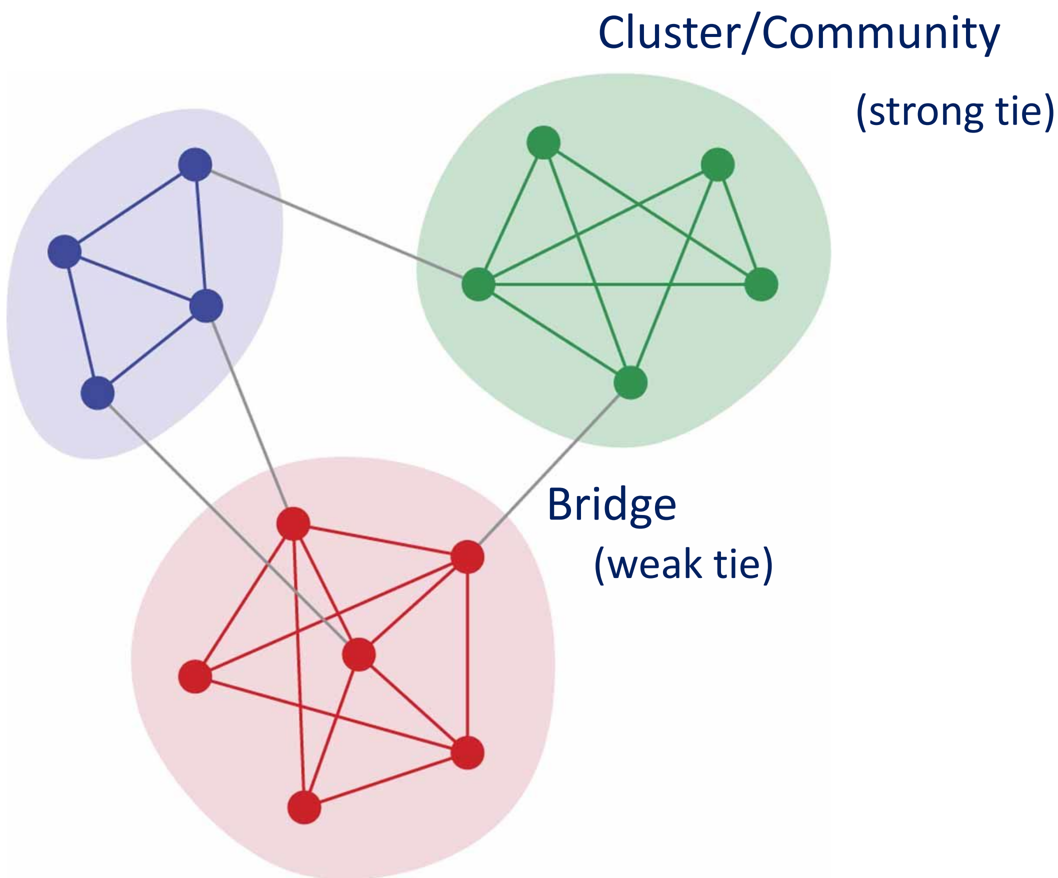
# Network Science

## #09 Community detection

# Network communities

# Conceptual picture of a network




- We often think of networks looking like this
- But, where does this idea come from?

# Granovetter’s explanation

Granovetter, The strength of weak ties [1973]
https://www.jstor.org/stable/pdf/2776392.pdf

**Q**: How do people discovered their new jobs?

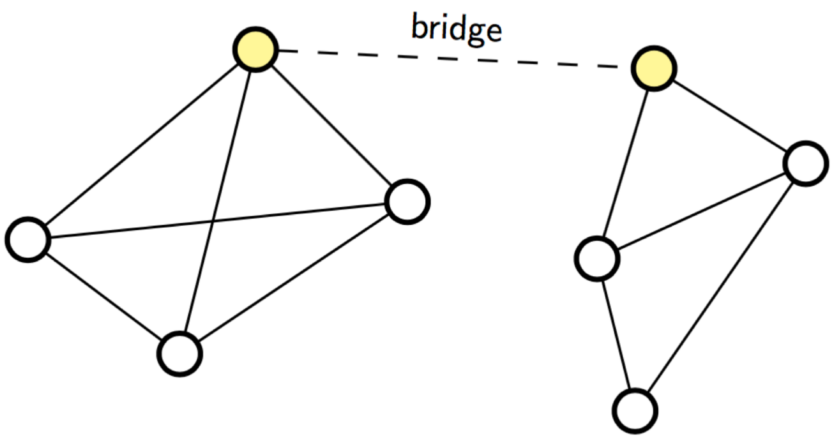
**A**: Through personal contacts, and mainly through acquaintances rather than through close friends

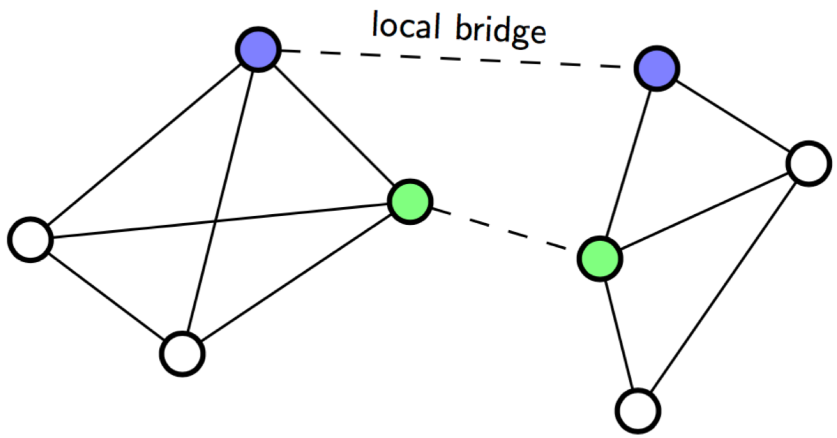
Remark: Good jobs are a scarce resource

Conclusion:

- Structurally embedded edges are also socially strong, but are heavily redundant in terms of information access
- Long-range edges spanning different parts of the network are socially weak, but allow you to gather information from different parts of the network (and get a job)

# Local bridges




- An edge (*i*,*j*) is a **bridge** if deleting it *i* and *j* fall into different components

  this is extremely rare, e.g., because of small world properties

- An edge (*i*,*j*) is a **local bridge** if, by deleting it, *i* and *j* have a span (distance) greater than 2, i.e., if *i* and *j* **do not have friends in common**

  common friends imply belonging to a triadic closure

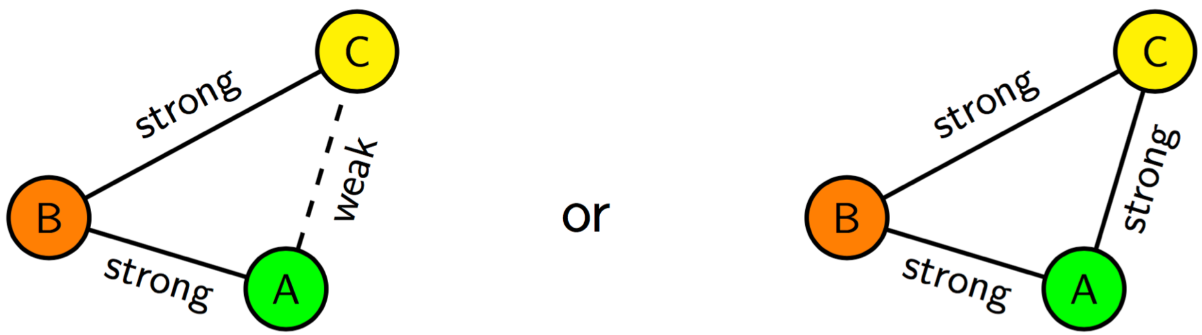
MIME

# Strong triadic closure

Assume two categories of edges:

- Strong ties (close friends)
- Weak ties (acquaintances)

Remark. If node B is strongly tied with A and C, then A and C are very likely to be connected (either weakly or strongly), that is




Strong triadic closure property – If a generic node B is strongly tied with A and C, then A and C are connected (either weakly or strongly)

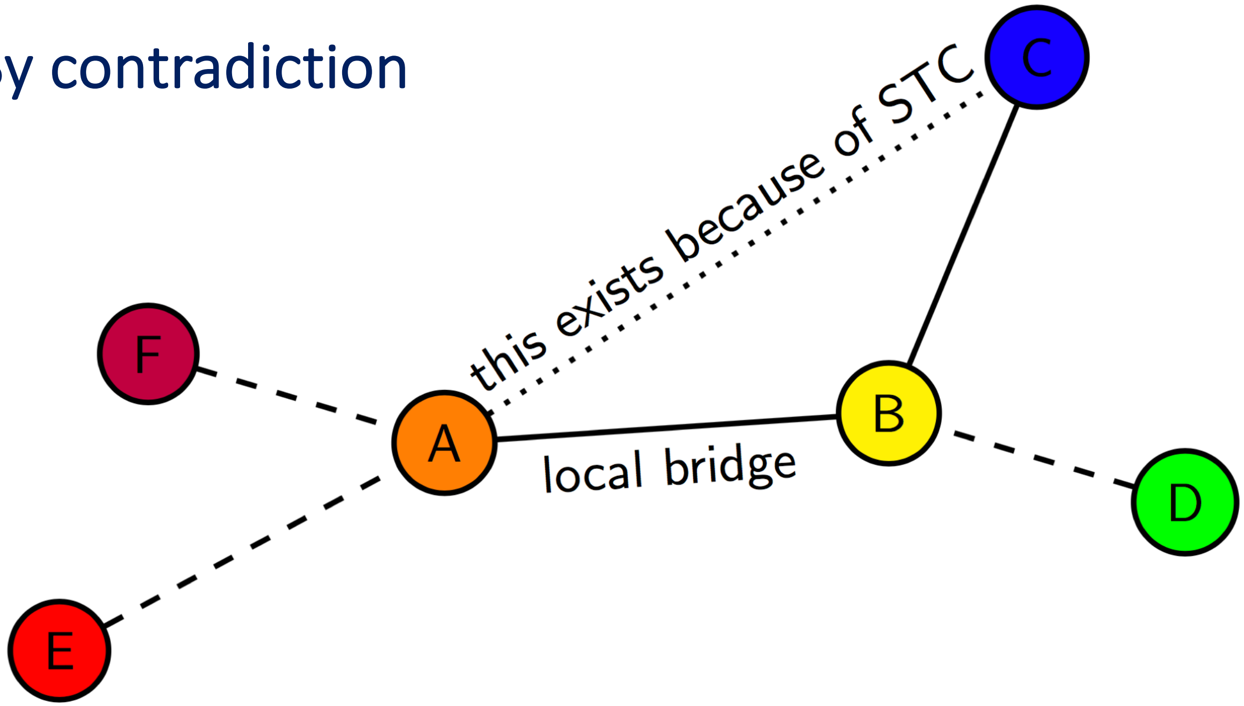
# Granovetter's claim

Claim:

