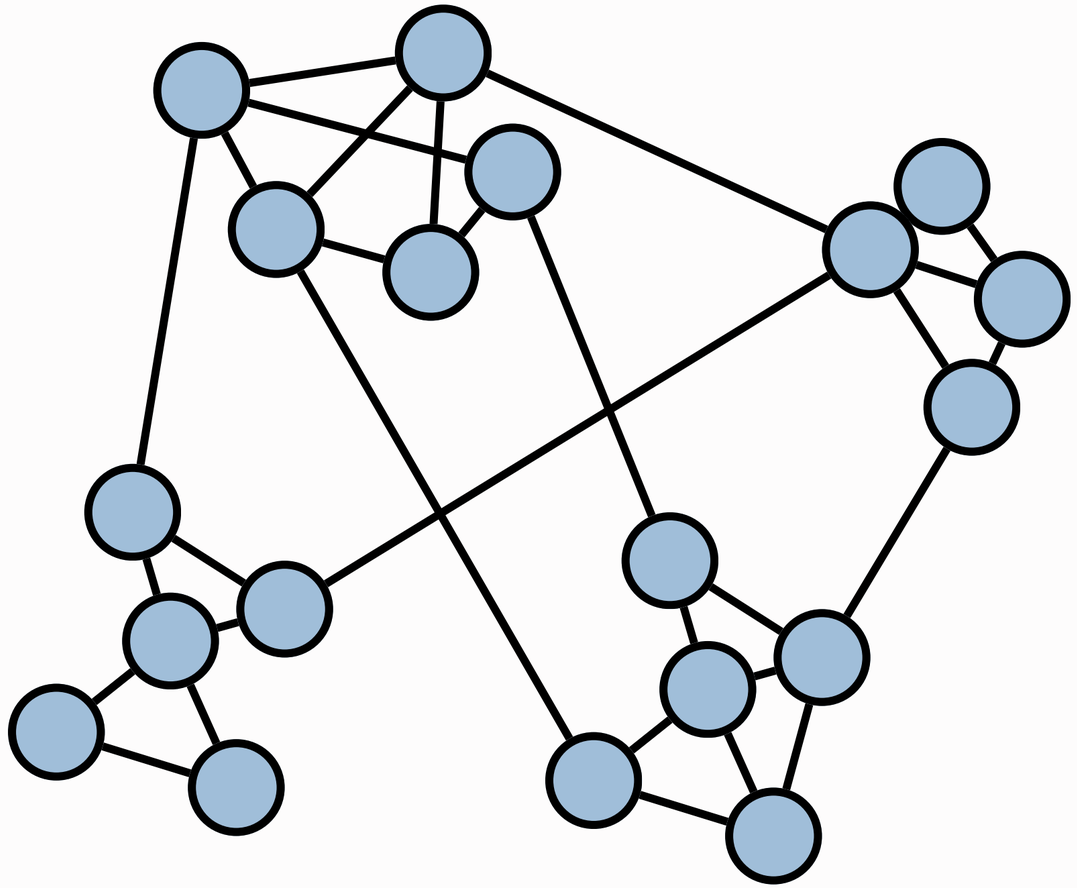
Testing Granovetter's Hypothesis

Granovetter's analysis suggests that social networks have a community structure

- in 1973 no data were available to test the hypothesis
- nowadays instead we have access to several sources
  - online social networks
  - phone networks
  - email networks