- Under the strong triadic closure property, local bridges are weak ties (if at least one of their nodes belongs to at least two strong ties)

Proof:

- By contradiction



# Community detection

- Granovetter's theory suggests that networks are composed of tightly connected sets of nodes (i.e., communities), loosely connected between them
- We want to be able to automatically find such densely connected group of nodes
- Applications in

  Social networks

  Functional brain networks in neuroscience

  Scientific interactions

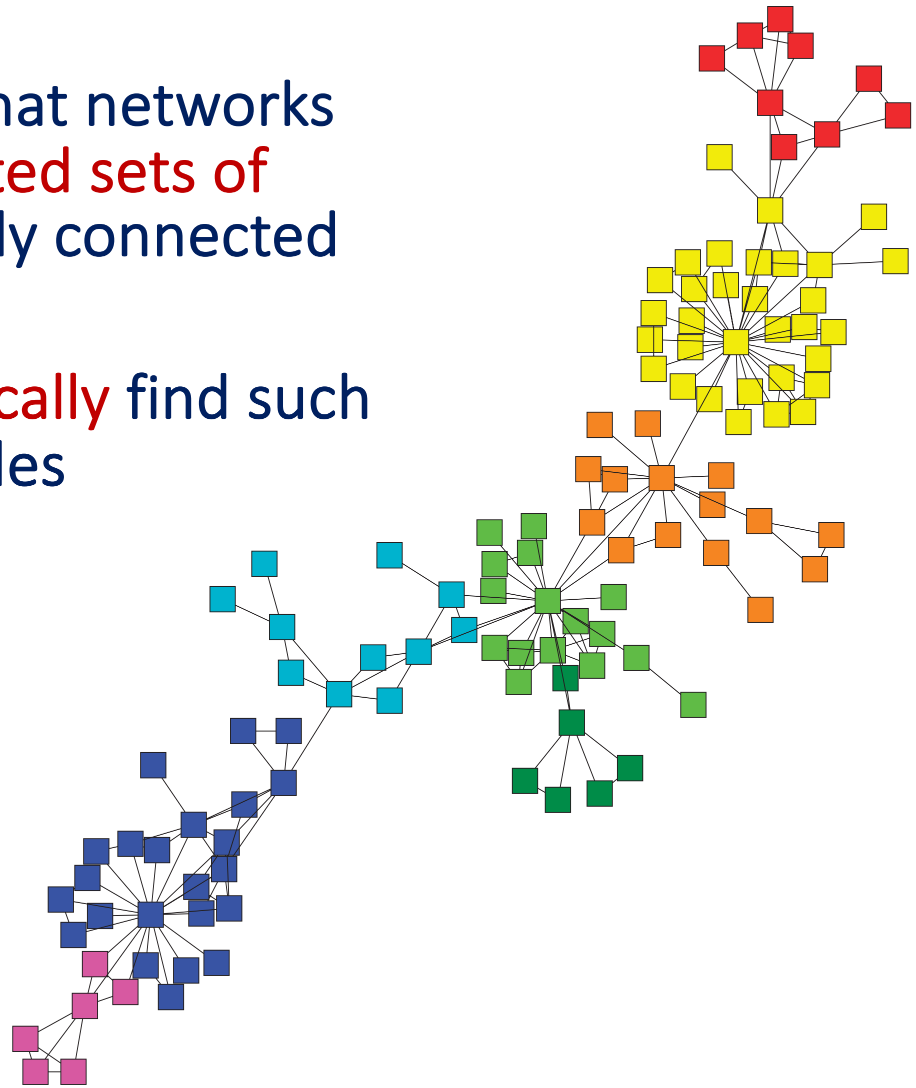
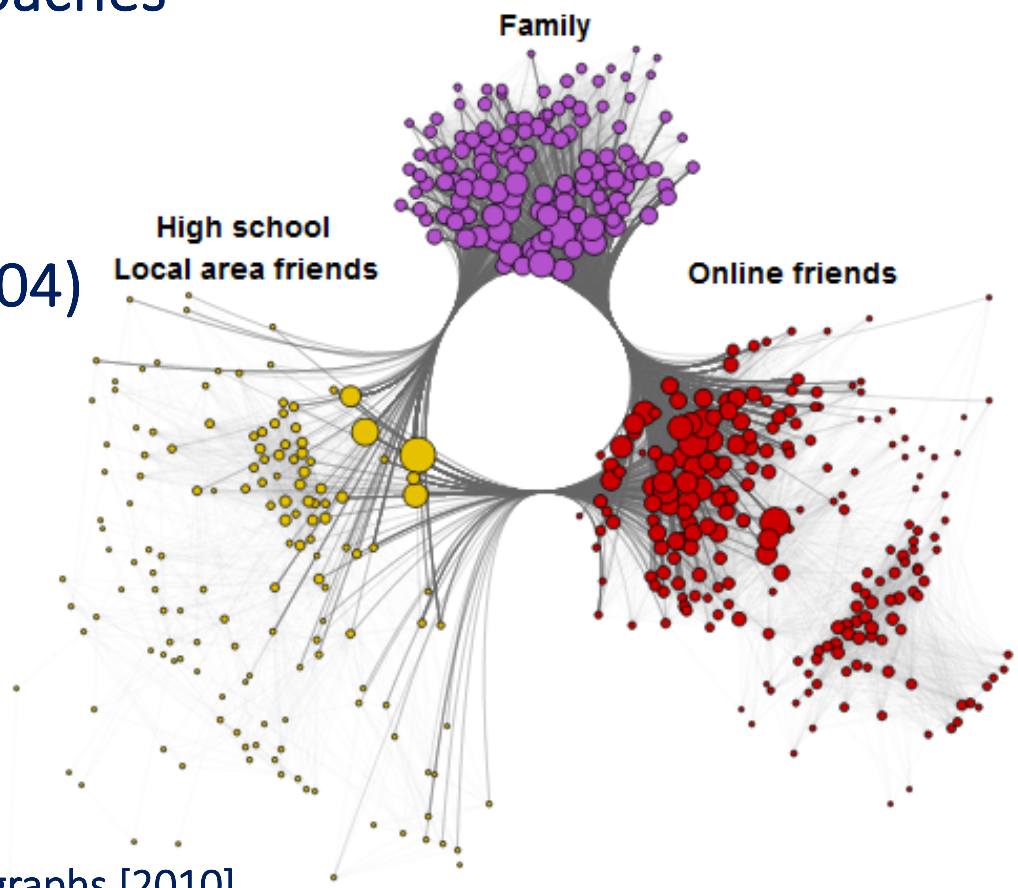


**Figure 2 |** A network of collaborations among scientists at a research

# Community detection

Some relevant algorithms/approaches

- ❑ Dendrograms
- ❑ Girvan-Newman (2001)
- ❑ Modularity optimization (2004)
- ❑ Spectral clustering (2002)




Find a complete list in:

Fortunato, Community detection in graphs [2010]

https://www.sciencedirect.com/science/article/pii/S0370157309002841

# Overlapping communities

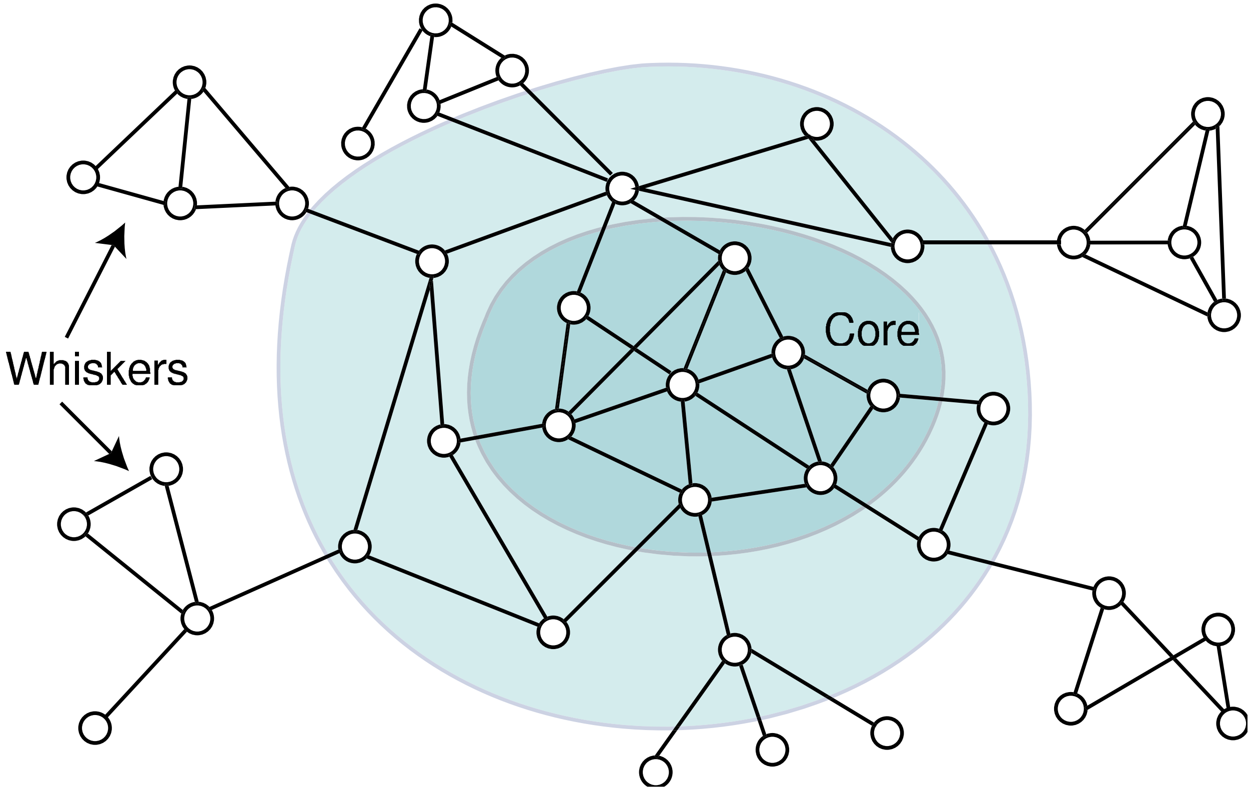
Lescovec, Lang, Dasgupta, Mahoney, 2008

Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters

https://arxiv.org/abs/0810.1355

# The core-periphery model

Small, peripheral clusters




Caricature of network structure

Can we find a justification for this?

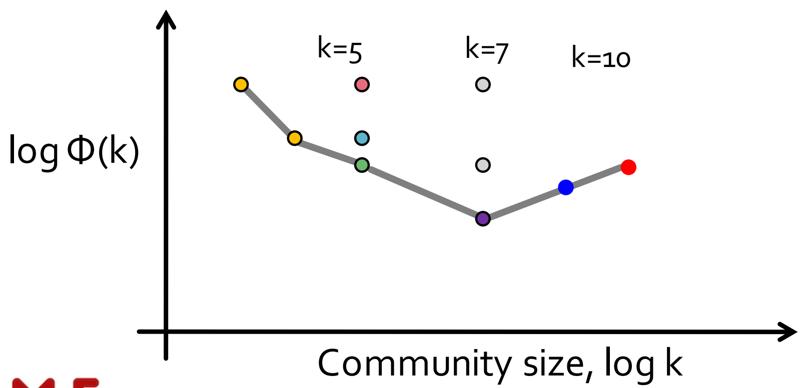
# Network community profile

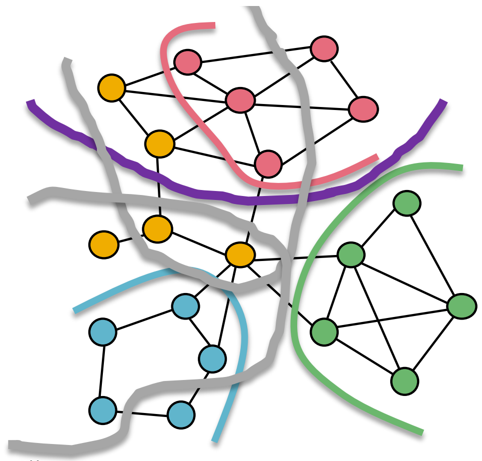
Conductance $\phi(S)$ – a metric for clusters

- $S$ is a good cluster if it has many edges internally and few pointing outside

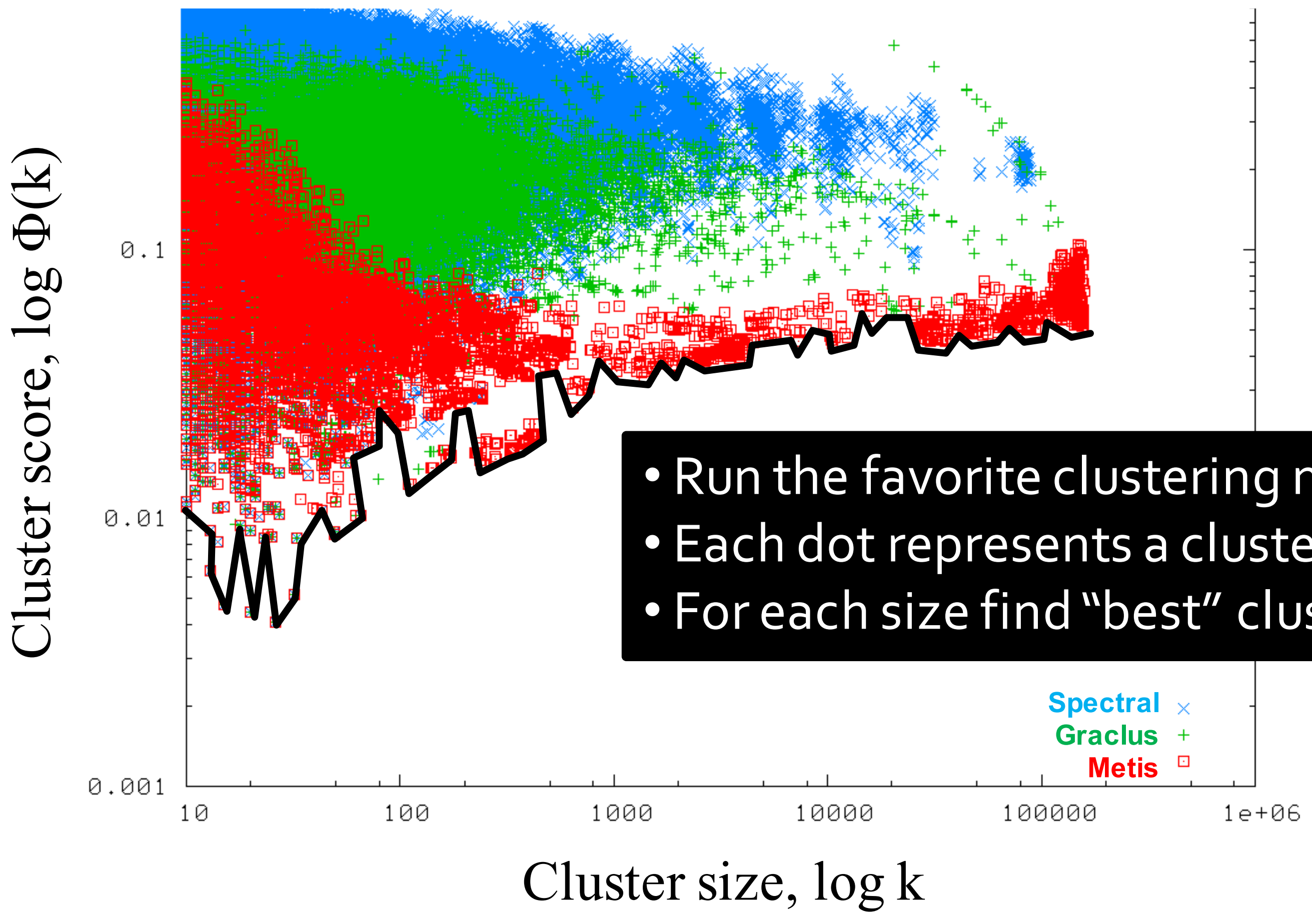
Network community profile – a metric for networks

- $\boldsymbol{\Phi}(k) = \min_{|S|=k} \phi(S)$
- Shows the best score for communities of order $k$



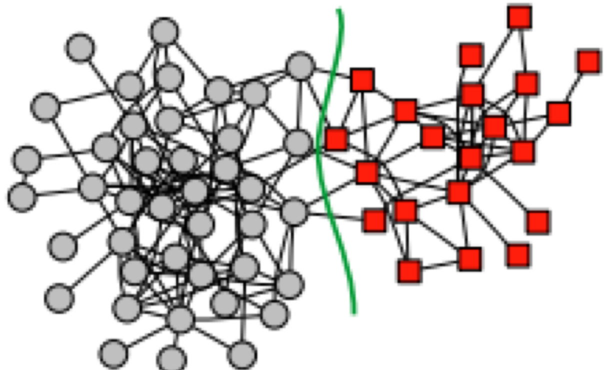



# Network community profile

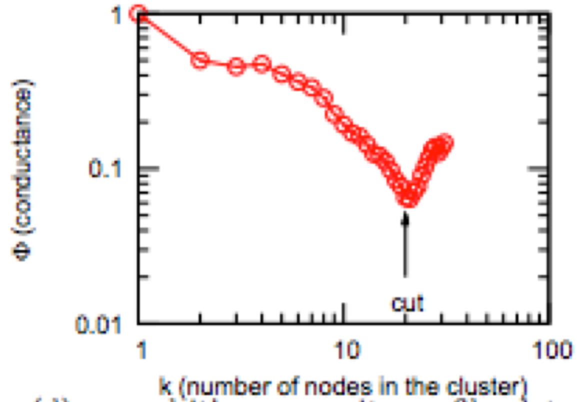



- Run the favorite clustering method
- Each dot represents a cluster
- For each size find "best" cluster