Granovetter's hypothesis was first tested in 2007 using a cell-phone network

- 20% of EU country's population
- Edge weight: number of phone calls



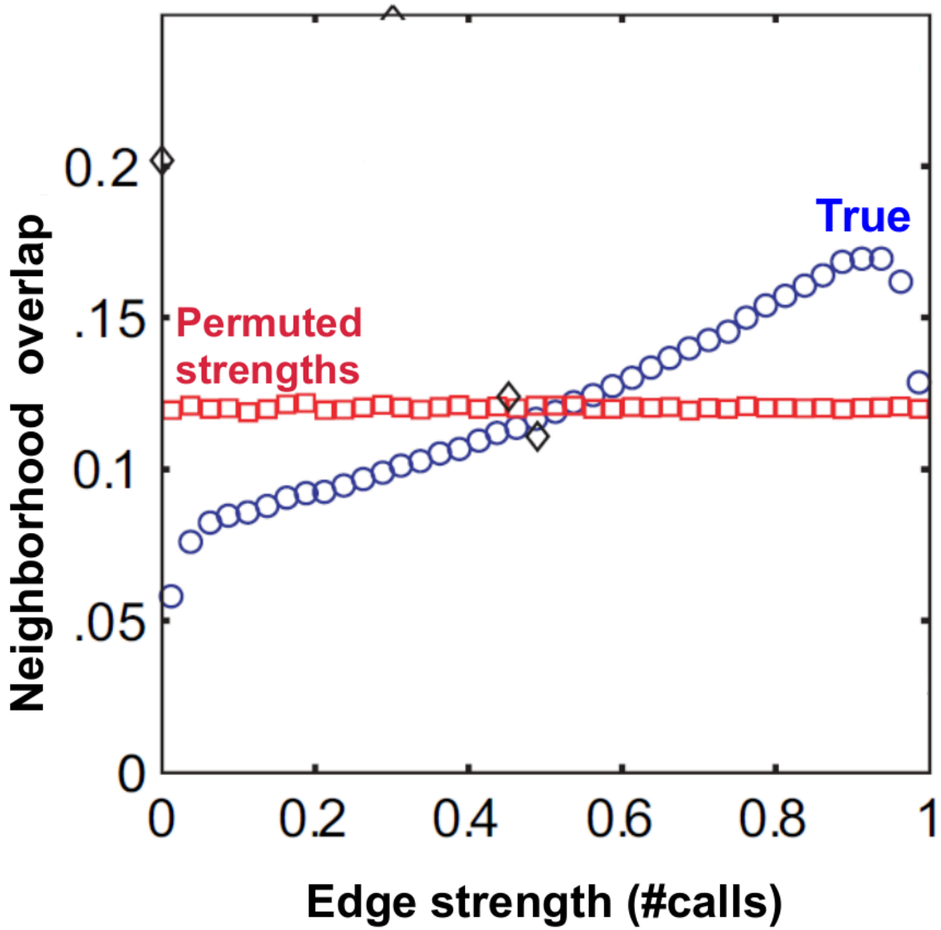
Onnela, J-P., et al. "*Structure and tie strengths in mobile communication networks.*" PNAS 104.18 (2007): 7332-7336.

Overlap vs Strength

The analysis of the phone networks reveals that

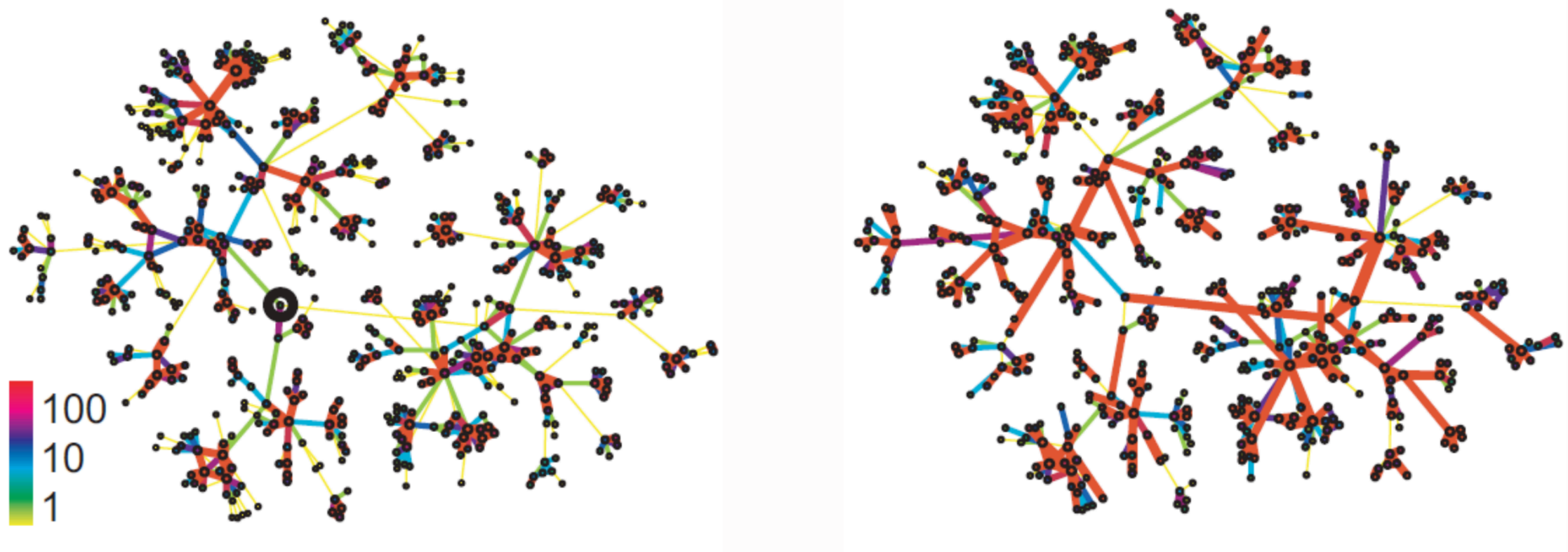
- there is a correlation between strength and overlap
- the strongest connections are also the most redundant
- randomly permuting the strengths instead leads to no correlation