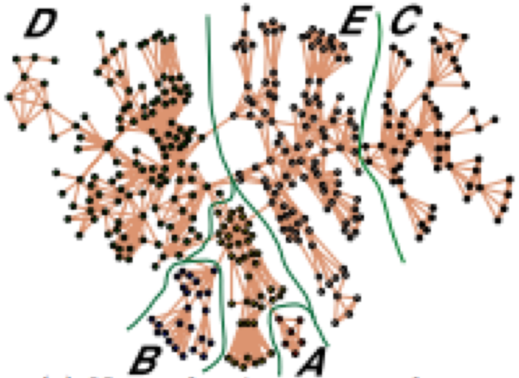
# Examples



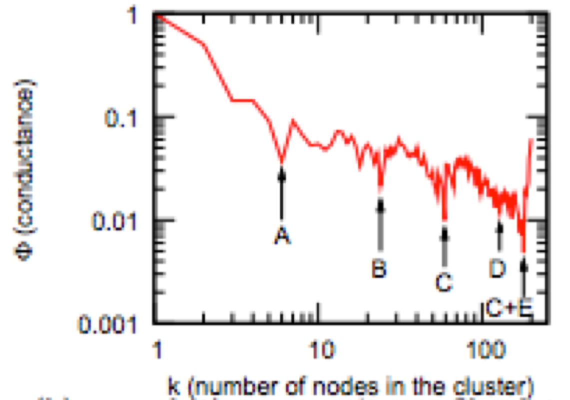
(c) Dolphins social network ...




(d) ... and it's community profile plot

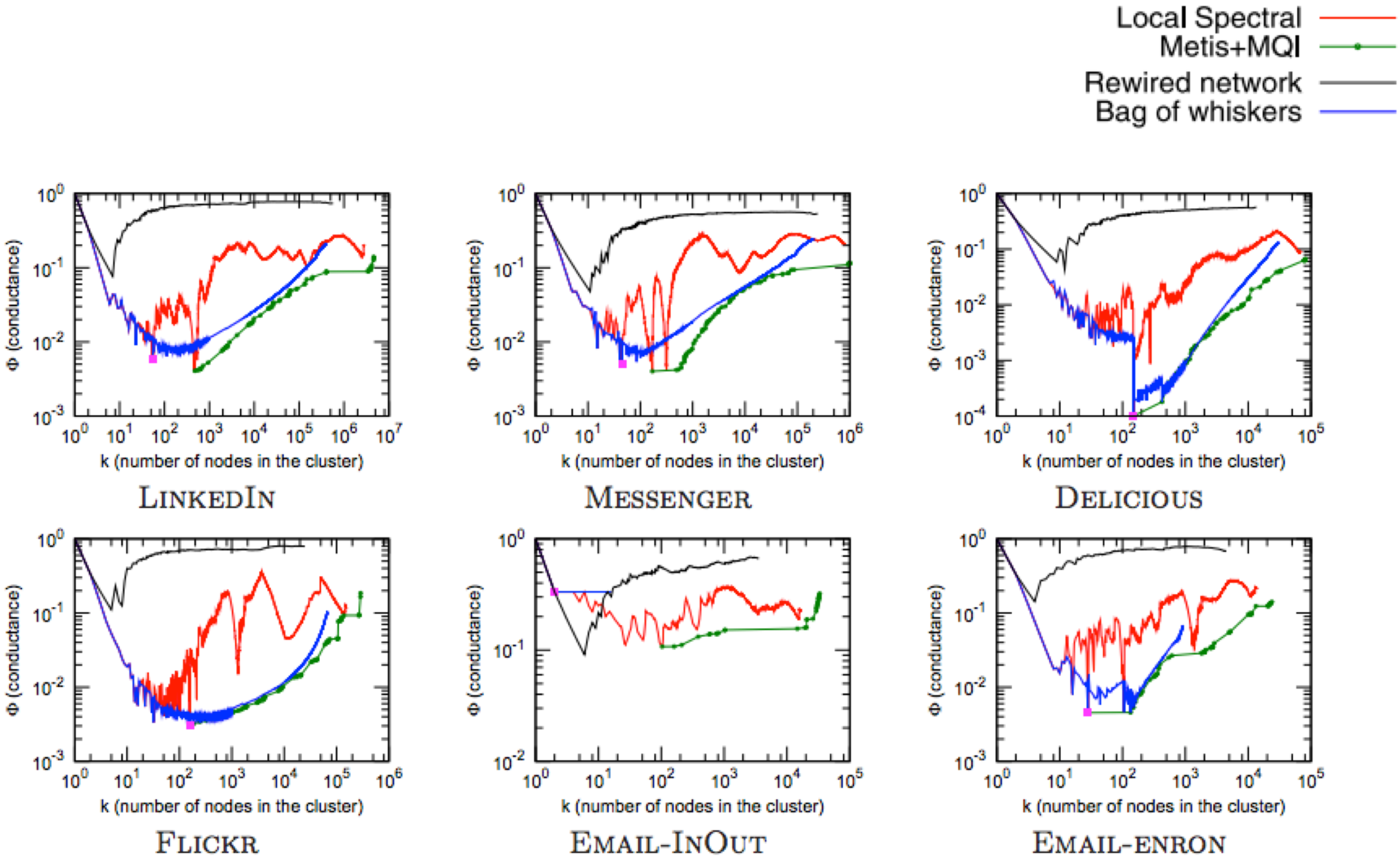



(g) Network science network ...

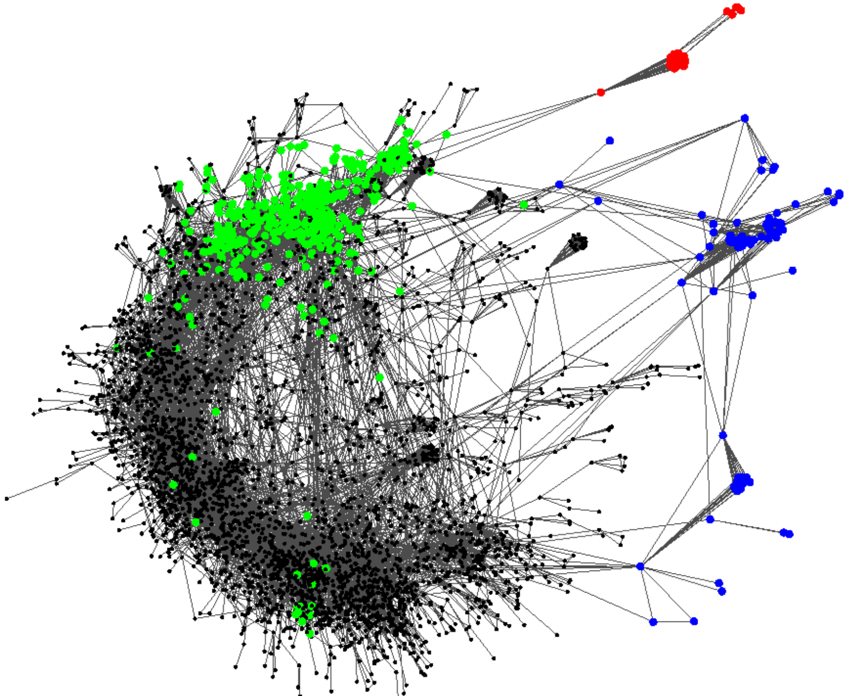


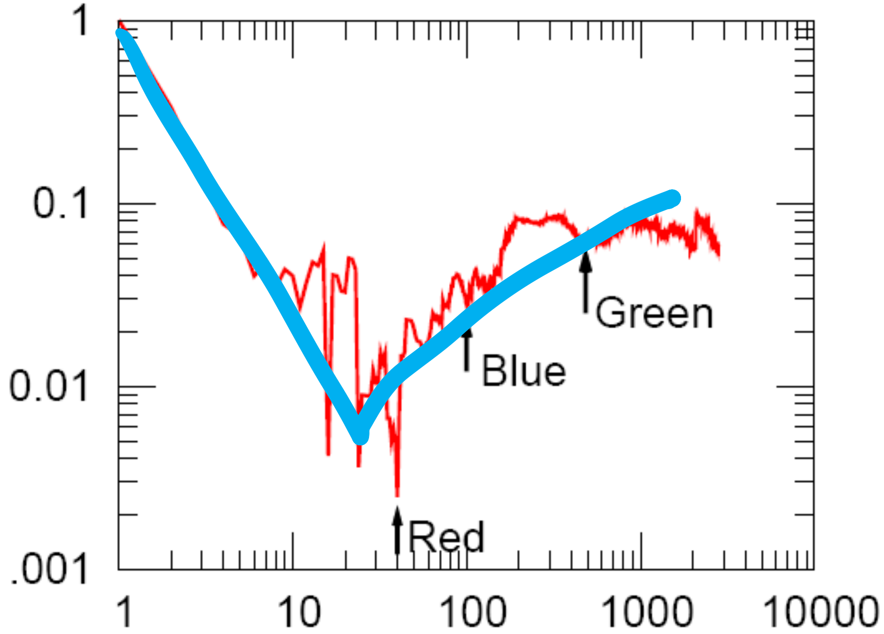

(h) ... and it's community profile plot

# Social network examples



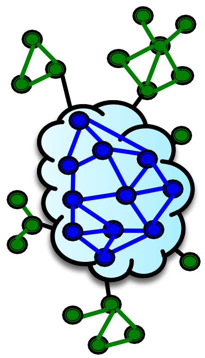
# V shape of NCP

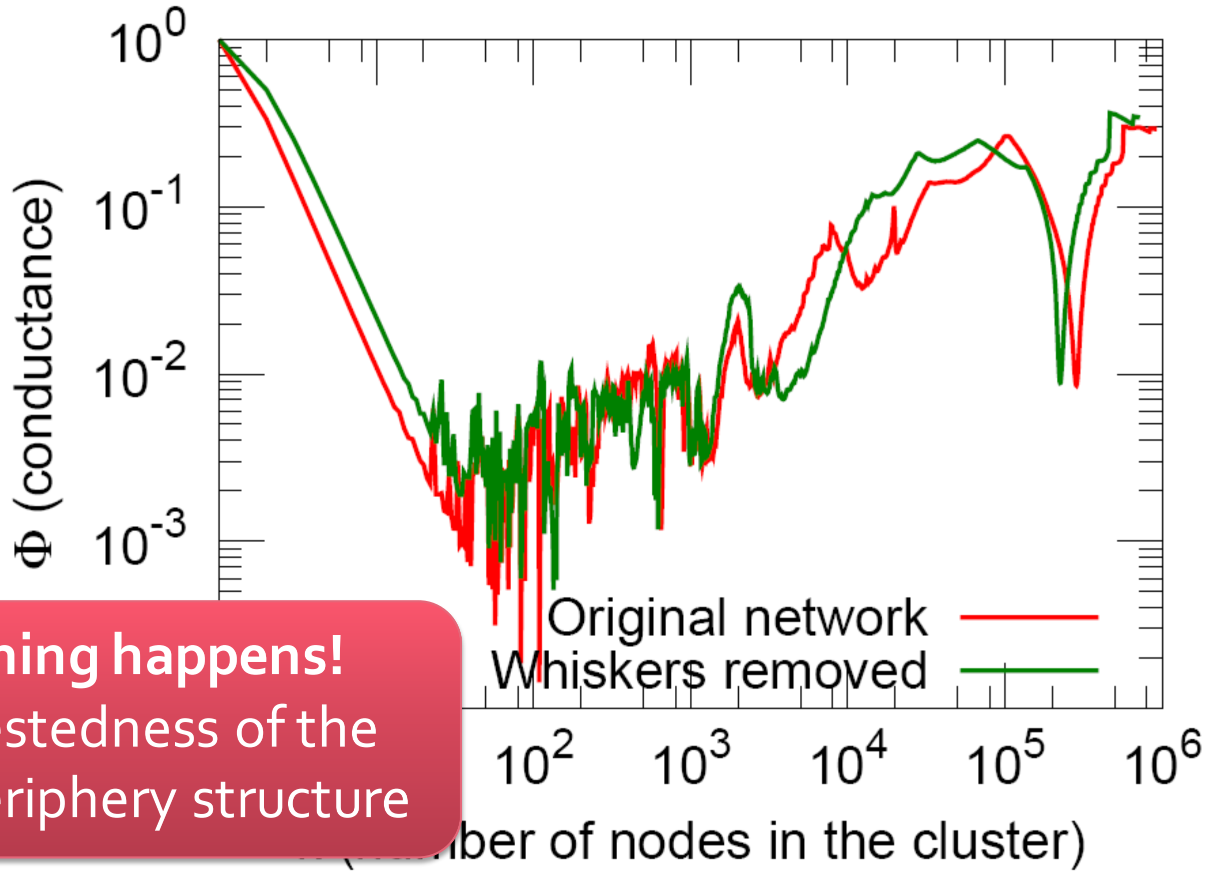





- Dips in the graph correspond to the good clusters
- Slope corresponds to the dimensionality of the network
- The V shape is common in large (social) networks
- Best clusters have about 100 nodes
- Large clusters get worse and worse performance

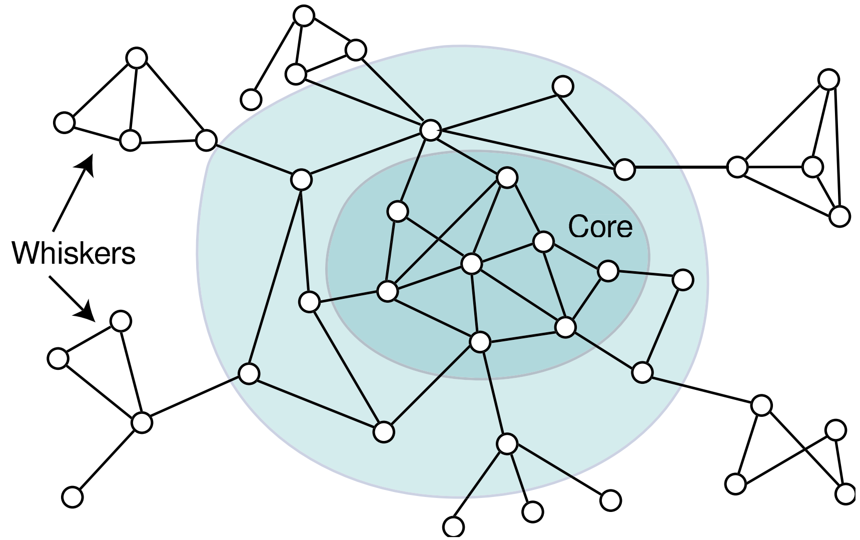
# What if we remove good clusters?






**Nothing happens!**
⇒ Nestedness of the core-periphery structure

# Overlapping communities model

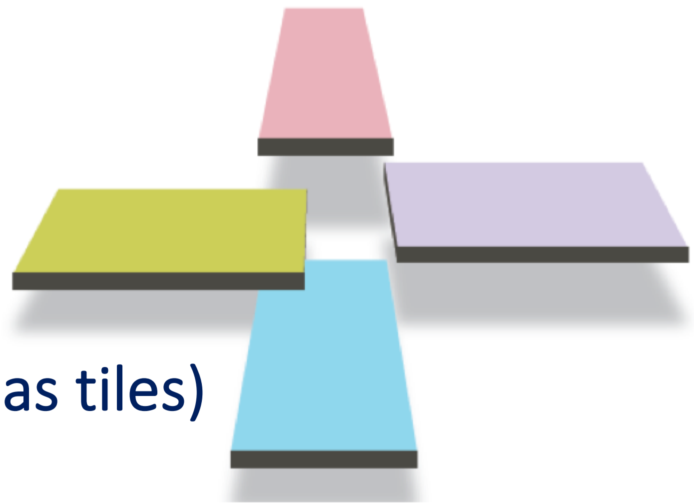



Wiskers

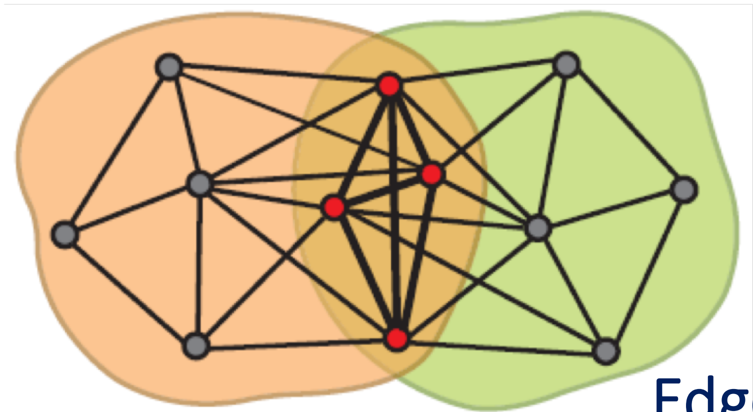
- are typically of size 100
- are responsible of good communities

Core

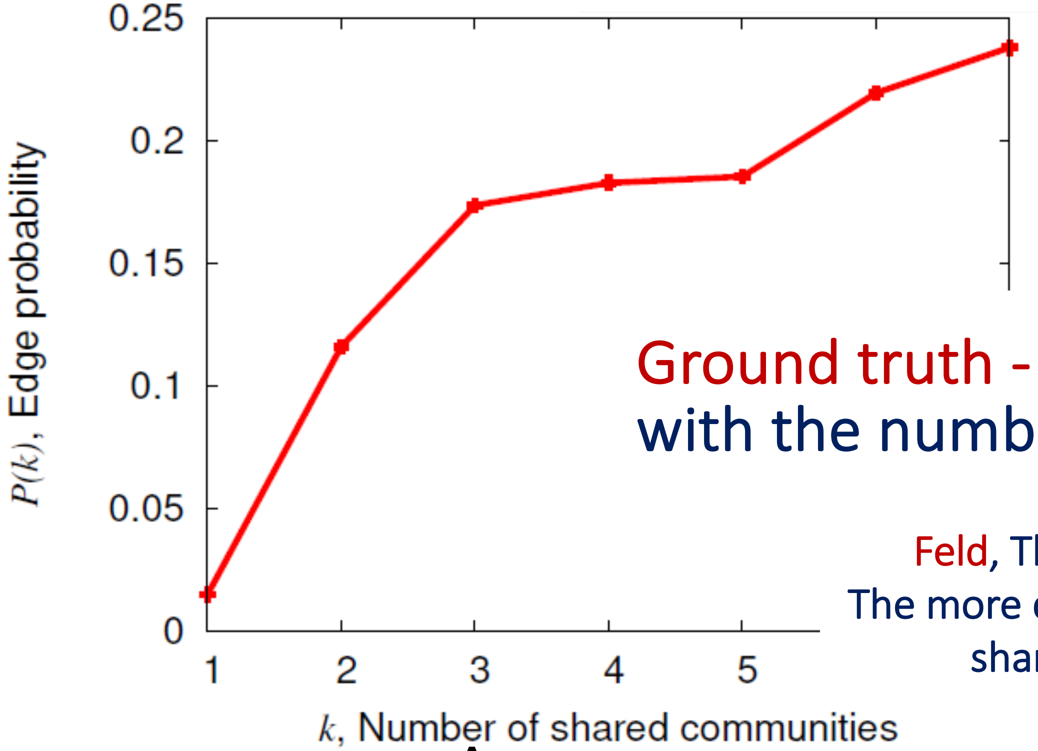
- denser and denser region
- contains 60% nodes and 80% edges
- a region where communities overlap (as tiles)