This confirms Granovetter's hypothesis



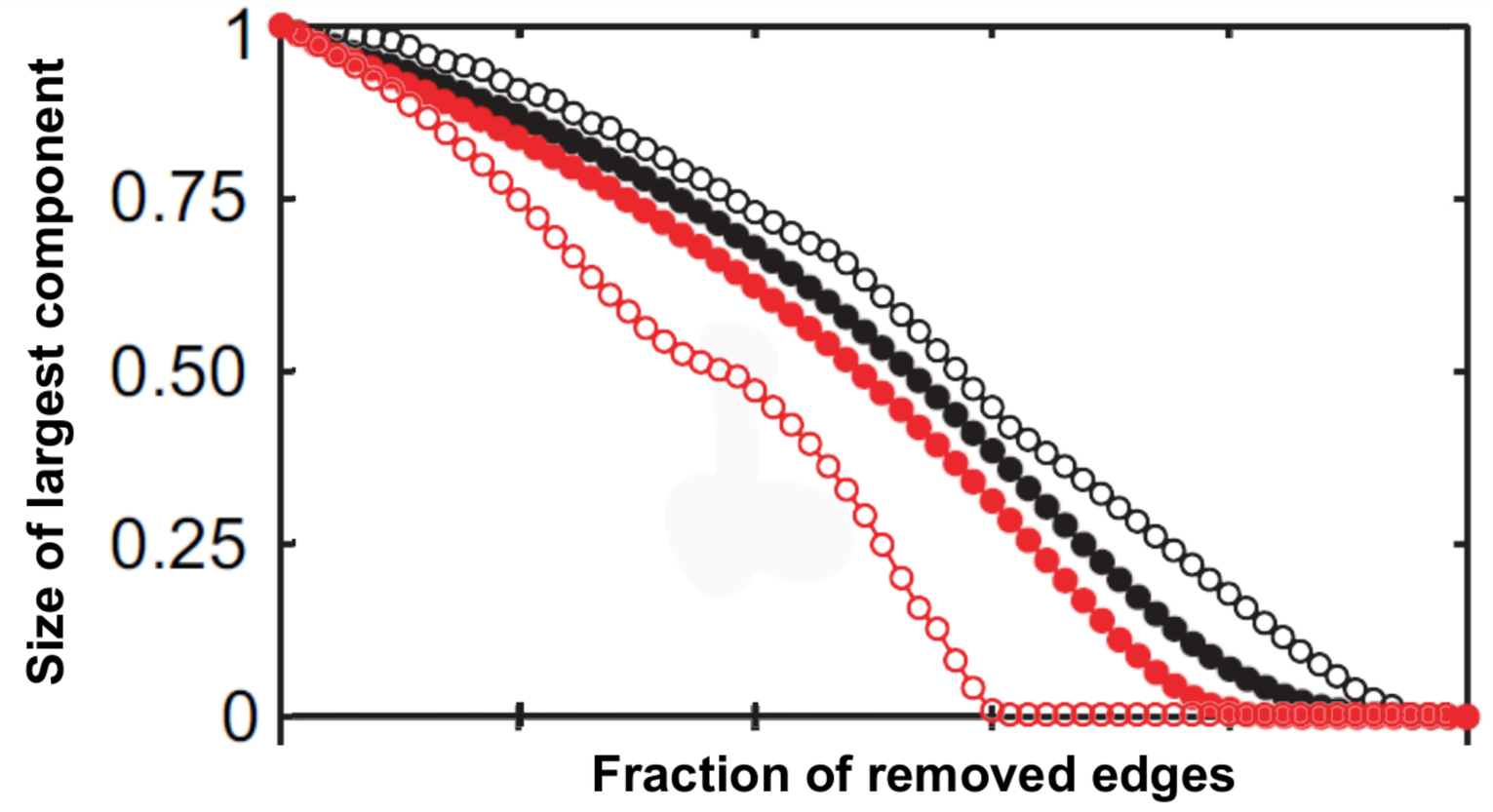
Visualizing the Phone Network

On the left is represented the network around a random individual. High strength connections are very embedded, while bridges have a low strength. On the right the same network but with randomly permuted strengths



The Importance of Bridges

Low Overlap edges act as bridges connecting different communities. Removing edges in increasing value of overlap disrupt the network the fastest



- Low to High Strength
- Low to High Overlap
- High to Low Strength
- High to Low Overlap

Conclusions

Communities in Networks

Real networks are characterized by communities and community detection is used to detect them. Modularity can be used to asses the quality of partitions.

Community Detection Algorithms

We introduced some of the most known community detection algorithms, pointing out their limitations and strengths.

Homophily and Communities Formation

Homophily plays a central role in the formation of communities. We introduced various models that link this tendency to the emergence of communities.

The Strength of Weak Ties

Weak ties connect communities in social networks acting as bridges, while edges within communities are characterized by an high redundancy.