# Overlapping communities model



Edge density is bigger in the overlap



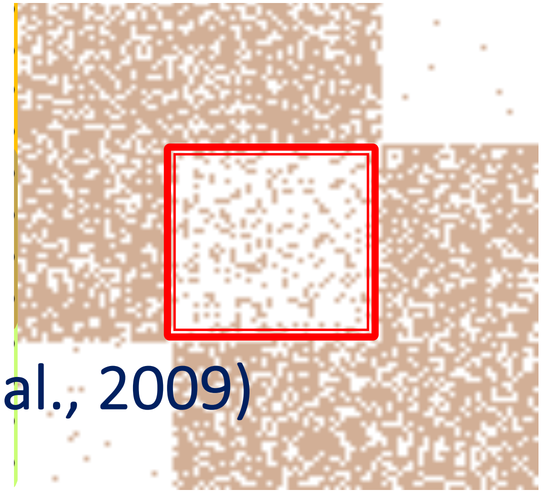

Amazon

Ground truth - Edge probability increases with the number of shared communities

Feld, The focused organization of social ties, [1981]
The more different foci (communities) that two individuals share, the more likely is that they will be tied

# Overlapping communities model

most assume a wrong overlapping model !

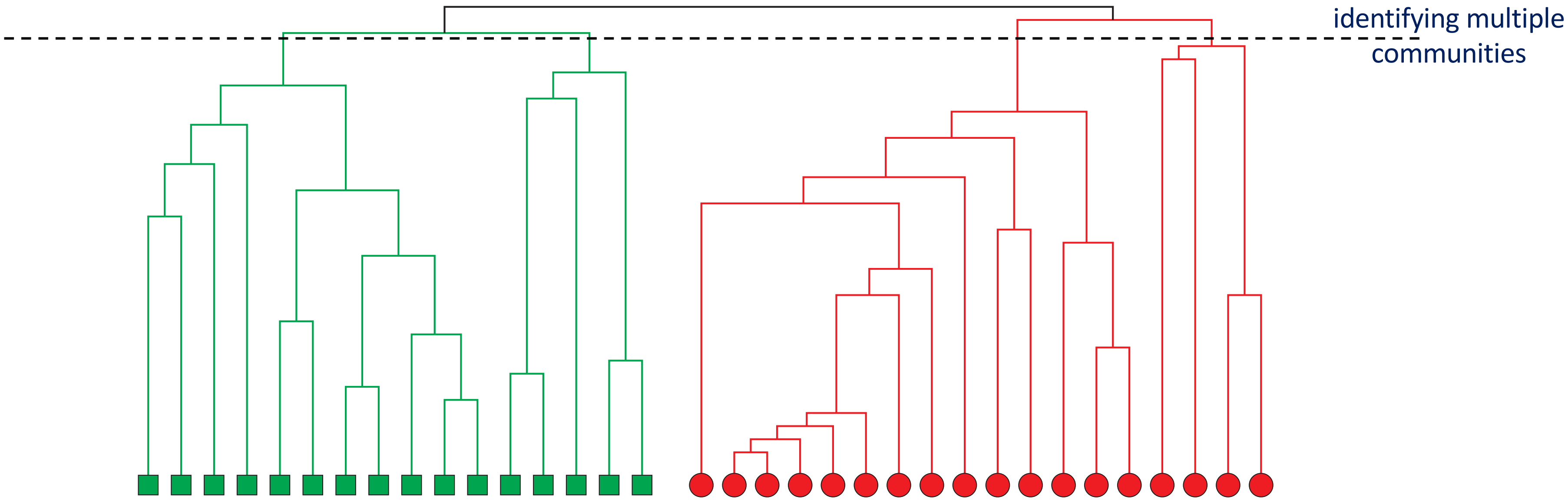


Available algorithms

- Clique percolation (Palla et al., 2005)
- Link clustering (Ahn et al., 2010) (Evans et al., 2009)
- Clique expansion (Lee et al., 2010)
- Mixed membership stochastic model (Airoldi et al., 2008)
- Bayesian matrix factorization (Psorakis et al., 2011)
- …
- BigCLAM (Yang and Lescovec, 2013)
- …

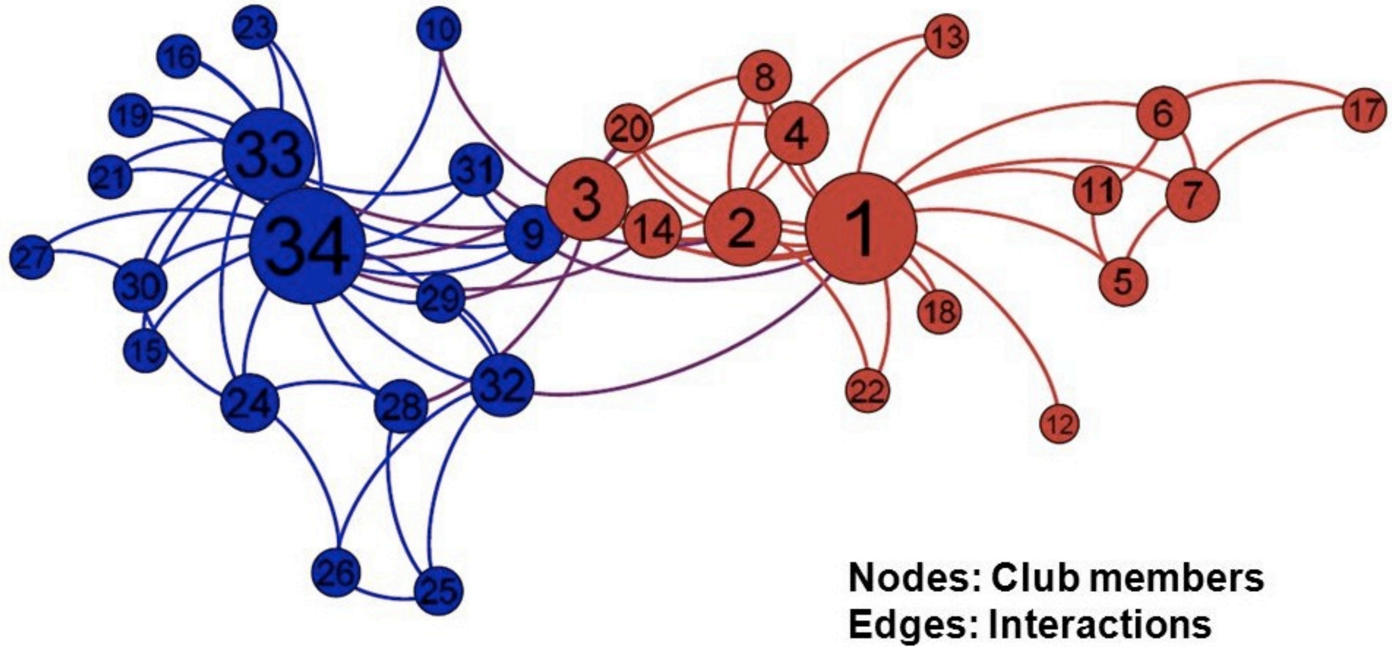
# Dendrograms

# Dendrograms




- A (agglomerative) hierarchical clustering algorithm
- Progressively add edges, from the strongest and ending with the weakest ones
- Example for Zachary's Karate club network

# Zachary's Karate club (social) network




- Ground truth
- Observe social ties and rivalries in a university club
- During observation conflict led the group to split
- Split could be explained by a minimum cut

# Pros and cons of dendrograms

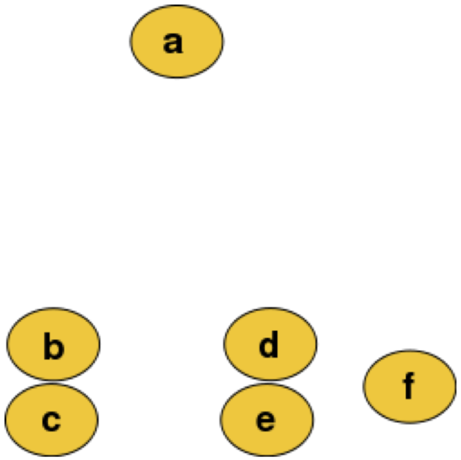
Pros and cons

- Performance strongly depends on the chosen weight (local weight definitions typically provide weak solutions)
- Can be agglomerative or divisive, but adding strongest weights is in general weaker that deleting weaker ones
- May provide poor results
- Useful method, far from perfect

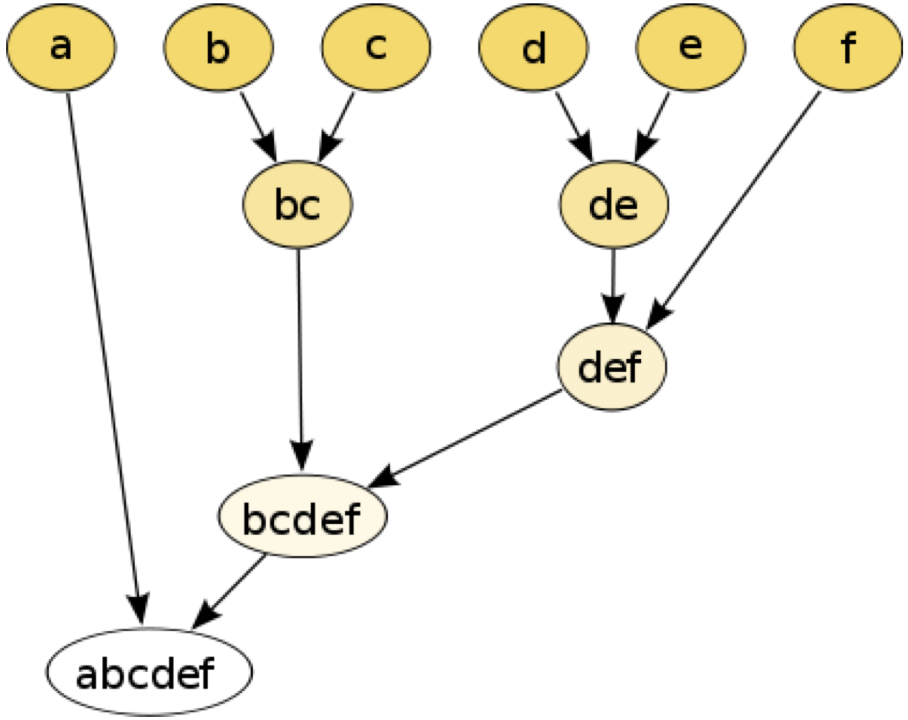
# Agglomerative hierarchical clustering example

Weight: Euclidean distance

Data:




Dendrogram



# Edge betweenness

Girvan and Newman, Community structure in social and biological networks [2001]
https://www.pnas.org/content/99/12/7821

- Use the concept of edge betweenness

$$b_{ij} = \sum_{(k,\ell)\in\mathcal{N}^2} \frac{\sigma_{k,\ell}(i,j)}{\sigma_{k,\ell}}$$

where $\sigma_{kl}$ is the # of shortest paths connecting $k$ to $l$, and $\sigma_{kl}(i,j)$ the subset of these including edge $(i,j)$

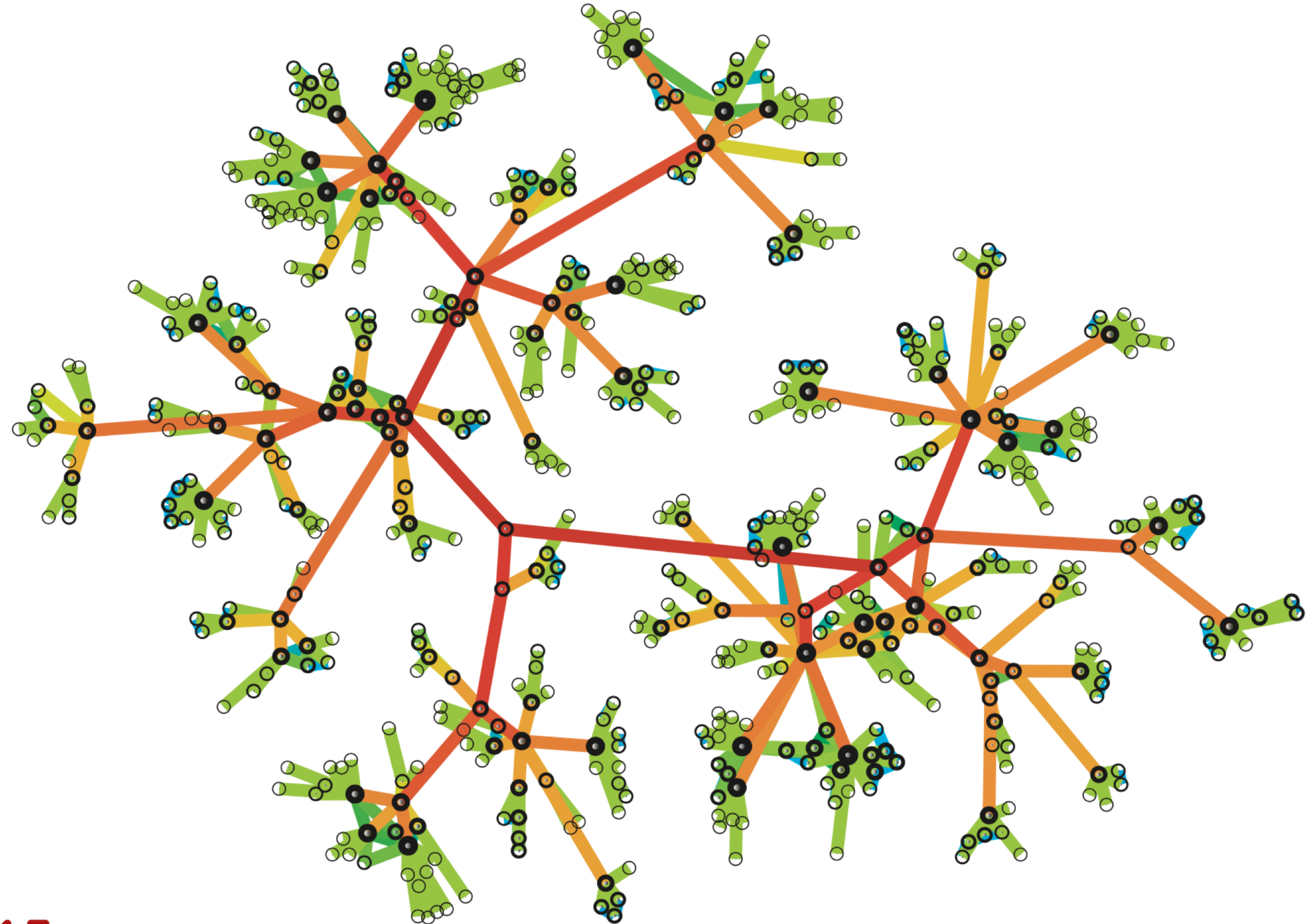
- Expresses centrality of a link in the network
- Can be normalized to range [0,1]

$$(b_{ij} - b_{\min})/(b_{\max} - b_{\min})$$

- Generalization of vertex betweenness (Freeman 1977) (Anthonisse, 1971)

MIME.

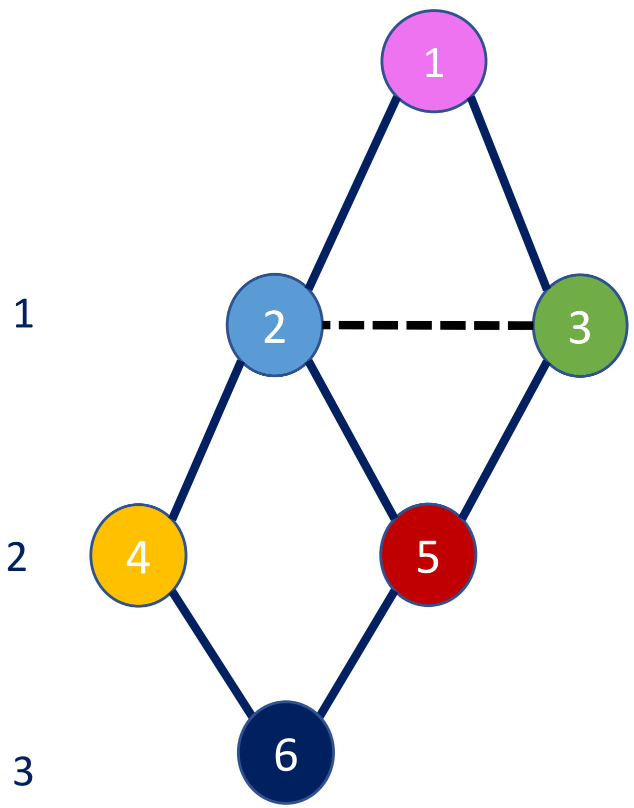
# Edge betweenness in a cellular call network



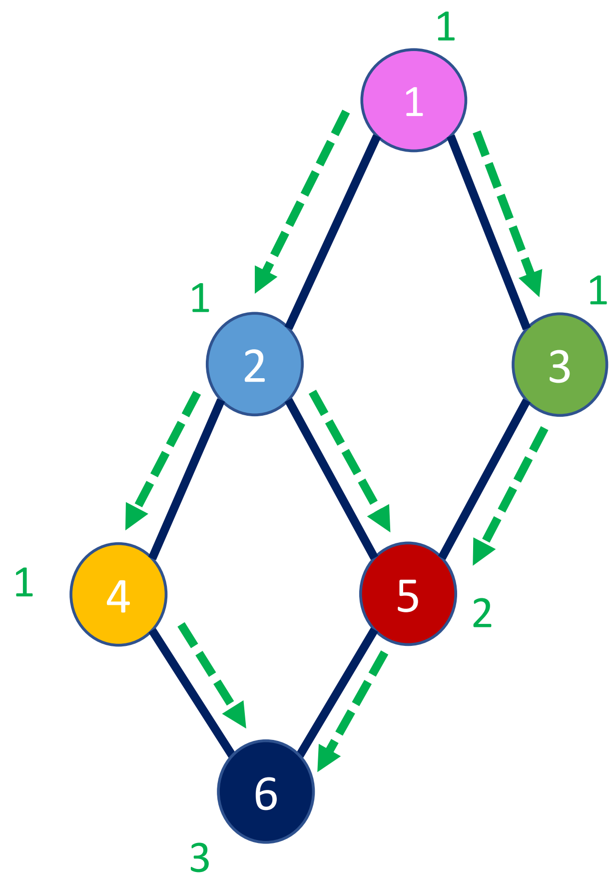
MIME

# Calculating betweenness

Breadth first search (from 1)

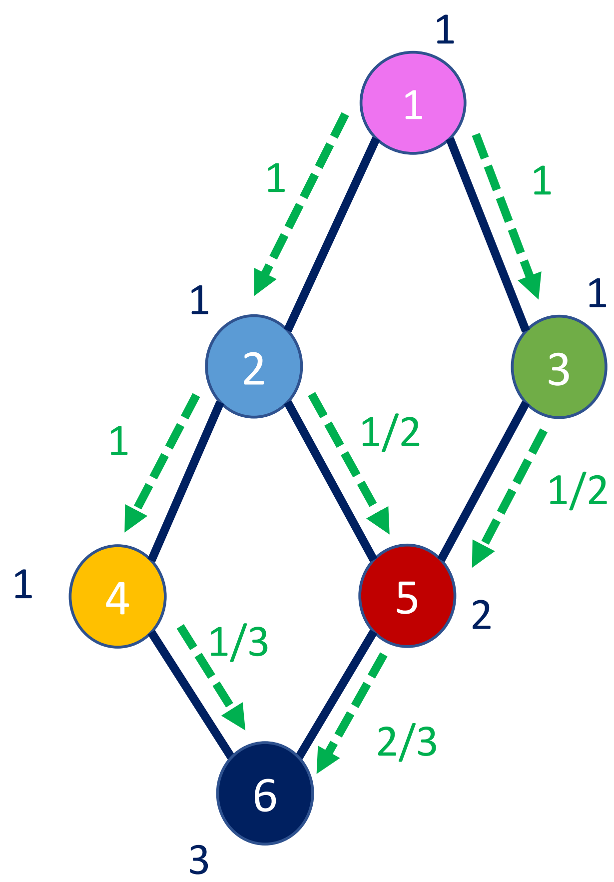



Count # of shortest paths (from 1)

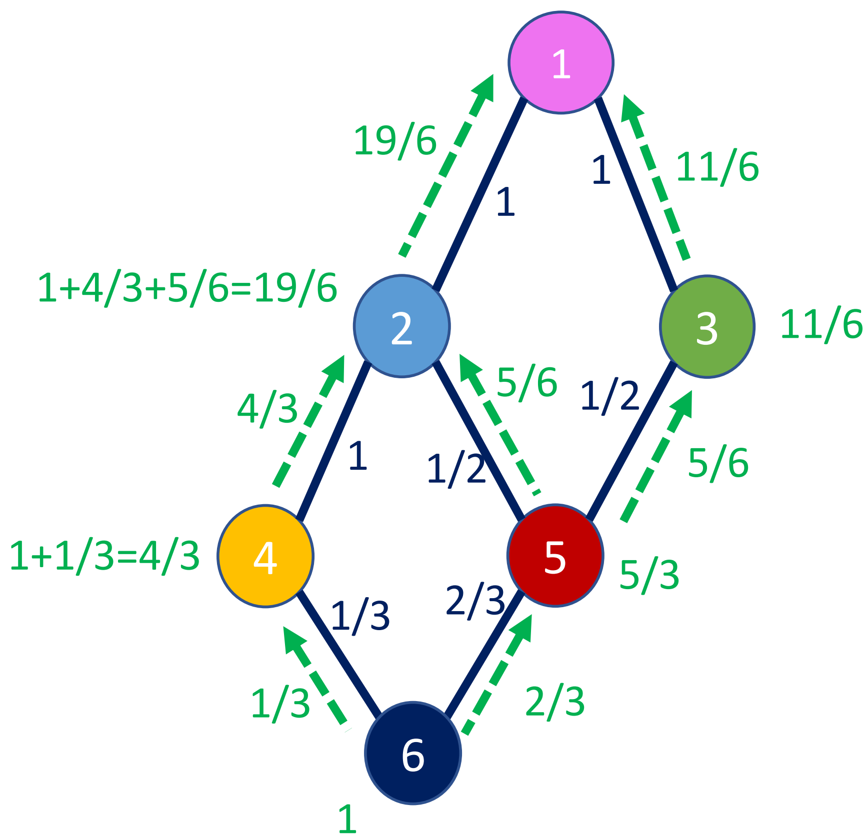


# Calculating betweenness

Measure edge flow (fractions)




Measure edge betweenness (from 1)




... then repeat for all other nodes!!! O(LN)

# Girvan-Newman method

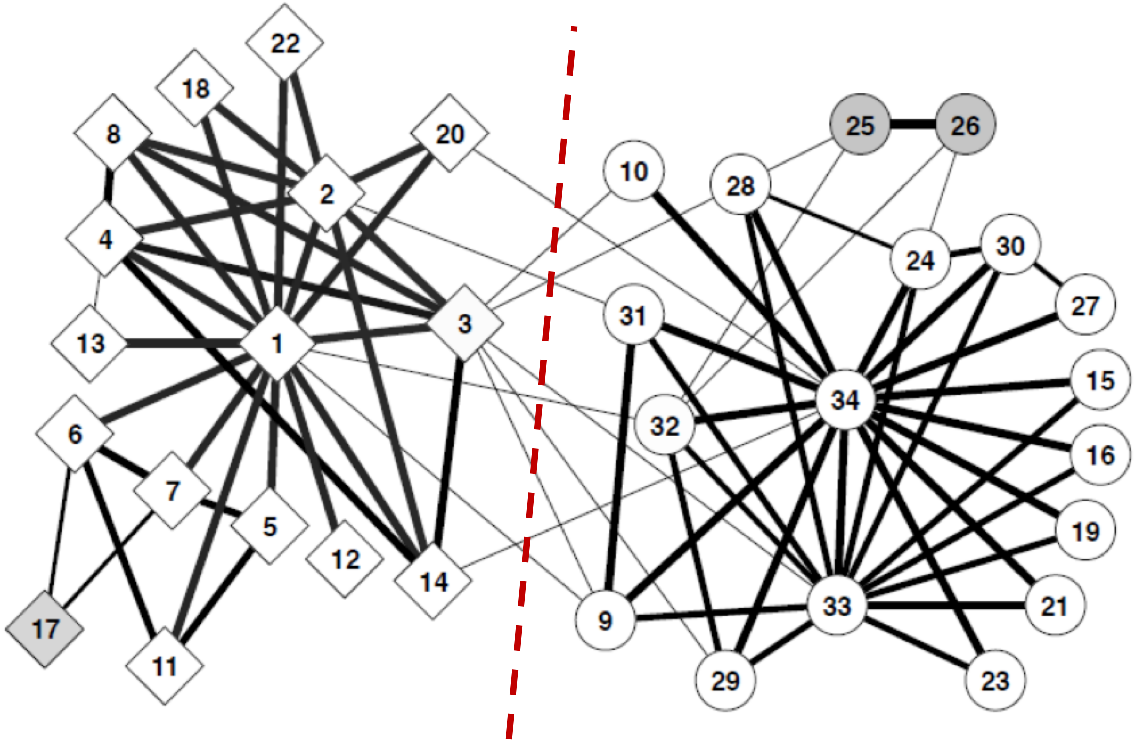
- Repeat until no edges are left in the graph
- (Re)calculate edge betweenness in the current graph – complexity O($LN$) by using a smart algorithm
- Remove edges with highest betweenness
- Connected components are communities

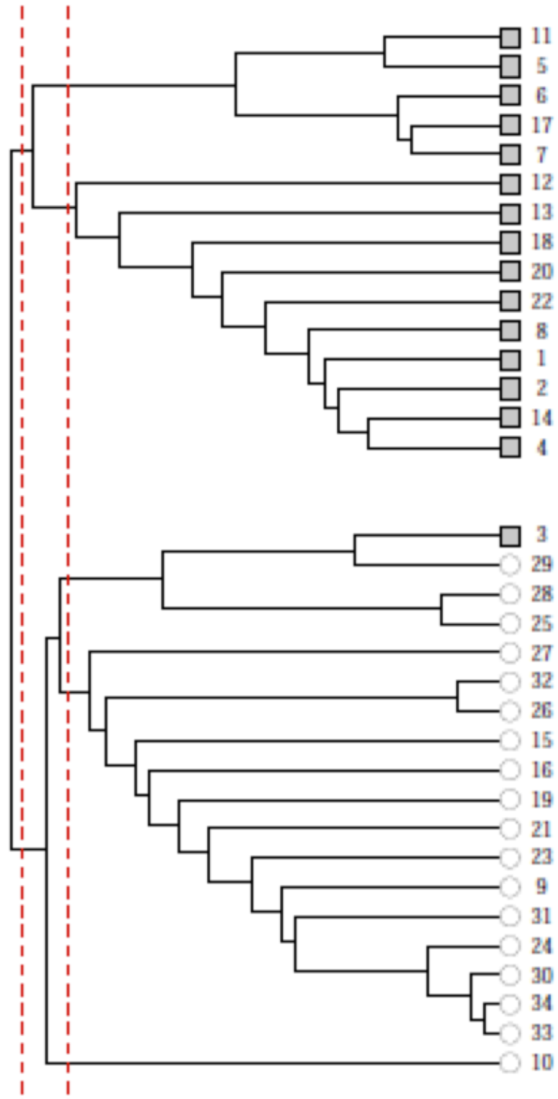
It is a (divisive) hierarchical clustering algorithm

Complexity O($L^2N$)

Recalculation step is essential to detect meaningful communities

# Zachary's karate club example




1 - instructor
34 - president
Correct but node 3

# Questions ?